Infectious Diseases of Poverty Research Data Management

Opportunities and challenges to capacity strengthening in Africa



Amélie M. Julé

Green Templeton College, University of Oxford

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (D. Phil.) in Clinical Medicine

Trinity Term 2018 – July 8, 2018

Infectious Diseases of Poverty Research Data Management: Opportunities & challenges to capacity strengthening in Africa

Amélie M. Julé – Green Templeton College, University of Oxford

Thesis submitted in Trinity Term 2018, in partial fulfilment of the requirements for the degree of Doctor of Philosophy (D. Phil.) in Clinical Medicine.

Abstract

Data Management (DM) is a critical, yet often neglected step of research studies, essential to guarantee the integrity of data and validity of conclusions derived from it. To enable meaningful data use (and re-use), investigators and their institutes must therefore be equipped with appropriate DM capacity. Little empirical research so far has explored DM practices in clinical research or the barriers and enablers of their implementation, especially in Low and Middle Income Countries (LMICs). Using a mixed-methods design, this thesis addresses this important and timely gap.

A systematic review and meta-analysis of schistosomiasis treatment efficacy studies first demonstrates the pitfalls of aggregated data for re-use, and uncovers varied data management and analysis practices within this research community. The case of schistosomiasis (a neglected, under-researched tropical disease) is contrasted with that of malaria. Through its well-established data-sharing platform, the World-Wide Antimalarial Resistance Network (WWARN) enables an in-depth exploration of datasets and accompanying study documentation from 108 antimalarial trials performed in Africa. This exploratory analysis shows that DM capacity development needs particularly concern comprehensive annotation and transparent handling of study datasets.

Two qualitative case studies in Dakar, Senegal and Brazzaville, Republic of the Congo complement this systematic description of DM across Africa. Through in-depth interviews and observations of local practices, I examine clinical research staff's knowledge, attitudes, experiences and perceptions of DM, as well as the constraints within which they operate DM systems (DMS). Findings confirm that robust DMS depend on adequate financing and strong infrastructure, but also on an overall enabling environment where the importance of DM is recognised.

Results from this work highlight priority areas for clinical research DM capacity strengthening in LMICs, emphasising the necessity to raise awareness of DM's complexity and to establish DM as a valued professional role.

Acknowledgements

These DPhil years did not come without their own challenges, and many people were critical in helping me to overcome these. Duly acknowledging all contributions would require another chapter to this thesis (!), but a few need mentioning.

Primarily, I address my heartfelt thanks to my four supervisors: Prof. Trudie Lang, Prof. Philippe Guérin, Prof. Vicki Marsh and Prof. Piero Olliaro. Each of you has been influential in keeping me on course at one point or another; and I am truly grateful you never expressed doubts over my ability to complete this work (even when I personally doubted it so many times!).

I received generous funding from the Medical Research Council UK through the Nuffield Department of Medicine Prize studentship, which enabled me to conduct my research, but also to attend diverse and highly-inspiring courses and conferences worldwide.

For their scientific contribution and varied help, I thank all my colleagues from The Global Health Network and IDDO (past and present). The advice of Hazel Ashurst, Georgina Humphreys, Laura Merson, Aileen Sheehy and Kasia Stepniewska was particularly significant for the quantitative piece on malaria. Martin Walker and Julia Halder provided tremendous help in the collation and analysis of data for the schistosomiasis meta-analysis. Sada Aliyeva, Liam Boggs, Tamzin Furtado, Nina Jamieson, Alexis Segrt, Iveta Simera, Francois van Loggerenberg and Lauren Whelan provided administrative and moral support in countless occasions.

For the qualitative work, I am particularly grateful to my local collaborators in Senegal, Prof. Oumar Gaye and Dr. Magatte Ndiaye, and in the Congo, Prof. Francine Ntoumi and Dr. Félix Koukouikila-Koussounda. Penda Touré and Carine Kades significantly helped with the logistics of my fieldwork, thus making these first ever visits in Africa very smooth. I also thank all the study participants, especially the data managers at UCAD, for they were not only helpful with my research, but also determined to make me experience the Senegalese 'terranga'.

I thank my fellow DPhil students and Postdocs from the Trop Med Department, Astrid, Gina, Prabin, Makoto, Shirley, Delphine and Mandy – lunch and coffee breaks with you in (and out of!) the basement were a real breather. My friends' support extended much beyond Oxford, and I particularly thank Oli, Anna, Marine, Tanya and Yoann for 'remotely being there'. I missed being able to reach out to you on a daily basis, but it was a relief to know I could always count on you.

I thank my parents for letting me pursue my demanding goals (even when you were worried how I would cope with the stress of it all!); and my sister Mathilde for first inspiring me to aim for the sky.

Finally, I want to thank three Alex who play(ed) a significant role in getting me to this point. Alexandre Bodin – I doubt I would have become a 'normalienne' without your support at the end of three exhausting years of 'classes prépa', and your help getting me through the final stage of the 'concours' is something I will never forget. Alessandro Gozzi and the whole YogaVenue team – discovering yoga during my DPhil has been a lifesaver, and the positive energy from the Oxford yoga community you bring together made my day more than once (in fact, 348 times!) when I so greatly needed it. The very best for the end, Alexander Mietke, my soon-to-be husband – I trust you have seen the best and worst of me during this DPhil, and I feel blessed you want to keep holding my hand throughout the years and adventures to come.

Contents

I Introduction, background & methodology **1**

1 Strengthening the capacity for clinical research data management: rationale and gaps **2**

1.1 Justifications for this research 3

1.1.1 The significance of data management for data sharing 4

1.1.2 Balancing research powers in the data sharing era 7

1.2 Research question and aims 8

1.3 Working definitions and scope 10

1.3.1 What is clinical research? 10

1.3.2 What does research data management encompass? 12

1.3.3 Why focus on infectious diseases of poverty? 15

1.4 Literature review and past work 17

1.4.1 Handling 'routine' clinical data 19

1.4.2 Data centralisation, harmonisation and curation: managing research data for sharing and re-use 23

1.4.3 Guidelines: defining 'good' practice *versus* strengthening the local capacity 25

1.4.4 Moving forward the capacity for clinical research data management in LMICs 27

2 Inquiry strategy: the value of mixed methods and systems thinking to characterise data management 29

2.1 Researcher's positioning: A pragmatic perspective with a post-positivist influence . . . 30

2.2 Applying a 'complex systems' lens to data management . . . 32

2.3 A globally inductive approach based on mixed methods . . . 35

II Quantitative work 39

3 Introduction to the scoping review and meta-analysis of antischistosomal trials 40

3.1 Schistosomiasis: biological, clinical and epidemiological facts . . . 42

3.1.1 *Schistosoma* lifecycle . . . 42

3.1.2 Physiopathology and diagnosis . . . 44

3.1.3 Disease spread and burden . . . 44

3.2 Preventive chemotherapy and the threat of drug resistance . . . 45

3.2.1 Therapeutic options . . . 45

3.2.2 Control strategy . . . 46

3.3 Specific objectives of the first quantitative piece . . . 47

4 Methods: Systematic approach to scope the amount and variety of existing drug efficacy data 49

4.1 Identification of relevant references . . . 50

4.1.1 Literature search . . . 50

4.1.2 Eligibility screening . . . 52

4.2 Data collection and preparation . . . 55

4.2.1 Estimation of participants' numbers . . . 56

4.2.2 Corrections for homogenisation of Egg Reduction Rates (ERRs) and Mean Egg Counts (MECs) . . . 57

4.2.3 Assignment of Global Positioning System (GPS) coordinates to study sites . . . 60

4.3 Data synthesis and analysis . . . 61

5 Results: Existing trial data and issues with aggregated measures of antischistosomal efficacy 63

5.1 Eligible studies and participant groups . . . 64

5.2 Quantitative synthesis . . . 67

5.2.1 Estimation and characterisation of existing IPD on schistosomiasis treatment efficacy . . . 67

5.2.2 Heterogeneity in methodology and reporting . . . 70

5.2.3 Homogeneity in data management tools . . . 73

5.3 Meta-analysis . . . 74

5.3.1 Tested hypothesis: clinical and public health aim . . . 74

5.3.2 Spatiotemporal coverage of included groups . . . 75

5.3.3 Range of reported praziquantel efficacy . . . 77

5.3.4 Descriptive analyses prior to statistical modelling . . . 80

5.3.5 Univariate modelling from aggregated data . . . 85

6 Discussion: Significance for antischistosomal efficacy data management and sharing 88

6.1 Interpreting guidelines: variability in study methods and data analysis 90

6.2 Scope and directions for the development a data-sharing platform . . 92

6.3 *De novo* data collection and continuous efficacy monitoring needs . . 94

6.4 Conclusion on the schistosomiasis research landscaping exercise . . . 96

7 Exploration of data management strategies and outcomes in clinical research on malaria 100

7.1 Introduction to the second quantitative piece 101

7.1.1 Malaria: background and relevance as 'case' disease 101

7.1.2 Specific objectives and hypotheses tested 105

7.2 Proposed quantitative methods to explore clinical research data management systems 108

7.2.1 Selection of antimalarial trials from the WWARN repository (sampling) 108

7.2.2 Data extraction 109

7.2.3 Statistical analyses and response variables 111

7.3 Results: Variations in malaria research data management 113

7.3.1 Included trials and temporal trends in retrieved data 113

7.3.2 Correlations (or lack thereof) between study attributes and data management indicators 119

7.4 Discussion: Lessons from the exploratory analysis of WWARN data . 124

7.4.1 Data management aspects in need of greater attention 125

7.4.2 Threats to the validity and generalisability of findings 129

7.4.3 Conclusion on the exploration of data management strategies and outcomes in malaria research 134

III Qualitative work 136

8 Introduction to the in-depth case studies 137

8.1 Motivations for a qualitative exploration of clinical research data management systems 138

8.2 Case study sites and their contrasting national environments 138

8.2.1 'UCAD Parasitology' and its long history of medical education and research in West Africa . . . 139

8.2.2 'The Foundation' (FCRM) as a pioneer and catalyst for clinical research in Middle Africa . . . 141

8.3 Specific objectives of the qualitative piece . . . 142

9 Qualitative methods 144

9.1 Sampling of cases and participants . . . 144

9.1.1 Selection of case research institutes . . . 145

9.1.2 Selection and recruitment of participants . . . 146

9.2 Qualitative data collection procedures . . . 147

9.2.1 Protocol adaptations in light of the situation at sites . . . 147

9.2.2 In-depth, semi-structured interviews . . . 148

9.2.3 Informal conversations and participant observations (ethnographic approach) . . . 149

9.2.4 Ethical considerations and informed consent . . . 150

9.3 Qualitative data processing and analysis . . . 151

9.3.1 Data preparation . . . 151

9.3.2 Analytical strategy . . . 152

10 Results: Challenges and opportunities in data management at research institutes in Africa 153

10.1 Participants recruited and institutes visited . . . 154

10.2 Establishing a permanent, institutional data management workforce . 160

10.2.1 In-country human resources to draw on . . . 160

10.2.2 Constraints and opportunities in staff recruitment imposed by organisations and project cycles . . . 165

10.3 Available material and its influence on data management procedures . 168

10.3.1 Similarities in data management challenges and opportunities across institutes and countries 168

10.3.2 The case of the Congo: Limitations of DMS predominantly relying on spreadsheet software 174

10.3.3 The case of Senegal: Barriers and enablers of transitioning from paper-based to EDC systems 176

10.4 Balancing change and continuity in clinical research DMS 179

10.4.1 The role of change agents in improving local practices 179

10.4.2 Working as a team in an evolving environment 181

10.4.3 Documenting to ensure continuity 183

10.5 Developing oneself, developing others: barriers and enablers of career development 185

10.5.1 The 'art' of people management 185

10.5.2 Barriers and enablers to training 190

11 Discussion of qualitative findings **197**

11.1 Shedding light on data management systems using the Aragón framework 198

11.1.1 Technology: The need for servers and role of change agents in upgrading data management practices 200

11.1.2 Finance: The impact of participants' values and norms on funding accessibility and use 204

11.1.3 The unexpected influence of organisational hierarchy on human resources establishment 205

11.1.4 Relationships, communication and informal rules complement formal management skills and processes 206

11.2 Raising the profile of clinical research data management 208

11.3 Trustworthiness of the qualitative work 211

11.3.1 Transferability of findings to other settings 212

11.3.2 Sampling effects and participants' openness 214

11.4 Conclusions on the qualitative piece of work 217

IV General discussion and conclusion 219

12 Strengthening data management systems: global lessons from the mixed-methods study 220

12.1 Characterising the variability of clinical research data management practices in LMICs 221

12.1.1 Data management processes vary among studies and institutes, and are documented with variable depth 222

12.1.2 Data management tools used by local investigators in LMICs are not so varied 225

12.1.3 Data management personnel and actors vary greatly 227

12.2 Clarifying the reasons behind variability in data management practices 228

12.2.1 Hard components 229

12.2.2 Soft tangible components 235

12.2.3 Soft intangible components 241

12.3 Understanding what is needed to develop the capacity in LMICs . . . 246

12.3.1 Developing a culture of data management crossing disease and discipline boundaries 247

12.3.2 Targeting data management training to audiences 249

12.3.3 Empowering change agents 252

12.4 Conclusion 253

V Appendices **256**

A Description of datasets compiled for the schistosomiasis scoping review and meta-analysis **257**

A.1 Datasets A and B: Search results and two-stage screening process . . 258

A.2 Datasets 1 to 5: Quantitative synthesis data 264

A.3 Datasets 6 and 7: Meta-analysis data 274

A.4 Quality assurance 283

B Supplementary methods and analyses of malaria data shared with WWARN **284**

B.1 Development of the analytical framework to characterise data management practices 286

B.1.1 Elements drawn from the literature 287

B.1.2 Input from experts 289

B.1.3 Pilot application and instrument refinement 290

B.2 Communications with data contributors 293

B.3 Data preparation for statistical testing (correlation analyses) 295

B.3.1 Study 'complexity' – trial design 295

B.3.2 Study 'complexity' – trial size 297

B.3.3 Study setting and sources of support 299

B.3.4 Interconnectedness of investigators in the malaria community (network analysis) 302

C Topic guides and intermediate qualitative data analysis **306**

C.1 Methodological tools and procedures 306

C.1.1 Participants' information brochure 306

C.1.2 Interview topic guides 311

C.2 NVivo coding tree and other charts produced during analysis 316

C.2.1 Organisational context (structure and resources) 316

C.2.2 People (behaviour and knowledge) 319

C.2.3 Data (procedures) 322

D First-author articles **336**

D.1 First publication: Systematic Review of Antischistosomal Trials . . . 336

D.2 Second publication: TDR Global Competency Framework for Clinical Research 354

Abbreviations

ACT	Artemisinin-based Combination Therapy
ADMIT	Association for Data Managers in the Tropics
ANOVA	Analysis of Variance
ANSD	Agence Nationale de Statistique et Démographie (National Agency for Statistics and Demography)
CANTAM	Central Africa Network on Tuberculosis, HIV/AIDS and Malaria
CDC	Centers for Disease Control and Prevention (USA)
CDISC	Clinical Data Interchange Standards Consortium
CERVE	Centre d'Etudes sur les Ressources Végétales (Centre for Vegetal Resources Studies)
CHEPSAA	Collaboration for Health Policy and Systems Analysis in Africa
CI	Confidence Interval
CR	Cure rate
CRF	Case Report Form
DFID	Department for International Development (UK)
DHIS-2	District Health Information System 2
DM	Data Management
DMP	Data Management Plan
DMS	Data Management System
DNDi	Drugs for Neglected Diseases initiative

ECRIN	European Clinical Research Infrastructure Network
EDC	Electronic Data Capture
EDCTP	European and Developing Countries Clinical Trials Partnership
EDISON	Education for Data Intensive Science to Open New science frontiers
ERR	Egg Reduction Rate
FCRM	Fondation Congolaise pour la Recherche Médicale (Congolese Foundation for Medical Research)
FDA	Food and Drug Administration (USA)
FMPOS	Faculté de Médecine, Pharmacie et Odontostomatologie de l'UCAD (Faculty of Medicine, Pharmacy and Odontostomatology at UCAD)
FOSTER	Fostering the practical implementation of Open Science in Horizon 2020 and beyond
GAUL	Global Administrative Unit Layers
GCDMP©	Good Clinical Data Management Practices©
GCP	Good Clinical Practice
GDP	Gross Domestic Product
GM	Geometric Mean Egg Count
GPS	Global Positioning System
HDSS	Health and Demography Surveillance System
HIC	High Income Country
HIS	Health Information System
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome
HR	Human Resources
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICMJE	International Committee of Medical Journal Editors
ICT	Information and Communication Technology
IDDO	Infectious Diseases Data Observatory
INDEPTH	International Network for the Demographic Evaluation of Populations and their Health

IPD	Individual Participant Data
IRD	Institut de Recherche et Développement (Institute for Research and Development)
IRSEN	Institut de Recherche en Sciences Exactes et Naturelles (National Institute for Research in Exact and Natural Sciences)
IRSSA	Institut National de Recherche en Sciences de la Santé (National Institute for Research in Health Sciences)
ISED	Institut de Santé et Développement (Institute for Health and Development)
ISO	International Organisation for Standardisation
IT	Information Technology
LLITN	Long-Lasting Insecticide-Treated Net
LMIC	Low and Middle Income Country
LNSP	Laboratoire National de Santé Publique (National Public Health Laboratory)
MARCAD	Malaria Research Capacity Development
M&E	Monitoring and Evaluation
MEC	Mean Egg Count
MMV	Medicines for Malaria Venture
MNCP	Malaria National Control Programme
MORU	Mahidol Oxford Tropical Medicine Research Unit
MS	Microsoft®
MSF	Médecins Sans Frontières (Doctors Without Borders)
NTD	Neglected Tropical Disease
ODK	Open Data Kit
OECD	Organisation for Economic Cooperation and Development
OxTREC	Oxford Tropical Research Ethics Committee
PC	Preventive Chemotherapy
PD	Pharmacodynamics

PERCH	Pneumonia Etiology Research for Child Health
PI	Principal Investigator
PK	Pharmacokinetics
PMI	President's Malaria Initiative (USA)
PNLP	Programme National de Lutte contre le Paludisme (Malaria National Control Programme)
PNT	Programme National de la Tuberculose (Tuberculosis National Programme)
POLR	Proportional Odds Logistic Regression
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PZQ40	praziquantel, 40 mg/kg of body weight
QC	Quality Control
RCT	Randomised Controlled Trial
RR	Relative Risk
SA	Statistical analysis
SAC	School-Aged Children
SAP	Statistical Analysis Plan
SAREC/SIDA	Swedish International Development Agency
SCDM	Society of Clinical Research Data Management
SMC	Seasonal Malaria Chemoprevention
SNIS	Système National d'Information Sanitaire (National Health Information System)
SOP	Standard Operating Procedure
SQL	Structured Query Language
STHs	Soil-Transmitted Helminthiases
TDR	UNICEF/UNDP/World Bank/WHO Special Programme on Research and Training in Tropical Diseases
UCAD	Université Cheikh Anta Diop, Dakar

UCAD Parasitology	Service de Parasitologie et de Mycologie de l'UCAD (Department of Parasitology and Mycology at UCAD)
UCAD Public Health	Service de Médecine Préventive et de Santé Publique de l'UCAD (Department of Preventive Medicine and Public Health at UCAD)
UK	United Kingdom
USA	United States of America
USB	Universal Serial Bus
USD	United States Dollar
WARC	WWARN West Africa Regional Centre
WHO	World Health Organization
WHO-AFRO	WHO's Regional Office for Africa
WWARN	WorldWide Antimalarial Resistance Network

Part I

Introduction, background

& methodology

Chapter 1

Strengthening the capacity for clinical research data management: rationale and gaps

Data is the primary 'raw' output from research: after appropriate processing and interpretation, research data should enable investigators to advance knowledge in their field or to advise practitioners, as appropriate. In the context of clinical research, accurate data is essential for protecting human subjects enrolled in a trial or study, but also to protect all subsequent patients from harm that a change in clinical care based on invalid research conclusions may entail. Data management encompasses the processes whereby 'real life' information gets captured and converted into an analysable format [1], and as such it is the first means to guarantee the integrity of research data and the reliability of conclusions drawn from it.

In this opening chapter, I justify the timely need for a research project focussing on this key (yet largely overlooked, "undervalued and underfunded" [2]) aspect of clinical research. I then present my research question and aims, clarify the scope of this doctoral work, and review the literature of relevance to an understanding of the 'clinical research data management capacity strengthening' field.

1.1 Justifications for this research

Despite its crucial role in guaranteeing the validity of clinical research results, data management often suffers from a lack of attention (or awareness) from investigators, who rarely receive adequate training to perform this demanding set of activities [3,4]. Interestingly, data management is gaining more attention in the current context of demands for data sharing, as the research community realises the additional management and curation efforts that data sharing entails [5–7], and as the value of re-using data that has dubious quality gets questioned [8,9]. Regardless of discipline, data sharing and its barriers or enablers are much more widely discussed than data management in the scientific literature (Figure 1.1). This active discussion among researchers, funders and publishers around 'best practices' in data sharing provides a timely opportunity to jointly attract attention on data management. Nonetheless, the current primary focus on how to enable data sharing considers the issue the wrong way round: good data management is the essential precursor to purposeful data sharing [10], and as such, capacity strengthening efforts should target data management first and foremost.

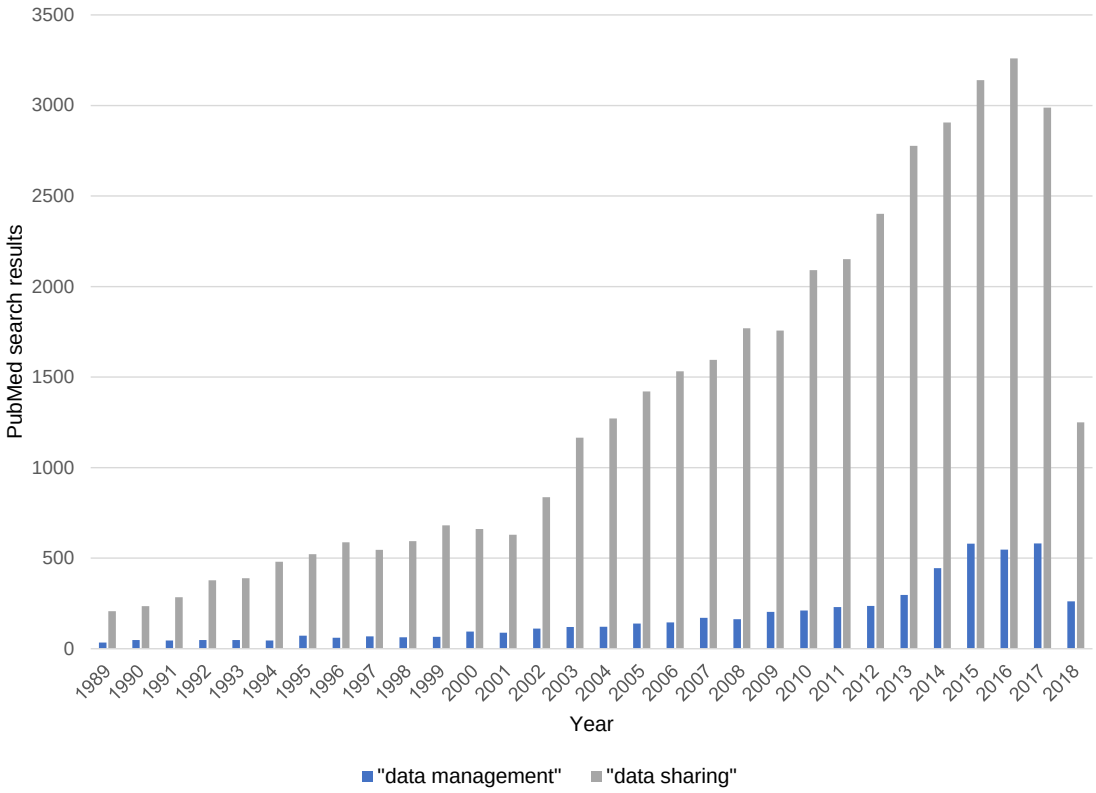


Figure 1.1: **Number of PubMed search results per year for "data management" and "data sharing" in the past 20 years** (search performed on 19 May 2018).

In the remainder of this section, I demonstrate the particular importance of data management within the context of data sharing, and I illustrate why a push for the latter makes it even more essential to strengthen the clinical research data management capacity of investigators and institutions based in **Low and Middle Income Countries** (**LMICs**).

1.1.1 The significance of data management for data sharing

Although strategies to address challenges around authorship concerns, ethical issues and governance frameworks remain debated [8, 11–14], expectations for researchers

to share data underlying their findings have greatly increased in the past few years, and mechanisms to facilitate the data-sharing enterprise have been developed. The number of platforms for dataset deposition has exploded, thus the recent need to also create directories of such repositories [15–18]. In parallel, multiple scientific journal publishers, research funders and academic institutions have adopted some data policy including a statement on data sharing; see for example, positions from the **International Committee of Medical Journal Editors** (**ICMJE**) [19] and varied other academic research stakeholders [20–22]. Furthermore, the pharmaceutical industry, alongside regulatory agencies and international organisations (including, the **World Health Organization** – **WHO**) are promoting and implementing data-sharing policies for clinical trials' data [12, 23, 24].

These growing demands and opportunities for data sharing are supposed to readily result in better management and validity of research data, as the scrutiny of peer review gains in breadth and depth, and as occasions to reproduce findings increase [25, 26]. Nonetheless, this promise is unlikely to be realised unless researchers are equipped to face the challenges of additional data management efforts, as necessary to reach the standards required for data sharing [27]. Because it is the first guarantor of continuous data quality, appropriate data management from the onset of a study is not only important but rather essential in the context of data sharing and re-use [10]. Sharing poor-quality data defeats one of the main motivations of research data sharing, which is to maximise (valid) findings and knowledge that can be derived from a set of data [8, 9]. Figure 1.2 illustrates the feedback loop between data management and sharing (via data quality), which makes the former essential

to the latter [10]. Strengthening researchers' skills and research organisations' capability to manage data ahead of its sharing is critical, and has potential to bring multiple benefits beyond the primary, most essential goal of ensuring research data is of high quality [5]. Of note, as the key link between data management and sharing, data 'quality' is also of interest to this thesis but not at its core; hence, the meaning and dimensions of data quality in clinical research are discussed in Appendix B.

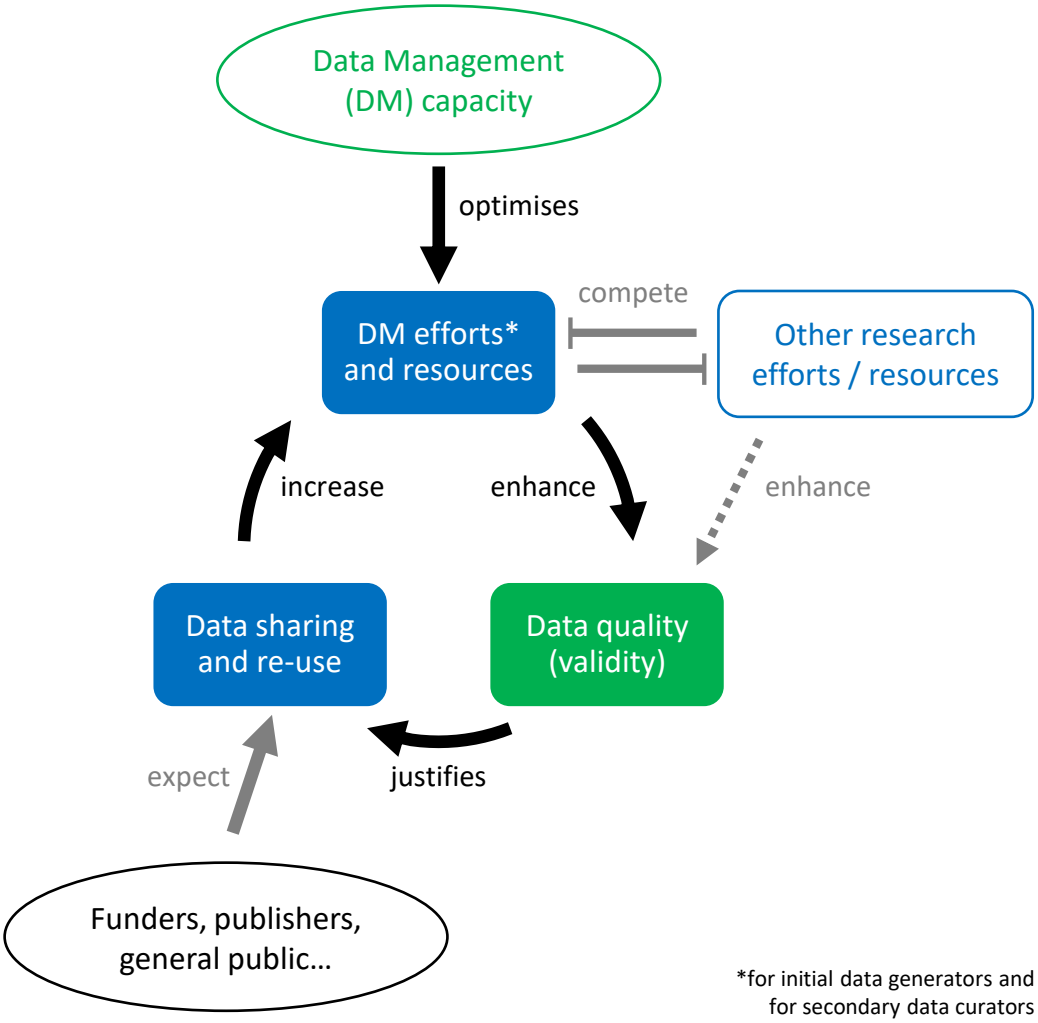


Figure 1.2: **The feedback loop between data management and sharing**, in the context of competing or driving factors.

A major benefit of getting data management right from scratch is to save substantial time and efforts later on. Clear and well-documented data management procedures help users understand the end data and assess its reliability; and indeed, secondary analysts were found to mostly rely on initial **Data Management Plans** (**DMPs**) and study protocols to evaluate the accuracy of datasets they reuse [28]. In absence of careful planning and transparent data handling, the final stages of preparing high-quality data for sharing can become a time-consuming and resource-intensive process [6, 7]. In reviewing surveys of researchers' opinions on data sharing in the UK and **United States of America** (**USA**), Kratz *et al.* thus found that the excessive time it takes to prepare data for sharing was the most consistent justification for not doing so [25].

This burden of data sharing on research resources has been widely discussed in terms of the financial cost involved [11, 14, 21, 26]. Yet, bridging the gap in research data management capacity will require more than an increase in funding efforts; and not the least, the development of adequate competencies for investigators and research support staff [3, 4, 29].

1.1.2 Balancing research powers in the data sharing era

Researchers in LMICs often perceive their needs for data management capacity strengthening as greater than those of their counterparts in **High Income Countries** (**HICs**) [5]. This perceived imbalance in capacities acts as a disincentive for LMIC-based researchers to engage with data sharing, and drives their fear of data

'exploitation' by well-resourced research teams, who could rapidly analyse data without participating in the initial collection effort [11, 12, 21].

Past research has documented the many barriers to the conduct of clinical research in LMICs [30,31], thus suggesting that the capacity gap is not only 'perceived' but also real, and that it can be somewhat mitigated through knowledge exchange and 'symbiotic' research collaborations. Feedback from The Global Health Network users and surveys of the wider community of clinical researchers [32] further highlight data management as a particular area of need for capacity strengthening, especially in LMICs. Data sharing partnerships should therefore embed solutions to promote the development of a local data management capacity [33–35], in order for the process to be seen as "fair" [2] and "trustworthy" [36], thus fostering a mutually beneficial culture of clinical research data sharing amongst both HIC and LMIC actors. Multi-country partnerships should also acknowledge that limitations and opportunities differ with investigators' research environment and the resources it provides [29,37]; hence partners should work together towards addressing challenges in locally appropriate ways.

1.2 Research question and aims

Strengthening research data management capacity in LMIC settings is a recognised need [29, 32, 35]; yet, the practicalities, enablers and barriers towards achieving this goal remain elusive. In light of this important gap (made all the more timely by the exceptional attention that data sharing receives), I proposed the following research

question for my doctoral thesis:

What are the opportunities and challenges in clinical research data management for investigators in Low and Middle Income Countries, and how could capacity-strengthening initiatives address them?

This open question calls for a descriptive and exploratory type of research, focussed on identifying the difficulties and facilitators of 'good' clinical research data management in LMICs. In addition, my research outputs should provide pragmatic evidence to inform the design and implementation of capacity-strengthening initiatives, targeted to the specific needs and reality of conducting clinical research in LMICs. Building on the existing body of knowledge and experiences (further reviewed in section 1.4 below), the three following aims were thus deemed relevant and achievable:

1. To characterise the variability in data management practices across clinical research on specific diseases relevant to LMICs;

2. To clarify the reasons behind variability in data management practices, as well as their implications for data use and re-use;

3. To gain an understanding of what is needed to develop the data management capacity of clinical research staff and institutions based in LMICs.

1.3 Working definitions and scope

'Data management' is a broad concept, which does not have a single, widely-acknowledged definition. The same goes for 'clinical research', thus the necessity to propose working definitions for those terms, and to clarify the scope of work accordingly. In this section, I discuss the intended scope, as implied in the research question and aims; however, I progressively narrowed this scope down, in view of the cases I identified as appropriate to enable meeting the aims stated above. I will comment on the relevance and specificities of the retained diseases (schistosomiasis and malaria) and geographical area (African countries) in later chapters, because these 'case boundaries' were not decided *a priori*. Indeed, I selected these restricted cases while my research project unfolded, and while the limitations or advantages of one case over the other became clear.

1.3.1 What is clinical research?

Most (but not all) of the work on clinical research data management performed as part of this thesis focusses on practices underlying clinical trials assessing drug efficacy. The WHO defines a clinical trial as "any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes" [38]. This definition helps to determine which studies involving human subjects require registration on a trial registry; and it is also the one recognised by the ICMJE [39, 40], hence the one adopted in this thesis.

Clinical research is a wider term, covering trials but also clinical studies based on other designs, including observational studies, such as case-control or cohort ones (authors from diverse medical specialities have published methodological articles describing the different study designs in clinical research, see for instance [41, 42]). The efficacy of an intervention may still be at the core of an observational study, but with the latter design, the intervention's assignment to human subjects is not the decision of the investigator. All study types critically contribute to the understanding of health and disease states, and as such they call for equally reliable data management and other study processes [43], a point to which I will come back in the concluding discussion (Part IV). As a first example from the literature, the **Pneumonia Etiology Research for Child Health** (**PERCH**) study provides an excellent example of applying high data management standards (as traditionally applied in clinical trials) to an observational, case-control study [44].

Note on the emphasis on clinical trials

Clinical trials represent only a fraction of the varied study designs that are relevant in infectious diseases of poverty research [45]; however, they present several attributes that justify their choice as an area of focus for this work on clinical research data management in LMICs. Trials are an easily identifiable type of clinical study, which is highly regulated and accordingly reliant on plenty of guidelines and procedures. The **International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use** (**ICH**) develops guidelines for good and ethical clinical trial practice [46]; and the **Food and Drug Admin-**

istration (**FDA**) provides an example of national regulatory framework for such clinical research [47].

I expect that this rather structured regulatory and operational framework should result in more readily comparable processes across clinical trials (including in terms of data management), thus facilitating the identification of key variations and commonalities in practices from one setting to the next. Regulations around clinical trials also mean that some investigators in LMICs perceive them as cumbersome and complicated, especially where locally-adapted interpretations of international frameworks like the ICH-GCPs are lacking [43, 48]. It would be interesting to know if and how these limiting perceptions transfer to data management. Finally, trials are of great relevance for infectious diseases of poverty such as malaria: the threat of (recurring) antimicrobial resistance requires continued monitoring of drug efficacy, which primarily relies on methodologically sound clinical trials [49, 50].

1.3.2 What does research data management encompass?

Data management involves numerous activities, and as such it takes different meanings for different stakeholders [51] and at different stages of the research data lifecycle (Figure 1.3). Building on the definition proposed by the **Clinical Data Interchange Standards Consortium** (**CDISC**) [1], I will use the following definition:

Data management constitutes the whole of the activities by which the analysis dataset is systematically created, based on data that has been

previously collected (whether primary data directly from the field, or secondary data from partners or data repositories) and specifically selected for its potential to enable answering the study question.

This definition excludes the data collection- and data analysis-related activities *per se*, but includes all activities which may happen 'in between' those two ends of the research study lifecycle, as well as their planning.

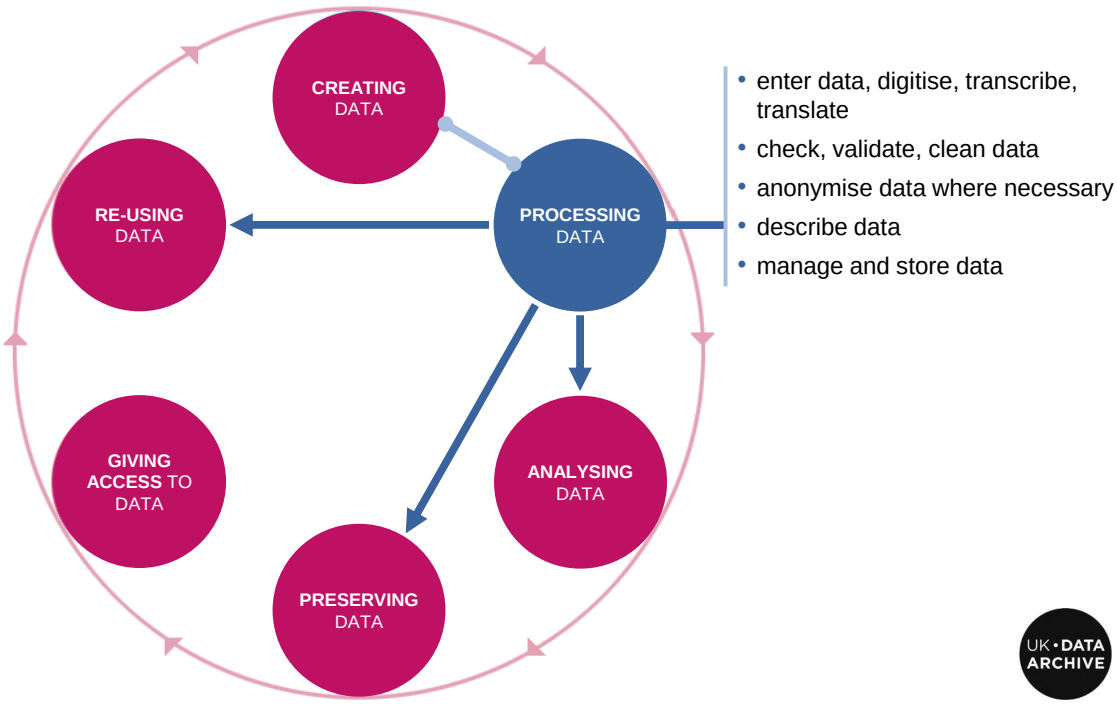


Figure 1.3: **Research data lifecycle, adapted from the UK Data Archive [52].** The "processing data" step and its interconnections with other steps will be most closely examined in this investigation of data management practices.

The data management activities most closely examined during this doctoral work therefore fall within the "processing" phase of the generic research data lifecycle proposed by the UK Data Service (formerly, UK Data Archive) [52], with a particular

interest in how this processing step connects with data analysis, relates to previous planning steps, and impacts on future conservation and re-use (Figure 1.3). Since I started my doctoral studies, the UK Data Service has updated its proposed representation of the research data lifecycle [53]. Despite important improvements, I here present the past version, as it guided my thinking throughout the planning and implementation of my research project. In the general discussion (Part IV), I will discuss the new take on the research data lifecycle in light of my own findings.

Variations on the term 'data'

In line with the FDA, CDISC defines data as "representations of facts, concepts, or instructions in a manner suitable for communication, interpretation, or processing by humans or by automated means" [1, 54]. Throughout this thesis, when used in reference to the topic under study, the term 'data' generically refers to any form of structured information that could support a quantitative analysis. Unless otherwise specified, data (as the study topic) is indeed assumed to be quantitative (whether numerical or categorical), because this is the data type most commonly generated in clinical research, especially during trials. Qualitative data (as collected for the purpose of this thesis) requires different data management and quality control procedures, and such text data is beyond scope as a study topic.

A dataset is understood as a collection of data file(s) and table(s) put together over the course of a study. Information (data) embedded in the study dataset is what concretely supports study conclusions. Similarly, a database is a set of elements

belonging to the same class (that is to say, sharing the same properties), generally organised in an informatics or other system that facilitates the search and retrieval of information [1]. An organised array of datasets may constitute a database; and a clinical trial data table may be seen as the simplest example of database, as it is an organised array of study participants' data.

Finally, in this thesis, meta-data is flexibly understood as any "contextual information required to understand data" [54], or in other words, any data description [1]. Meta-data may take the form of annotations embedded within the digital data file itself, especially as automatically generated by different software; but it may also take the form of a separate, manually created document that sheds light on the data (that is to say, a variable dictionary *per se*, a term interchangeably used with 'data dictionary', and also referred to by some researchers as 'codebook').

1.3.3 Why focus on infectious diseases of poverty?

'Infectious diseases of poverty' is a term used to refer to communicable (and often preventable) diseases that affect the poorest layers of the population in a disproportionate way, and further drive their poverty as a result of treatment costs and/or reduced ability to earn a living [45, 55]. At the macro-level, LMICs have an excess of deaths and morbidity due to diseases of poverty compared to HICs; and at the micro-level (including within HICs), individuals with lower revenues are also more prone to these diseases [56, 57]. In practice and within this thesis, the 'infec-

tious diseases of poverty' phrase often serves to refer to the 'big three', high-profile communicable diseases (**Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome** – **HIV/AIDS**, malaria and tuberculosis) jointly with the many **Neglected Tropical Diseases** (**NTDs**) [55]. NTDs are a varied group of 20 illnesses (as of May 2018 [58]) that are particularly under-studied as compared to the number they affect [59]. Similarly to malaria, most NTDs are caused by parasites with complex lifecycles, and whose vectors strive in tropical climates and contexts of poor hygiene and sanitation [60, 61].

In recent years, calls have multiplied to emphasise the need for LMICs to drive the health research agenda relevant to their population and current epidemiological situation [45, 48, 62–64], so that they may become the generators rather than recipients of the clinical data they need [65]. In addition, clinicians and investigators who daily come in contact with patients suffering from NTDs or other tropical diseases are most likely to take an interest and seek a culturally acceptable solution to these infections [66]; and indeed, a disproportionate share of scientific articles authored by researchers affiliated with institutions based in LMICs concerns NTDs [59]. This research focus and burden of infectious diseases of poverty in LMICs highlight their significance for local investigators, and the relevance of using these diseases as a case study for an exploration of clinical research practices and data management capacity in low-resource settings.

1.4 Literature review and past work

Throughout my doctoral studies (conducted between October 2014 and June 2018), I regularly visited the literature on the three interrelated concepts of data management, data quality and data sharing. While my interest lies primarily on the first, Figure 1.1 illustrated that academic articles on data management in the biomedical sciences are rather few, let alone those concerning LMICs. More specifically, a PubMed search on 19 May 2018 for ("data management" AND "developing country[MeSH Terms]") returned no result at all; and one search for ("data management" AND ("low and middle income" OR "low income" OR "middle income")) returned 68 articles. In parallel and as already mentioned, the push for research data sharing leads to the formulation of multiple data policies and guidelines by different stakeholders (see FairSharing [18] for a listing). Such documents may advise on how to manage data and prepare it for sharing, although guidance often lacks in clarity and remains open to interpretation [21, 67]; or, in the words of Wilkinson *et al.* "what constitutes 'good data management' is ... largely undefined, and is generally left as a decision for the data or repository owner" [10]. Finally, data-sharing policies and guidelines rarely specify how researchers or institutes could develop the capacity to implement the proposed 'good' data management practices.

In the context of health research, 'capacity building' can be defined as "the process of empowering individuals, institutions, organisations and nations to: define and prioritise problems systematically; develop and scientifically evaluate appropriate solutions; and share and apply the knowledge generated" [68]. The phrase is

often used interchangeably with 'capacity development' and 'capacity strengthening'; however, 'capacity building' suggests that capacity needs to be created from scratch, rather than reinforced where some infrastructure or skills pre-exist [69]. As particularly emphasised by Potter and Brough [70], developing capacity requires action at multiple levels, and deciding whether to primarily invest in training individuals or in providing them with an enabling institutional environment is a complex issue. Franzen *et al.* recently reviewed the major initiatives established to develop health and clinical research capacity in LMICs [71]. These mostly include focussing resources on one disease (vertical project) or a few champion institutions (centres of excellence), or engaging in North-South collaborations (partnerships and networks). The authors concluded that many of these approaches suffer from a lack of ownership and leadership from intended LMIC recipients [71], and fall short of considering the 'systemic approach' promoted by Potter and Brough [70].

In a similar fashion to Franzen *et al.* [71], I set out to identify and critically appraise articles on clinical research data management practices in LMICs, and especially those focussing on the opportunities and challenges to their implementation or improvement. However, faced with only few articles precisely discussing clinical research data management capacity strengthening in LMICs (as evidenced by test, iterative searches of different journal databases), I opted for a traditional review rather than for a focussed and systematic one [72]. This approach enabled me to include pertinent perspectives from 'neighbouring' fields (without seeking to exhaustively review each), thus providing me with a broader overview of major gaps and current priorities for different health or research data management stakehold-

ers. More precisely, I purposively looked for articles from the scientific literature and reports from the grey literature helping to answer one of the two following questions:

- What do we know about the current practices and capacity for clinical research data management in LMICs?
- What initiatives have been tried to strengthen these, and what were the challenges and opportunities encountered?

Relevant articles and reports belonged to three major themes and research fields, whose main interests and standpoint on data management capacity strengthening I successively summarise in sections 1.4.1 to 1.4.3, before concluding on that preliminary work in section 1.4.4.

1.4.1 Handling 'routine' clinical data

Health Information Systems (**HIS**) rely on high-quality data from varied sources to get an accurate picture of the current state of health and disease in a population and at a given time [73, 74]. In contrast with research, data underlying HIS is collected in a 'routine' context of standard clinical care, with the objective to monitor the performance of existing healthcare services rather than to acquire entirely new knowledge [75]. Here, I outline the core challenges and opportunities for LMICs in the management of clinical datasets collated outside of a research context, as documented by past work on HIS.

Recording and reporting data from health facilities to the national level

Some recent articles on clinical data management examine the challenges and opportunities of establishing registries to capture patient data electronically at the point of care. In particular, a review by Bommakanti *et al.* [76] highlights the barriers and enablers in implementing and managing electronic registries in LMIC-based hospitals to record cases of trauma. Barriers include limited financial resources, preventing from hiring personnel (whose role would include keeping oversight of the clinical data management system), and from buying or sustaining appropriate material (including, renewable licences for database software). Despite these barriers, the authors conclude that electronic registries provide an opportunity to optimise the use of limited human and technological resources in clinical settings, hence compensating for general administrative and organisational issues in hospitals [76].

Systems for direct, electronic recording of patient data however remain rare in LMICs, and much of the routine data that needs centralising at the national level originates from paper registries. On top of their healthcare delivery duties, healthworkers must regularly fill forms, reporting the number of different disease cases at their facility in a more or less aggregated manner (by age group, by sex, etc.), depending on the requirements of those requesting health information [73]. The process results in work overload for local healthworkers in LMICs, where overall shortages in medical staff are frequent [77], and where health facility managers often juggle multiple responsibilities [78]. **Human Resources** (**HR**) management and supervision are thus frequently raised as critical factors in the performance of HIS

and in the quality of routine data [79–81]. Beyond issues at the data collection point, the HIS literature tells us little about the challenges and opportunities in clinical data management 'processing', or about strategies to build robust data systems [82], which are of greater interest to this thesis.

Capturing health data through population surveys

HIS also draw on longitudinal data collected through regular population surveys [73, 74], whose individual-level data underlies **Health and Demography Surveillance Systems** (**HDSS**). The **International Network for the Demographic Evaluation of Populations and their Health** (**INDEPTH**) has played a pioneering role in developing resources and building the capacity for HDSS data management [83]. As part of this effort, INDEPTH has developed different data management procedures, which, although they have limitations for direct application in clinical research, can be adapted towards this use [84].

In recent years, HDSS have leveraged on the increasing availability of mobile phones and other hand-held devices to collect and report data, including from remote, rural locations. Direct **Electronic Data Capture** (**EDC**) assisted by these technologies is a hot topic in relation to data collection in LMICs (including beyond the case of HDSS); and several authors have reviewed the challenges and opportunities that their use entails [85–87]. The body of literature on EDC emphasises the advantages of this technique for end-data quality, as well as its high potential for prompt data collection and reporting [85, 87–91]. Challenges relate to the variable

reliability of data collection tools and **Information and Communication Technology** (**ICT**) infrastructure for data transmission, whose associated cost can also threaten the sustainability or scale-up of projects [85, 86]. A further barrier relates to the skills needed, whether for health and field workers little accustomed to new technologies [86, 87], or for managers programming the EDC tool [89, 92].

Building on the overall success of EDC and mobile technologies for health data collection and reporting, clinical investigators in LMICs are starting to adopt such systems for research studies, although it remains infrequent so far [93, 94].

Research and capacity strengthening efforts in HIS

Data management capacity strengthening elements are often incorporated in implementation science or operational research initiatives, whereby individuals involved in HIS receive training on developing a pragmatic research project, generally making use of the health data they readily have access to. While data management importantly becomes part of such training, the latter is primarily oriented towards raising awareness of research, encouraging the use of data in health policymaking, and improving data analysis skills [82, 95, 96].

1.4.2 Data centralisation, harmonisation and curation: managing research data for sharing and re-use

The literature on research data management reflects a particular interest in the late phases of sharing and long-term archiving of data. Perrier *et al.* [97] performed a systematic review on research data management (without restriction of discipline or country), which supports quantification of this observation: 68% of the 301 articles (and 10 "companion reports") identified by the authors focus on the 'giving access to data' step of the UK Data Archive's research data lifecycle introduced above (Figure 1.3). Of note, only nine articles in the same systematic review were found to report studies performed in Africa, Asia or South America; which emphasises the lack of past work on the case of LMICs and tropical countries.

Some organisations thus explored and reported some challenges in bridging the data management capacity gap, but mostly within the context of establishing systems for data centralisation and sharing [3, 29, 98]. As an example in a LMIC setting, the **Mahidol Oxford Tropical Medicine Research Unit** (**MORU**) highlighted the difficulties in hiring staff that it faced when setting up a centralised clinical research data management unit in Thailand [98]. Cheah *et. al.* explains that data managers are rare in the country, and they relate this HR gap to a failure from past (foreign) trial sponsors to involve Thai staff in core data management duties [98].

In the UK, the Wellcome Trust commissioned a series of reports, including one reviewing existing opportunities to strengthen the data management and program-

ming skills required for data sharing and open science [3]. This report lists multiple courses and ongoing partnerships for UK (and European) research staff, pointing notably towards the listing provided by **FOSTER** (**Fostering the practical implementation of Open Science in Horizon 2020 and beyond** [99]). CDISC also makes a move towards providing universal guidance on data management for clinical research [100], with a primary objective to harmonise data formats across studies; however, training on and implementation of CDISC processes can be complex and costly. Such data management capacity-building opportunities risk being difficult to access for LMIC-based researchers, due to the geographical location of courses (if face-to-face), and/or to their high cost.

Structural issues remain with role definitions and career paths in research data management [3, 101]; however, an interesting perspective from the open science movement and literature on data sharing is its more 'mature' thinking on the professionalisation of different 'data-intensive roles', from IT staff and librarians up to data scientists [3, 102]. Detailed competency frameworks exist for the latter group, such as the **Education for Data Intensive Science to Open New science frontiers** (**EDISON**) framework, built in consideration of previous ones [103, 104] and inclusive of a 'data management' competency area. Such initiatives and similar reflections on the desirable 'data competencies' for researchers in the context of an evolving research landscape [105] may help to raise the profile of the data science profession, with potential benefits for the similarly indistinct clinical research data management one.

1.4.3 Guidelines: defining 'good' practice *versus* strengthening the local capacity

Data management (and quality) is an area of active discussion in the computer and information sciences; however, in line with the interests of this field, the discussion and proposed solutions tend to be quite technical. This body of literature may be helpful to a specialised audience, but it is unlikely to significantly help the starting clinical research team, especially in LMICs where access to technologies and a skilled workforce can be an issue [30, 31, 106].

Several publications provide useful and detailed accounts of the procedures and ICT systems they put in place in specific research studies, including mostly in clinical trials [93, 107–110], but also in effectiveness [111] and observational studies [44, 112, 113]. Similarly, some authors describe the development and testing of new data management tools, and jointly discuss pilot applications in different clinical research contexts [114]. Of note, important requirements of data management solutions for clinical research that did not arise in the HIS literature include the necessity for clinical research databases to provide reminders for scheduled patient visits, and to allow integration of data from different sources (for example, X-ray scans or laboratory data in addition to clinical assessment data) [44, 111, 114]. Another important consideration for multi-country trials cited by Watson *et al.* for the PERCH study, which involved seven countries, was the necessity to balance harmonising study data management procedures with adaptation to local settings and different pre-existing systems [44].

While some of those articles also remain quite technical for a beginner-level audience in need of capacity strengthening, they nonetheless provide useful, real examples of applying 'good practices' in clinical research data management, and are essential as they contribute to disseminating such practices and tools with demonstrated value and applicability. A further limitation in terms of capacity-strengthening impact is that published procedures often underlie studies led by actors in HICs, hence their one-off implementation is unlikely to strengthen local capacity for data management in a sustainable manner [63], especially as core data management activities (data validation) are often performed outside of the local LMIC context [115].

The **Society of Clinical Research Data Management** (**SCDM**) has developed extensive **Good Clinical Data Management Practices** (**GCDMP**©) [116]. Interestingly, membership payment used to be necessary for access to these guidelines, but their download is now free, after registration on the SCDM website, which should promote their wider circulation. Guidance for data management is also embedded within the ICH-**Good Clinical Practice** (**GCP**) [46], but strictly speaking the latter apply solely to clinical, interventional trials. Moreover, those ICH-GCP were developed with 'Northern' research contexts in mind, and can be difficult to transfer to LMIC contexts [43, 48].

Following a baseline survey of research data centres in Europe [117], the **European Clinical Research Infrastructure Network** (**ECRIN**) developed and further revised extensive guidelines for clinical research data management [118,119].

The ECRIN guidelines provide a comprehensive description of the organisational requirements that must be satisfied to guarantee good data management, and are a notable exception in their focus on the institutional capacity for clinical research data management. The implementation of a certification scheme for ECRIN data centres has contributed to improving clinical research data management practices across Europe [120]. Nonetheless, the ECRIN guidelines assume the potential availability of fully-fledged infrastructure, and fall short of considering how data centres located within more resource-constrained settings should work towards applying the same standards as HICs.

1.4.4 Moving forward the capacity for clinical research data management in LMICs

The literature review presented throughout section 1.4 focussed on the current trends in data management capacity development across different but related fields. Overall, it demonstrated the limited availability of training and guidelines that are tailored to the reality of performing clinical research in low-resource settings. Indeed, the challenges and opportunities of data management in LMICs have been partly explored, but mostly in the context of HIS or data sharing. Major issues mentioned in this body of the literature are rather typical of LMIC settings, and primarily relate to shortages in staff and data management skills, as well as to limitations of the ICT infrastructure [29–31, 106, 121].

In contrast, the barriers and enablers of good data management in LMICs remain unknown as far as clinical research is concerned. The variability in data management procedures underlying clinical studies on diseases relevant to LMICs has not been explored in a systematic manner, thus providing limited understanding of 'who does what and why' in that context. This gap in knowledge makes it difficult to know where to start with clinical research data management capacity strengthening in LMICs. Initiatives should address the specific capacity development needs of individuals and institutions, while taking account of their wider research environment [70], whose imposed constraints must be clearly understood. Using a mixed-methods study design described in the next chapter, I set out to address this important gap, and to understand the perspectives from local researchers on the barriers and facilitators they typically encounter in handling clinical research data.

Chapter 2

Inquiry strategy: the value of mixed methods and systems thinking to characterise data management

In the introduction chapter, I gave the rationale for the proposed project and I reviewed past work on similar topics. I now present the methodology and hypotheses that guided my own investigation of clinical research data management, and I clarify my 'worldview' (or research paradigm). A research paradigm is a broad theoretical framework that coherently links the ontological, epistemological and methodological assumptions underlying a study [122]. It is characterised by its specific understanding of what reality is (ontology), as well as its accepted ways to access such reality

(epistemology) and associated means to acquire knowledge (methodology) [123,124]. The present doctoral work was driven by the will to address a concrete, yet complex, 'real world' issue that is determined by social factors among others. As such, the study had a strong pragmatic orientation, especially regarding the choice of methods. Yet, my worldview remains strongly post-positivist [122], which greatly influenced the study planning, conduct and subsequent interpretation of findings, as I explain below.

2.1 Researcher's positioning: A pragmatic perspective with a post-positivist influence

I acknowledge a preference for and thus potential bias towards the use of quantitative methods, owing to initial training in natural sciences and to a mostly post-positivist worldview [123, 124]. Within this paradigm, deductive thinking prevails. Theory and hypotheses precede data generation, with a fixed research protocol established prior to the start of evidence gathering. Data may disprove or support the initial set of hypotheses, but formally a hypothesis cannot be 'proven' or ascertained as true. With the underlying assumption that mechanisms occurring in the natural world are deterministic, post-positivism focuses on understanding cause-effect relationships between a few variables. This requires carefully controlling for the confounding effect of other factors, thus a reductionist approach to the problem under study.

Despite the elegance of experimental and statistical methods traditionally associated with the post-positivist paradigm, I recognise their limitations for testing hypotheses in the 'real world', where varied and difficult to control factors impact on the investigated issue [124]. Aragón [125] qualifies such problems as 'soft', by contrast with 'hard' ones, for which a solution can be 'engineered'. In addition, designing an experiment requires sufficient prior knowledge of the concept under study, so that clear hypotheses can be outlined and experimentally tested. Just like the practitioner planning a social intervention has to "see 'enough' and understand 'enough' to make sense of [the] world such that [they] can act meaningfully and purposefully within it" [126], the researcher must know enough about their problem of interest before they can experiment. The literature review (section 1.4) highlighted the lack of theoretical understanding of clinical research data management practices in LMICs. In that context, a systematic, exploratory mapping of the problem and an inductive approach seem more likely to contribute to the current state of knowledge than an explanatory and deductive approach.

The present doctoral work is therefore positioned within the pragmatic paradigm [122, 123]: the applied nature of the question, the relative absence of past work to build upon (especially at the onset of the study in October 2014), as well as the importance of providing a practical solution to the problem raised, altogether justify the use of methods that go beyond what the initially favoured post-positivist paradigm allows. I recognise that highly-variable individual and socio-political contexts will strongly affect the functioning of clinical research data management systems; and these intricate contextual factors need to be examined through the study

of specific, relevant cases, with a depth that the sole use of quantitative methods would hardly allow. Mixed methods [123, 127] (outlined in section 2.3 below) thus underlie the present thesis: I collected and combined quantitative and qualitative data using different techniques, to gain a most complete understanding of clinical research data management practices in LMICs, and of their associated challenges and opportunities.

2.2 Applying a 'complex systems' lens to data management

One original contribution of this thesis is to propose an examination of clinical research data management as a 'complex system', or "an interconnected set of elements that is coherently organised in a way that achieves something" [128]. With this definition, Meadows [128] emphasises the importance of acknowledging that a system is characterised by its "elements", but also by the "interconnections" between these, and by a "function or purpose". A systems' lens helps to broaden the view on a concept (here, data management) from a rather mechanistic and linear perspective to a more holistic one, thus uncovering the influence of interrelationships with other concepts on the desired outcome(s) [126].

Clinical research data management includes a whole range of activities or "tasks", to quote the terminology put forward by CDISC [1]. The **Association for Data Managers in the Tropics** (**ADMIT**) expands the view of data management be-

yond a set of tasks, by proposing a framework and clinical trial data lifecycle that not only shows data management activities, but also the documents and policies supporting or constraining them, as well as the groups of actors involved in the process [129] (Figure 2.1). Similarly, Smith *et al.* [130] describe the components that need to be brought together when planning data management for clinical research (namely; hardware, software and personnel), and also explain that some organisational structure to bring together and oversee these is essential.

Together with Meadows' description of 'complex systems' these depictions of clinical research data management inspired the initial framework of thought and hypotheses guiding this doctoral work, which I graphically outline in Figure 2.2.

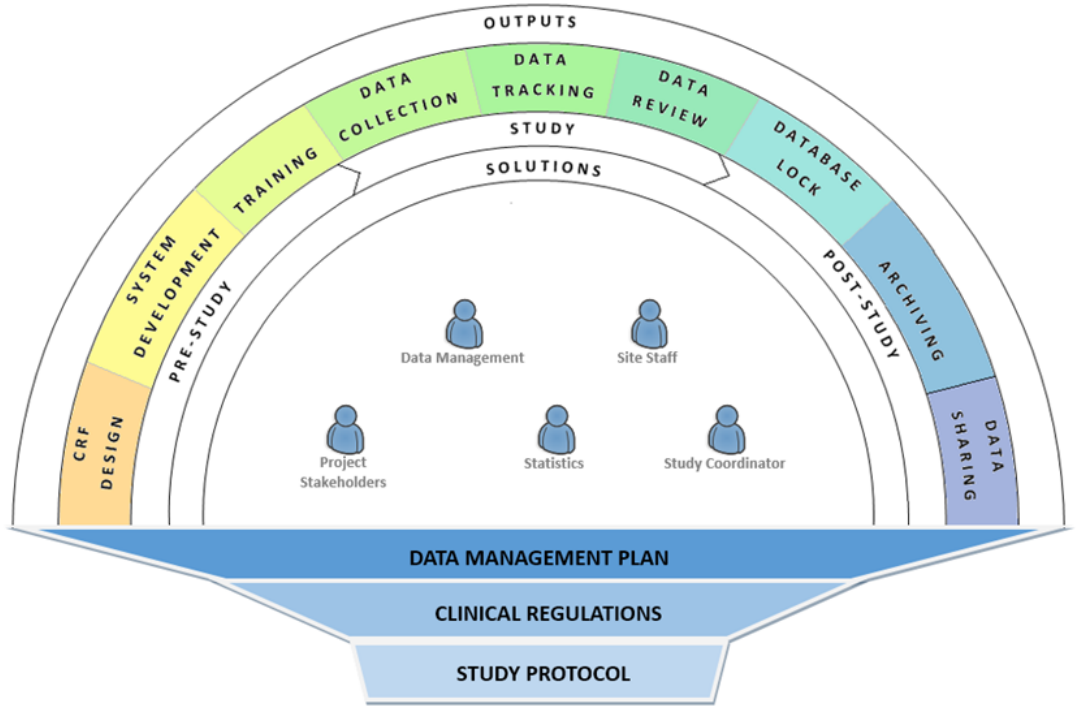


Figure 2.1: **The Association for Data Managers in the Tropics (ADMIT) research data lifecycle** [129].

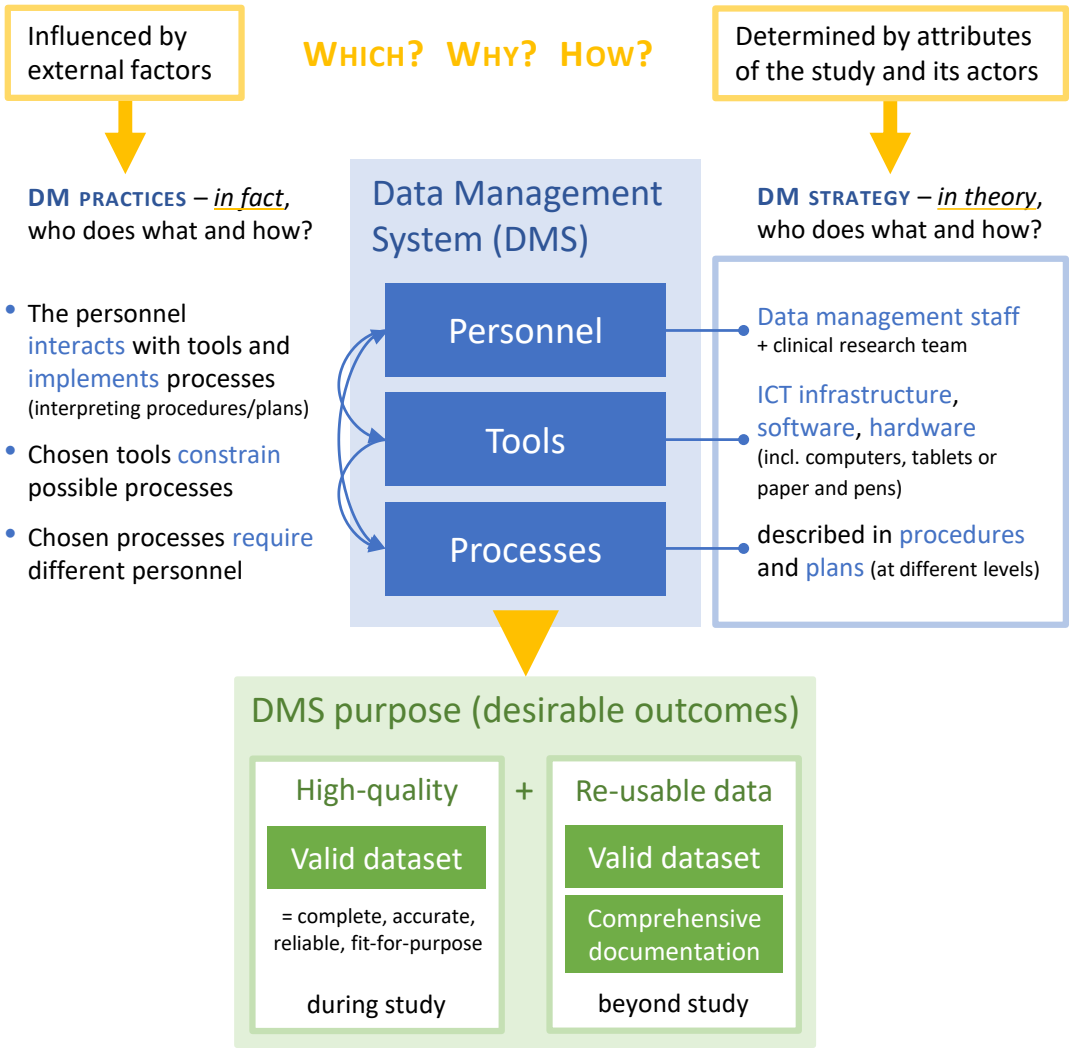


Figure 2.2: **Key components and outcomes of a Data Management System (DMS), and potential external factors influencing those**. An earlier version of my understanding and graphical representation of a DMS is included in Appendix B.

Throughout this thesis, a clinical research **Data Management System** (**DMS**) is defined by its purpose to provide high-quality data during and beyond the study lifecycle. Ideally, and within the context of data sharing, the study data obtained as a result of operating a DMS should also be re-usable, which requires as outputs a valid dataset, but also comprehensive documentation [36, 131]. A clinical research

DMS with the consistent capability to deliver a valid dataset for analysis is qualified as robust; or in other words, its functioning is resilient enough to perturbations to ensure the system's purpose is fulfilled as desired. This may necessitate to build in some flexibility within the DMS structure, for example to accommodate changes in available technological tools, as emphasised by some investigators describing their clinical study data management procedures [111, 113].

While the "purpose" of a clinical research DMS is clear, much of the inductive work performed in this doctoral thesis seeks to clarify the less evident influential "elements" and "interconnections" that make up this system, as well as the factors that may determine or perturb its functioning.

2.3 A globally inductive approach based on mixed methods

Three interconnected pieces make up this doctoral thesis (Figure 2.3). Each piece relies on different methods and datasets, whether primarily quantitative (Part II) or qualitative (Part III). Data collection and analysis occurred independently for the three pieces. The primary quantitative piece of work was conducted first, and the second one was performed concurrently with the qualitative piece. Results were eventually reconsidered and interpreted in view of each other (Part IV), thus making the overall study design a convergent, mixed-methods one [123].

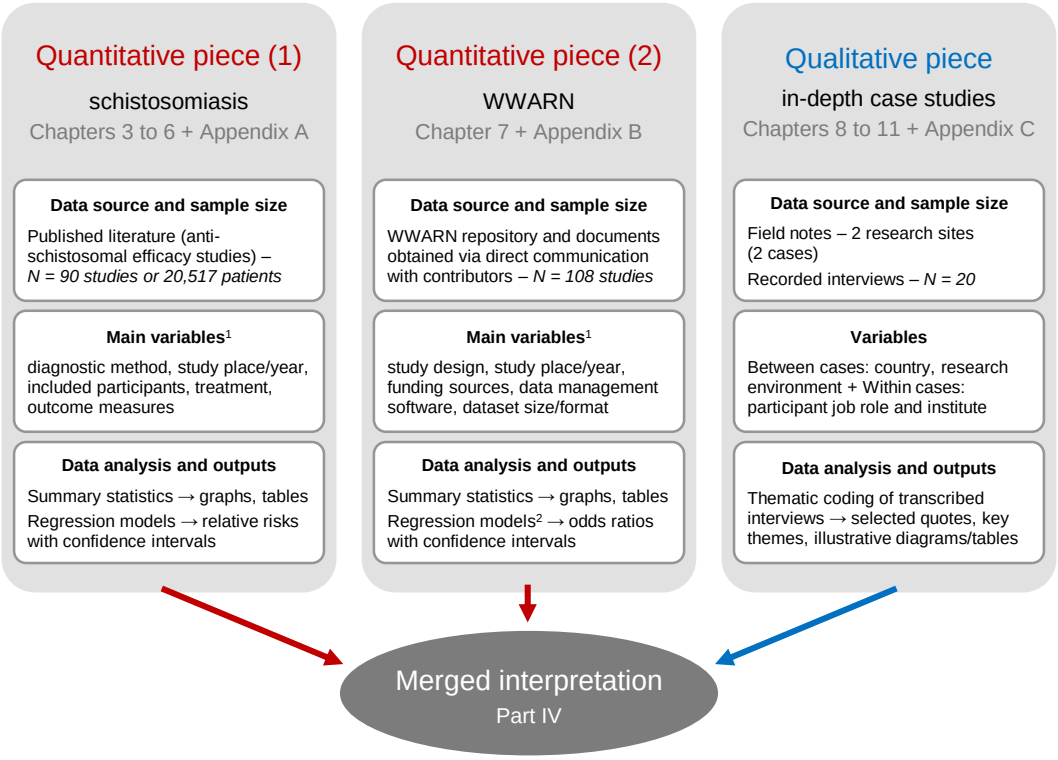


Figure 2.3: **Mixed methods convergent design.** The piece of quantitative work focussing on schistosomiasis was conducted first, while the two subsequent pieces on malaria largely overlapped in their planning and execution timeframe. All analyses were performed independently, and merging happened at the overall interpretative stage. WWARN: WorldWide Antimalarial Resistance Network; [1] See Appendix A and B for a full list of variables underlying quantitative piece (1) and (2), respectively; [2] Initially planned but not performed, in light of results from preliminary statistical analyses.

Taken together, the three pieces contribute to an inductive and holistic understanding of the multiple components at stake in the planning and implementation of clinical research data management in LMICs. Considered in isolation, each piece enabled me to investigate specific "elements" and/or "interconnections" of clinical research DMS in greater depth.

The first piece of analysis underscores the challenges of harmonising aggregated (published) data from clinical studies on a NTD mostly researched by investigators

based in LMICs (schistosomiasis) [59]. I particularly question the fitness-for-purpose and re-usability of data as accessible from the literature, to determine the value of establishing a data platform and of performing pooled analyses using **Individual Participant Data** (**IPD**) rather than aggregated population-level measures. I also reflect on the data management (and analysis) processes that led to the delivery of such data, and discuss the implications of my work for DMS and data sharing in schistosomiasis research. This piece of the thesis strictly draws on quantitative methods, using a mix of descriptive and regression statistics to shed light on heterogeneous data. Chapters 3 to 6 respectively cover the introduction, methods, results and preliminary discussion related to this work.

The second quantitative piece (Chapter 7) characterises the variety of data management strategies within a community of malaria researchers who have engaged in data sharing through the **WorldWide Antimalarial Resistance Network** (**WWARN**). In this work, I also explore associations between study characteristics and data management indicators, seeking to understand which factors influence the chosen clinical research data management processes and tools. This piece is mainly quantitative, based on a relatively large sample and (initially planned) logistic regressions to identify trends; however, I also drew on qualitative methods in the design of the data collection instrument (as described in Appendix B).

The third analysis (Part III) focuses on two specific sites in distinct LMICs, to complement observations from the quantitative work and explore additional (social) challenges and opportunities in clinical research data management for local investigators. The perceptions, experience, knowledge and attitudes of study staff are

indeed expected to impact on data management practices, as the recognition of what planning and running a DMS require may vary from one individual to the next. The latter piece is exclusively qualitative, using an 'ethnographic approach' and mostly relying on in-depth interviews with clinical research staff in varied roles.

Detailed methods are provided for each piece as I turn to its reporting, together with further background on the specific 'cases' used – whether disease (schistosomiasis and malaria), international network (WWARN) and research institutes (in Dakar and Brazzaville), or geographical area (Africa, especially Senegal and the Congo).

Part II

Quantitative work

Chapter 3

Introduction to the scoping review and meta-analysis of antischistosomal trials

Data sharing proponents often make the assumption that sharing all data and any data is a good thing; however, such an unspecific approach may drive the creation of "data dumpsters" [131], where disparate, hard to discover, and poorly documented data is piling up [10,13]. Not all ways of sharing research data are equally meaningful, and without appropriate operating procedures for both initial study DMS and final repositories, much of the shared research data risks becoming unusable [131]. Furthermore, not all data may be in demand for re-use, and some have suggested that clinical research data sharing should be driven by a will to answer specific questions that single datasets cannot address [11, 132]. In the case of health research,

repositories dedicated to a specific disease can help to address issues of discoverability and disparateness [131]. Creating meaningful data-sharing platforms nonetheless requires a strong commitment to exhaustively gather datasets regarding the target disease, and to systematically curate incoming data. To avoid wasting resources, it is therefore crucial to assess the feasibility and expected benefits of a disease-specific data platform before launching it.

As part of an initiative started by the WHO and its **Special Programme on Research and Training in Tropical Diseases** (**TDR**) [133], since pursued by the **Infectious Diseases Data Observatory** (**IDDO**), I investigated the potential value of developing a schistosomiasis treatment efficacy data platform. Schistosomiasis constitutes an excellent example of NTD, and the initial interest from WHO and TDR to support data management and sharing of clinical research data within the schistosomiasis community enhanced the impact and immediate relevance of focussing on this disease.

In this first chapter regarding the schistosomiasis piece of work, I provide some background on human schistosomiasis and its current control strategy, and I state the specific objectives pursued in this quantitative piece. The biological and epidemiological aspects reviewed here are necessary to understand the landscape of schistosomiasis research and inherent variations in study methods or data analysis approaches. Areas of variability in antischistosomal efficacy trials will be extensively discussed in upcoming chapters, in their relation to the parasite's biology and existing diagnostic tools introduced here, as well as to investigators' methodological and reporting choices.

3.1 Schistosomiasis: biological, clinical and epidemiological facts

Schistosomiasis (also known as bilharzia) is a parasitic disease that can affect humans and other primates, but also cattle, pigs, dogs and rodents [134]. The causative agents are trematode worms of the *Schistosoma* genus. Three species (*S. mansoni*, *S. japonicum* and *S. haematobium*) are responsible for most of the human disease burden [134–136]. In this thesis, the emphasis is on the two species striving in Sub-Saharan Africa, that is to say *S. mansoni* and *S. haematobium*.

3.1.1 *Schistosoma* lifecycle

Schistosoma parasites transit through two different hosts (human or other mammalians, and snails) to complete their lifecycle (Figure 3.1).

Microscopic cercariae released in water can penetrate the mammalian host's skin [134, 135]. Within the human body, they turn into schistosomulae (immature schistosomes). The latter migrate to the liver, where they become adults of seven to twenty millimetres in four to six weeks [134, 135]. Male-female pairs then reach their tissue of residence, where they can survive and reproduce for years [134]. *S. mansoni* and *S. japonicum* mature worms are mostly found in mesenteric blood vessels, whereas *S. haematobium* ones reside in veins surrounding the bladder [134, 135].

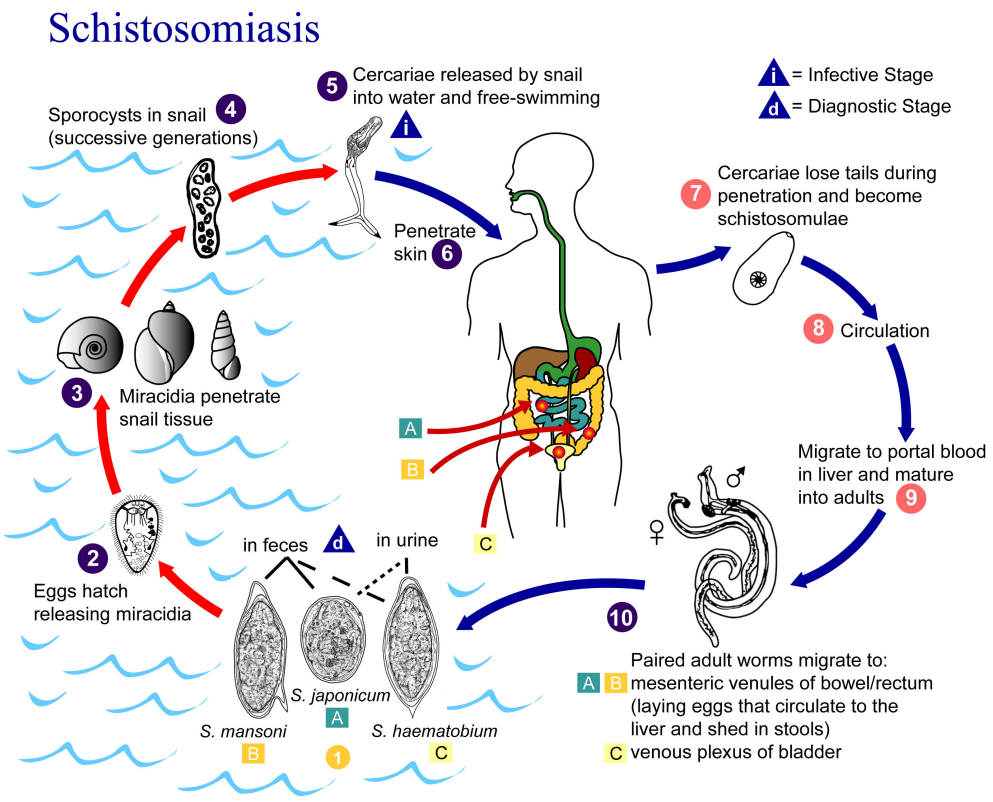


Figure 3.1: **Schistosomiasis lifecycle**. Reproduced from the USA Centers for Disease Control and Prevention (CDC) [135].

Schistosoma females shed eggs in large amounts (up to thousands a day per worm [134]), which are released in the intestine (*S. mansoni* and *S. japonicum*) or urinary tract (*S. haematobium*). When excreted and in contact with water, eggs hatch and release miracidiae, the parasite's form able to penetrate the second, intermediate host (snails) [134, 135]. Within the snail, miracidiae divide asexually to produce sporocysts, which further divide into cercariae, thereby completing the parasite's lifecycle [134].

3.1.2 Physiopathology and diagnosis

Eggs are responsible for most of the pathology, as they get trapped and trigger chronic inflammatory responses in tissues around the intestine, urinary or reproductive tract [134]. Long-lasting infections lead to fibrosis of affected tissues, with debilitating consequences and potentially death [134].

Schistosomiasis is usually diagnosed by counting eggs in excreta – urine or stool, depending on the infective species and its corresponding resident tissue. Counting eggs under microscopy is a diagnostic test of limited sensitivity, even when used jointly with concentration techniques before preparing microscopy slides [134]. The chosen number of independent samples taken per participant and of slides prepared per sample further affect sensitivity, and it represents a major source of variability in diagnostic approach across schistosomiasis studies [137]. Other diagnostic methods exist, including some showing greater sensitivity such as rectal snips or antibody detection assays. However, their use as diagnostic tools in the field is limited, owing to greater invasiveness of the clinical procedure or to cross-reactivity of the test with other worms, which also abound in many schistosomiasis-endemic areas [134].

3.1.3 Disease spread and burden

Natural habitats of the disease vector (snail) determine areas where schistosomiasis is endemic, together with the frequency and patterns of human contact with water: *S. japonicum* is present in focal areas of China and South-East Asia; *S. mansoni*

and *S. haematobium* are both found across Sub-Saharan Africa; and Brazil is solely affected by *S. mansoni* [134]. **School-Aged Children** (**SAC**) are most affected by schistosomiasis, and there is considerable individual variability in worm burden, which has been attributed to different immune responses and immunological control of the pathogen among individuals [134]. The latest global burden of disease study estimates that over 290 million are affected by schistosomiasis worldwide, corresponding to a loss of 3.06 million disability-adjusted life years (data for 2013 [138]).

3.2 Preventive chemotherapy and the threat of drug resistance

3.2.1 Therapeutic options

Praziquantel is effective against all species of schistosomes, and has been the drug of choice for schistosomiasis treatment since the 1970s [134]. The drug has a good efficacy and safety profile at the recommended regimen of 40 milligramme per kilogramme of body weight (mg/kg), as supported by evidence from several clinical trials and (aggregated data) meta-analyses [139, 140]. Praziquantel is active against adult worms, but not so much against immature ones or eggs [134]; hence *Schistosoma* eggs may still be excreted for weeks after treatment. In highly endemic settings, new infections may be acquired during that timeframe, thus leading to *de novo* egg excretion. As schistosomiasis diagnosis traditionally relies on egg counts,

the ideal timepoint to assess cure is debated, especially in the context of antischistosomal efficacy trials [134,141]. Capturing the 'true' efficacy of a drug indeed requires balancing the risk of observing resolving infections with the risk of observing newly acquired ones.

Oxamniquine is also a potent antischistosomal drug, but it is only active against *S. mansoni* and causes more undesirable side effects than praziquantel [134]. Finally, artemisinins show efficacy against immature forms of schistosomes [134]; however, preserving their antimalarial efficacy is critical, and their use is hardly possible in malaria-endemic areas, which greatly overlap with schistosomiasis-endemic ones.

3.2.2 Control strategy

Schistosomiasis control largely relies on 'preventive chemotherapy' [142, 143]. This strategy consists in the distribution of a single drug (praziquantel) at regular intervals (twice or once yearly, or every second year) to large numbers of individuals living in endemic areas [142]. WHO's strategic aim is to control morbidity by 2020, and eliminate schistosomiasis as a public health problem by 2025 [143]. In 2016 alone, 89.2 million people received preventive chemotherapy with praziquantel, and those were mainly schoolchildren in Africa (68.5 million) [144]. This is a substantial number, yet still far from the intended programme target of 206.4 million requiring preventive chemotherapy across 52 countries (of which, 42 African ones) [144].

Population-based therapy relying on a safe and effective drug has indeed shown some success [145, 146]; however, this strategy is not without risk of emerging drug

resistance [147]. Genetically-acquired resistance to drugs used against other worms (helminths) has been reported in cattle [148–150]; and resistance to praziquantel has been reported in a few studies in natural endemic settings [151–153]. As hardly any alternative treatment for schistosomiasis exists, it is imperative that responses to praziquantel be carefully monitored [154, 155].

3.3 Specific objectives of the first quantitative piece

The schistosomiasis research landscaping exercise contributes to addressing the first two aims set out for this thesis, because it explores the variability in data management (and other study) procedures across antischistosomal trials, and questions their impact on the re-usability of resulting aggregated therapeutic efficacy data. As mentioned above, this first quantitative piece took place within the context of a wider feasibility assessment of an antischistosomal IPD sharing platform, and its overall objectives were thus as follows:

I. To estimate the number of participants included in recent antischistosomal efficacy trials – that is to say, to estimate the global amount of IPD that could be retrieved and gathered within a dedicated data-sharing platform;

II. To characterise the variability in trial methods and resulting antischistosomal efficacy data, and the extent to which this variability may hamper data harmonisation and re-use;

III. To determine the added scientific value of collating IPD (disaggregated data), by seeking to address a question of clinical and public health relevance, using population averages reported in articles (aggregated data).

Chapter 4

Methods: Systematic approach to scope the amount and variety of existing drug efficacy data

This chapter outlines the protocol developed to meet the objectives stated above, and especially to estimate the volume and to evaluate the homogeneity of IPD arising from past schistosomiasis treatment efficacy studies. Such a 'landscaping exercise' primarily relies on a systematic review (here complemented by a meta-analysis), and is particularly useful to determine the value and feasibility of creating disease-specific data platforms. Since I performed this exercise for schistosomiasis, similar landscaping reviews focussing on visceral leishmaniasis [156] and on **Soil-Transmitted Helminthiases** (**STHs**) [157] have also been published, thus highlighting the merit of this approach.

Of note, the rather lengthy methodological procedures and assumptions detailed in Chapter 4 reflect the challenges encountered in performing the work, and already suggest the important limitations of aggregated data in schistosomiasis research.

4.1 Identification of relevant references

4.1.1 Literature search

The systematic review started with electronic searches for bibliographic references in the Cochrane Library, PubMed, CENTRAL and Embase. Keywords described in Figure 4.1 were used and combined as detailed in the next paragraph. No restriction on publication language was applied, but search results were filtered for 'human' as the species of study whenever possible. Additional references were identified through manual examination of bibliographic lists from past systematic reviews of antischistosomal trials [137, 139, 140, 158, 159].

Four independent searches were run in May-June 2015 by an information specialist from the Cochrane Infectious Diseases Group. The first three each focussed on one major pathogenic *Schistosoma* species, thus combining keywords (1.2., 1.3. or 1.4.) AND (2) AND (3.1.) from Figure 4.1. The searches for literature on *S. haematobium* and *S. mansoni* were updates of previously published Cochrane systematic reviews [139, 140], and were therefore limited to publications from 2000 up to the date of search (25 May 2015). The *S. japonicum* search used no such cut-off on publication date, as no Cochrane search pre-existed.

Theme 1 – **disease**: Schistosomiasis

(1.1.) schisto*

(1.2.) 'schistosoma hematobium' OR 'schistosomiasis haematobia' OR 'schistosom* haematobi*' OR 'urinary schistosom*' OR 'urogenital schistosom*'

(1.3.) 'schistosoma japonicum' OR 'schistosomiasis japonica' OR 'schistosom* japonic*' OR 'oriental schistosom*'

(1.4.) 'schistosoma mansoni' OR 'schistosomiasis mansoni' OR 'schistosom* mansoni' OR 'intestinal schistosom*' OR 'bilharz*'

Theme 2 – **study design**: Clinical (any)

(2) treatment OR trial OR random* OR double-blind* OR cohort OR prospective

Theme 3 – **drugs**: Antischistosomal (known or suspected)

(3.1.) praziquantel OR artesunate OR artemether OR metrifonate OR albendazole

(3.2.) oxamniquine OR mefloquine

Figure 4.1: **Search keywords** belonged to three themes reflecting the purpose of identifying schistosomiasis treatment efficacy studies (focussing on any drug and using any design). Albendazole has not shown efficacy against *Schistosoma*, but it was included in the search terms to retrieve relevant studies on STHs, which often assess intestinal schistosomiasis treatment efficacy aside of their major study objective.

The fourth and final search looked jointly at all species but focussed on different drugs that had been overlooked in initial searches, therefore combining keywords (1.1.) AND (2) AND (3.2.) from Figure 4.1. In addition, this later search was restricted to abstracts published since 2001, owing to the decision made in the meantime to restrict the review scope to more recent studies only. Indeed, when collecting datasets for secondary analyses, the likelihood of obtaining those dating back before 2000 is very limited [160]. It may therefore be misleading to consider older studies when estimating the volume of IPD that could contribute to a global database, which was one of the primary objectives pursued here.

As a first screening step to discard studies completed before 1 January 2000, the list of references compiled after completion of all (manual and electronic) searches

was shortened to cover only articles published since 1 January 2001. Duplicate references were also deleted at this stage using built-in functionalities of the bibliographic references' manager software Zotero [161]. The resulting references' list was marked as final and can be consulted at `https://github.com/Amelie-TGHN/schistoDB/blob/master/sourceDatasets/sDB_0search.csv`

4.1.2 Eligibility screening

Abstracts and, whenever necessary, full-text articles corresponding to the identified references were assessed against pre-defined eligibility criteria. Two consecutive assessments occurred, in June-August 2015 and September-December 2016 for the initial quantitative synthesis and for the subsequent meta-analysis, respectively. Eligibility criteria are outlined below, along with the rationale for their formulation; and further details on the two-step screening process are provided in Appendix A.1.

First set of criteria – applicable to the quantitative synthesis and subsequent meta-analysis

Abstracts of conferences run before 1 January 2014 were excluded upfront, to avoid duplicate consideration of one same study. Indeed, results presented at conferences may be published a couple of years later, and in such cases, the bibliographic search is likely to capture both conference abstracts and full publications. Case reports/series and non-primary research studies (meta-analyses, reviews, textbook chapters, opinion papers and comments) were also screened out.

Remaining studies were carefully examined to confirm that: (i) study drugs with (known or suspected) antischistosomal activity were administered to human subjects in an endemic setting; (ii) all or a subset of participants were positively diagnosed for infection with at least one of the three *Schistosoma* species of interest, before they received treatment; and (iii) a post-treatment diagnostic test was performed on the same participants to assess outcome within 60 days, where day 0 is the day of administration of the first (and in most cases, single) treatment dose.

Second set of criteria – applicable to the meta-analysis only

Conference abstracts were excluded altogether from the meta-analysis, owing to the limited contextualisation of study results that this short reporting format allows. In addition to the above, articles had to meet the following, more stringent criteria to be considered in the second part of the landscaping exercise: (i) have administered praziquantel at the single dose of 40 mg/kg; (ii) have assessed outcome at 14 to 60 days post-treatment; (iii) include schoolchildren (regardless of the actual age range they covered) or some participants aged 5-19 years (expected ages of primary schoolchildren in most countries); (iv) report at least one measure of drug efficacy as cure rate or egg reduction rate for a group comprising only the targeted ages.

For the purpose of preventive chemotherapy, WHO defines its target group of SAC as those children aged 5 to 14 years [162]. However, the experience of reading articles for the quantitative synthesis suggested that drug efficacy studies performed in schools enrol pupils within a wider age-range; including children from 4 up to over

20 years old. Because aggregated data does not enable selection for participants falling within the real target group, I adopted an extended, data-driven definition of SAC, to include individuals of 5 to 19 years old. Articles which did not report age ranges but specified that study recruitment happened at schools were also retained and their participants considered as SAC.

Notes on follow-up timepoints

When performing anthelmintic efficacy studies, WHO recommends assessing therapeutic outcomes within 14 to 21 days of treatment administration [141]. This is deemed necessary to allow some time for the drug activity to become observable using diagnostic methods relying on egg count, while still ruling out cases of early reinfection (as explained in section 3.2.1). However, the maturation time of schistosomes in humans induces a delay of at least four weeks between infection and egg excretion [163]. Within the 60 days post-treatment window, individuals who were initially cured but got reinfected are likely to excrete only low numbers of eggs (if any), and to appear negative upon diagnosis by egg count. Participants with outcomes assessed up to 60 days post-treatment (and arguably even beyond, provided appropriate statistical methods are used to account for this effect [164]) could in principle contribute to meta-analyses of antischistosomal efficacy, thus their inclusion in the count of estimated IPD.

When progressing to the aggregated data meta-analysis for this thesis, it became necessary to refine the considered window. In the vast majority of *S. haematobium*

and *S. mansoni* intervention studies, the therapeutic outcome is assessed at least once between three and six weeks post-treatment; except for studies primarily interested in evaluating diagnostic methods, where earlier measurements also occur. In view of the current WHO guidelines (assessment two to three weeks after praziquantel administration) as well as the reality of implementing antischistosomal studies in endemic settings, all trials with outcomes assessed within the 14-60 days period (inclusive) were retained in the meta-analysis. For the few studies which had available efficacy data at multiple post-treatment timepoints, the only outcome measure retained is the one at 21 days post-treatment, as this best matches WHO recommendations and was available for all such studies.

4.2 Data collection and preparation

Once eligibility had been ascertained, relevant data was extracted and recorded into a spreadsheet document (Microsoft Excel® 2000). All the necessary data was either readily extracted or indirectly derived from information provided in publications, except for the year when countries first implemented preventive chemotherapy. The latter was obtained from the WHO databank [162], or communicated by the WHO NTDs' department for countries that began preventive chemotherapy before 2006, as this is not recorded in the online databank.

Appendix A.2 and A.3 comprehensively detail all data tables, their constitutive variables and the assumptions made when publications ambiguously reported the chased information. In sections 4.2.1 to 4.2.3 below, I describe key variables which

required the most interpretation and played a major role in analyses performed to fulfil the primary objectives.

4.2.1 Estimation of participants' numbers

Some of the studies included in the landscaping exercise had a broad scope; for example, looking simultaneously at schistosomiasis and at other parasitic infections, in particular STHs. Articles for such studies often did not report the exact number of participants with a *Schistosoma* infection at baseline, thus the need to estimate it. Whenever possible, I recalculated the number of *Schistosoma*-positive participants from the specified prevalence of infection among participants screened at enrolment. If the latter was unknown, I used instead the prevalence in the study area (potentially established from pre-dating studies). Finally, in the rare case when there was no mention of prevalence at all, half of the study population was arbitrarily considered infected with any *Schistosoma* species at baseline; that is to say, I assumed a disease prevalence of 50% among enrolled participants.

For consistency and unless more specific information was available, the reported total of participants in subsets of the study population (for example, subset of participants successfully followed-up at later timepoints) was then multiplied by the same percentage of infection (exact prevalence in study population, reported prevalence in study area, or 50%). This assumes that the subset characteristic (for example, successful follow-up) and the initial infection status of the participant are independent, which is a reasonable though not necessarily correct assumption.

4.2.2 Corrections for homogenisation of Egg Reduction Rates (ERRs) and Mean Egg Counts (MECs)

The **Egg Reduction Rate** (**ERR**) is the usual population measure of antischistosomal efficacy [141], together with the cure rate. The ERR is derived from population **Mean Egg Counts** (**MECs**) at baseline and follow-up; such that $ERR = 1 - \frac{MEC_f}{MEC_b}$ (usually expressed in percent). In this formula, MEC_b and MEC_f are the average number of eggs found in the excreta of the study population as a whole, at baseline and at follow-up, respectively for MEC_b and MEC_f [141]. The cure rate is simply the proportion of participants infected at baseline and cured at a given follow-up timepoint; and it is a less robust measure of anthelmintic efficacy [165].

Both cure rate and ERR data were considered in the meta-analysis exploring spatiotemporal variations in praziquantel efficacy. For all studies providing both MEC_b and MEC_f within the target group (schoolchildren), I systematically re-calculated ERRs from reported MECs, hence some occasional differences between ERR values presented in results (section 5.3) and those reported in the underlying articles, due to rounding.

The quantitative synthesis uncovered some initial heterogeneity in ERR calculation and reporting, as this rate may be derived from arithmetic or geometric MECs [137, 166]. Closer examination as part of the meta-analysis unravelled further heterogeneity in the way MECs are calculated, which prompted me to apply some corrections, in a similar fashion as Zwang and Olliaro in a recent, similar meta-analysis [167].

Mathematical procedure

The following two main ways of calculating a **Geometric Mean Egg Count** (**GM**) were identified:

$$\begin{aligned} GM_1 &= antilog(mean(log(x_i+1))), \quad x_i \geqslant 0 \\ \Leftrightarrow log(GM_1) &= \frac{\sum_{i=1}^{N_{enrol}} log(x_i+1)}{N_{enrol}} \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} GM_0 &= antilog(mean(log(x_i))), \quad x_i > 0 \\ \Leftrightarrow log(GM_0) &= \frac{\sum_{i=1}^{N_{pos}} log(x_i)}{N_{pos}} \end{aligned} \tag{4.2}$$

where x_i is the egg count for individual i; N_{enrol} is the number of participants enrolled (infected at baseline and administered an antischistosomal as part of the study), and N_{pos} is the number of participants with a positive schistosomiasis diagnostic test by egg count (infected). Of note, at baseline $N_{enrol} = N_{pos}$, hence $GM_0 \approx GM_1$.

In contrast, at follow-up it is expected that $N_{pos} < N_{enrol}$, with the greatest difference between N_{pos} and N_{enrol} when the cure rate is high. For all cured participants $x_i = 0$ (negative diagnostic test), which means $log(x_i + 1) = 0$. Equation (4.1) is therefore equivalent to:

$$log(GM_1) = \frac{\sum_{i=1}^{N_{pos}} log(x_i+1)}{N_{enrol}} \tag{4.3}$$

By correcting the denominator of GM_0, it is possible to get GM_0^* approximately equal to GM_1. This modified follow-up MEC (GM_0^*) is expressed as:

$$log(GM_0^*) = \frac{\sum_{i=1}^{N_{pos}} log(x_i)}{N_{enrol}} \tag{4.4}$$

The value of $\sum_{i=1}^{N_{pos}} log(x_i)$ can then be derived from the known (reported) values of N_{enrol}, CR (cure rate) and GM_0 as follows:

$$\begin{aligned} (4.2) &\Leftrightarrow N_{pos} \times log(GM_0) = \sum_{i=1}^{N_{pos}} log(x_i) \\ &\Leftrightarrow (1 - \frac{CR}{100}) \times N_{enrol} \times log(GM_0) = \sum_{i=1}^{N_{pos}} log(x_i) \end{aligned} \tag{4.5}$$

In summary, for all publications reporting a MEC_f obtained from GM_0, a 'corrected' GM_0^* was recalculated as:

$$\begin{aligned} GM_0^* &= antilog(\frac{N_{pos}}{N_{enrol}} \times log(GM_0)) \\ \Leftrightarrow GM_0^* &= antilog((1 - \frac{CR}{100}) \times log(GM_0)) \end{aligned} \tag{4.6}$$

and a 'corrected' ERR was eventually recalculated using the reported MEC_b and the modified MEC_f (GM_0^*).

Of note, I identified another two methods to calculate GMs and corresponding ERRs. Some publications used $GM_{1b} = antilog(mean(log(x_i + 1)) - 1)$, $x_i \geqslant 0$ (which was not corrected for). Yet others used GM_{0b} which is the same formula as GM_0, except that the baseline GM is recalculated *a posteriori* for the subset

of participants found uncured at follow-up, before deriving the ERR. The latter amounts to excluding cured participants from the study population altogether, and was corrected in a similar fashion as GM_0.

Finally, it should be noted that when studies reported both arithmetic and geometric MECs, I only recorded and used geometric ones for the meta-analysis. This contrasts with current WHO guidelines [141], but was guided by the fact that in the studied period (mostly pre-dating these 2013 guidelines), more articles reported geometric ERRs than arithmetic ones.

4.2.3 Assignment of Global Positioning System (GPS) coordinates to study sites

The meta-analysis of spatiotemporal variations in PZQ40 efficacy required matching GPS coordinates to each measure of efficacy (cure rate and ERR) retrieved from eligible articles. Most publications mentioned the recruitment site(s) of participants, but the granularity of reported spatial information varied greatly (ranging from mention of a vast region down to a specific school), and identifying back the described place can be all the more difficult when the administrative division of the country has evolved since the study occurred.

Searches of Google Maps and use of **Global Administrative Unit Layers** (**GAUL** [168]) data enabled assignment of GPS coordinates to each site. When the

exact location could not be found on Google Maps with enough certainty, I calculated the centroid point of the smallest GAUL area that could be recognised, using the function gCentroid() from the 'rgeos' R package [169]. A categorical variable was also defined to record the precision of the assigned point estimate.

4.3 Data synthesis and analysis

Analyses were performed using R (version 3.1.3 and later) [170]. Source data tables, supporting information and analysis codes for the (published [166]) quantitative synthesis are all accessible at: `https://github.com/Amelie-TGHN/schistoDB`

The quantitative synthesis made use of descriptive statistics (counts and proportions) to characterise IPD. The meta-analysis is registered on PROSPERO (number CRD42016039648). In the latter part of the work, the predominately descriptive analytical approach was complemented by fitting a set of basic statistical regression models, separately to data on cure rate and on ERR from *S. mansoni* and *S. haematobium* studies. These regressions sought to determine whether spatio-temporal trends in praziquantel efficacy exist, which could be masked by potential confounding variables. Studies were not assessed for their risk of bias; hence the confidence in the efficacy estimate(s) provided by a study is simply reflected on figures and accounted for in models through its sample size.

To define regression models, I sequentially tested the impact of relevant covariates (year of study conduct, starting year of preventive chemotherapy, prevalence of infection at site, population MEC at baseline, time of follow-up assessment) with univariate models. I did not consider a covariate in further models when no significant effect (at the 0.05 level) was uncovered from univariate analyses. Alternately, the intention was to fit a multivariate model, keeping the year of study conduct and starting year of preventive chemotherapy (as their impact was at the core of the tested hypothesis), along with any covariate of suspected significance (confounders).

For all models, the response variable was defined as $1-$*efficacy*, where *efficacy* is either cure rate or ERR, expressed as a frequency rather than percentage. The decision to use $1-$*efficacy* was to ensure the domain of the response variance was $(0, \infty]$, since negative estimated efficacies are possible (as discussed by Walker *et al.* [171]). No specific parametric form was assumed for these response variables. They were instead defined with a variance proportional to the mean and weighted by the inverse of the participants' number in each group from which the efficacy measure was initially calculated; in other words, greater variances are linked to smaller treatment groups. This variance to mean relationship was chosen based on the observation that higher efficacies tend to be associated with lower variance [159,172]. All models were fitted using quasi-likelihood techniques.

Huber-White sandwich estimators [173] were used to estimate robust standard errors, accounting for the effects of potential correlation among efficacies measured either from the same study, or from multiple studies within the same country.

Chapter 5

Results: Existing trial data and issues with aggregated measures of antischistosomal efficacy

This chapter presents the findings from the scoping review and meta-analysis of antischistosomal efficacy trials. In the first section, I discuss the amount of studies identified and included, following guidelines from the **Preferred Reporting Items for Systematic Reviews and Meta-Analyses** (**PRISMA**) statement [174]. I then present the quantitative synthesis of all eligible studies (section 5.2), which aimed to determine whether enough participant data exists to justify the effort of creating a schistosomiasis treatment efficacy data-sharing platform, and to assess whether the variability in study procedures would hamper the collation of these IPD. Finally, I focus on a sub-sample of studies for the aggregated data meta-analysis

(section 5.3), whose 'clinical and public health' objective is to explore spatiotemporal variations in praziquantel efficacy. In the context of this thesis, the underlying aim is however to determine whether such an objective can be addressed using aggregated data, or whether access to IPD is essential.

While the findings in Chapter 5 have multiple implications for schistosomiasis research, they are of particular interest to show the relevance of creating an IPD platform dedicated to antischistosomal efficacy data. Overall, the landscaping exercise also highlights potential data management capacity strengthening needs among schistosomiasis investigators, which I mention here alongside the results, and further discuss in Chapter 6.

5.1 Eligible studies and participant groups

Searches for literature within electronic databases yielded a total of 1,268 references, and bibliographic lists of past meta-analyses [137, 139, 140, 158, 159] uncovered an additional 190 references (Figure 5.1). After screening the corresponding 914 unique references and applying the first set of eligibility criteria, 85 articles published since 2001 and five conference abstracts presented since 2014 were retained for the quantitative synthesis. Thirty-two of the 85 articles remained eligible when assessed against the second set of criteria required for consideration into the meta-analysis. The most frequent reason for exclusion at this late stage was the absence of an outcome measure reported for the targeted age group (schoolchildren), leading to the exclusion of 37 articles.

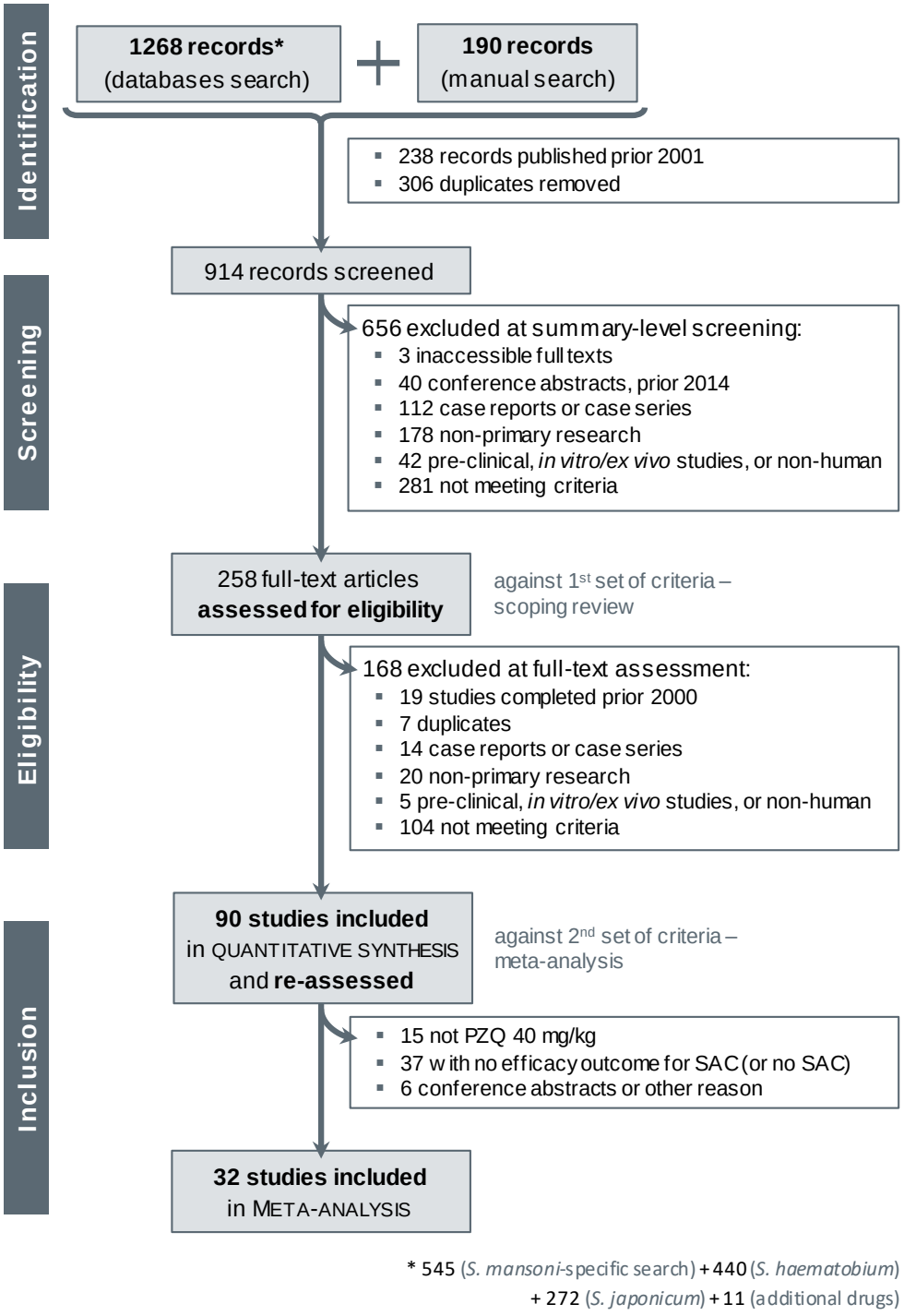


Figure 5.1: **Flow of references from identification during bibliographic searches up to inclusion**. First set of criteria – Clinical studies conducted in a schistosomiasis endemic setting and completed in 2000 or later were eligible for the scoping review (quantitative synthesis) if at least part of the included participants were: (i) diagnosed for infection with *S. haematobium*, *S. mansoni* or *S. japonicum* before treatment; (ii) administered a drug with known or suspected antischistosomal activity; and (iii) reassessed for infection within 60 days of treatment. Second set of criteria – Studies were then eligible for the meta-analysis if the tested regimen was praziquantel (PZQ) 40 mg/kg and a cure rate or egg reduction rate was reported among school-aged children (SAC).

The 90 studies of interest for the scoping review comprise 104 independent cohorts of participants, and 166 arms. A cohort corresponds to the largest possible subset of study participants recruited in the same country and who were subjected to the exact same study procedures. Seventy-nine studies applied a consistent protocol to all participants, and therefore consist of a single cohort each. The other 11 studies applied variable procedures to different subsets of participants (for example, introducing variations in the protocol to account for differences in participants' age or *Schistosoma* species found at distinct recruitment sites), and therefore account for up to four cohorts each.

A cohort may be further divided into study arms or subsets of participants differing by a (unique) characteristic, whose impact on the measured treatment efficacy is being tested. Distinct study arms received different drugs, doses and regimens; or received treatment at different times around the schistosomiasis season. This distinction between studies, cohorts and arms was unavoidable to allow extraction of published data at the different levels in which it is reported in articles. Proposed definitions match those in Halder *et al.* [157] (reproduced in Appendix A.2).

Finally, for the purpose of the meta-analysis, cohorts and arms were rearranged into groups. Once again, this additional unit was indispensable to ensure aggregated data was extracted from published articles at the most specific and comparable level. Throughout the schistosomiasis work, a 'group' refers to a subset of the whole study population, such that all individuals in the subset are SAC, treated with **praziquantel 40 mg/kg** (**PZQ40**), and recruited in the same place (study sites within 100 kilometres of each other at most) and at the same time (year and season).

Overall, extracting data from publications in a meaningful way to enable its synthesis and comparison was challenging, which is a first indication of the need to access more comparable, individual-level data.

5.2 Quantitative synthesis

5.2.1 Estimation and characterisation of existing IPD on schistosomiasis treatment efficacy

Altogether, the 90 studies included in the scoping review enrolled an estimated 20,517 individuals whose data could contribute to pooled analyses of antischistosomal efficacy. As specified in methods and although a precise estimate is suggested, the exact number of schistosomiasis patients included in each study was not always straightforward to extract from publications. I occasionally had to derive a value from indirect information about prevalence of *Schistosoma* infection in the study area, especially for studies jointly investigating distinct *Schistosoma* species or yet other parasites. Datasets from such mixed-endemicity or multi-parasite studies are likely to be more complex (covering more varied information), hence to require more curation efforts before data from relevant participants can be included into IPD meta-analyses. However, excluding those studies from the scope of a shared data platform would mean overlooking a fifth to a quarter of the estimated total of schistosomiasis patients treated in the context of a clinical study (Figure 5.2).

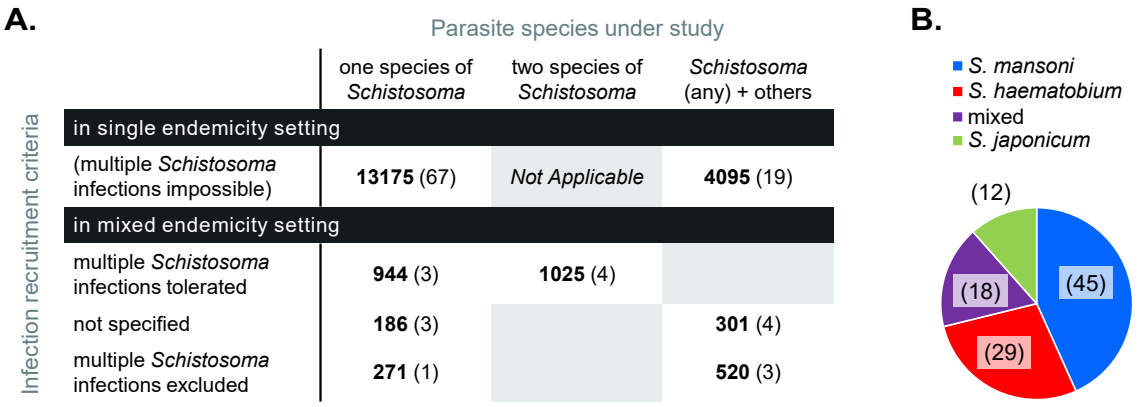


Figure 5.2: ***Schistosoma* infection considered for inclusion into a study, in view of the endemicity context and parasite(s) under investigation**. A. Counts of participants (out of 20,517) enrolled in a study focussing on one or several parasites, in a given endemicity context (single or mixed *Schistosoma* species infesting the recruitment site) and applying the corresponding criteria (tolerance or strict exclusion of patients infected with mixed *Schistosoma* species). Numbers in brackets are the corresponding count of cohorts (out of 104). B. Breakdown of cohorts by *Schistosoma* species endemic at the recruitment site(s). Mixed: *S. haematobium* and *S. mansoni* endemicity context.

I estimate that complete post-treatment follow-up data is available for about 17,500 of those over 20,000 enrolled participants, although the latter estimate was harder to derive due to inconsistent reporting of loss to follow-up. Throughout this chapter, counts of participants therefore refer to those with confirmed *Schistosoma* infection at baseline and who fully received their assigned treatment regimen, regardless of the possibility to subsequently follow them up. The received treatment was praziquantel (occasionally in combination with other drugs) for the vast majority of participants (18,302 individuals or 89.2% of the identified IPD), which suggests varied pooled analyses could be envisioned for this treatment.

Of the 52 countries requiring preventive chemotherapy [144, 162], I identified published clinical trial data for 26 (Figure 5.3). In particular, most of the data comes from six countries (Egypt, Uganda, Kenya, Niger, Zimbabwe and China):

each of these countries accounts for more than 8% of the estimated participants' total, and other countries account for much smaller shares. China alone holds most of the *S. japonicum* data (1,675 participants out of 2,327 in total for this species), while 17 and 18 countries share the 9,833 *S. haematobium* and 10,790 *S. mansoni* IPD, respectively.

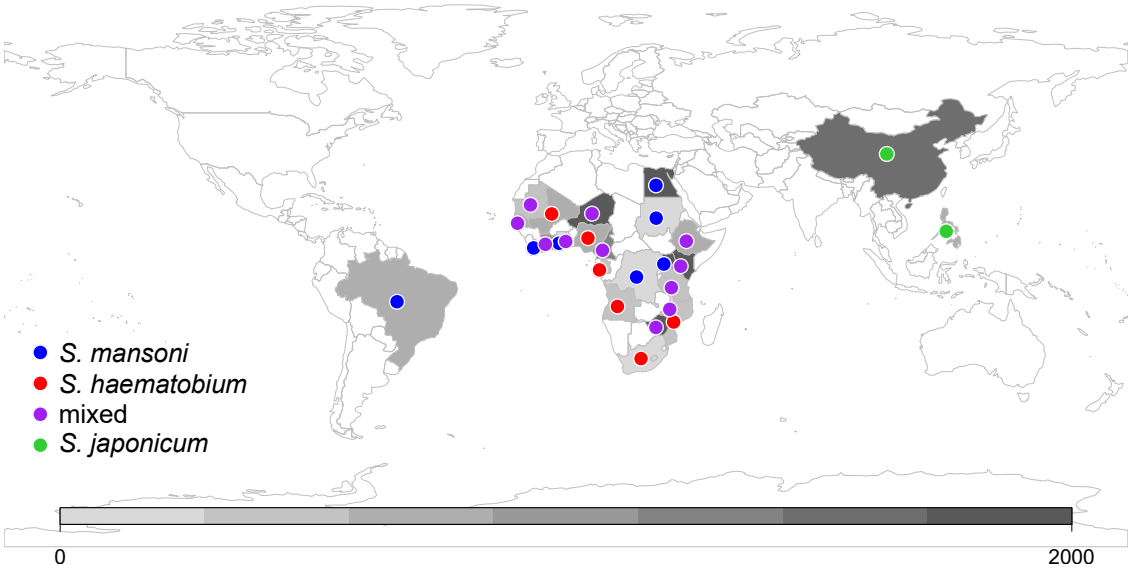


Figure 5.3: **Number of participants recruited and species studied by country**. The grey shading reflects the total of estimated patients enrolled in antischistosomal trials, from a low of 9 in Malawi to a high of 2,045 in Egypt. Dots highlight for which species the country holds published data. Mixed: data available for both *S. haematobium* and *S. mansoni*.

In addition to allowing exploration of variations in treatment efficacy across countries, the potential individual-level participant database could have tremendous value for studying outcomes within sub-populations sharing a relatively rare or under-studied individual characteristic, or for seeking out shared characteristics among those participants with suboptimal responses. Single trials are not usually powered to answer such questions, but one of the main advantages of pooling IPD

together is precisely to address this limitation of smaller datasets [175,176]. For example, Figure 5.2 illustrates that 26 cohorts consist of 4,916 patients coinfected with parasites other than *Schistosoma*. While schistosomiasis is most frequently studied in conjunction with STHs, it is noticeable that 15 cohorts (2,332 participants) were also systematically assessed for coinfection with *Plasmodium* and retained in the trial regardless of their diagnostic test result. Pooling IPD from those 15 cohorts together may enable the study of an hypothetical impact of malaria on antischistosomal treatment outcomes.

5.2.2 Heterogeneity in methodology and reporting

While a **Randomised Controlled Trial** (**RCT**) is considered the gold standard in terms of methodology to assess drug efficacy [177], other study designs also yield data that would be valuable to collate within an individual-level participant database. Forty-one of the studies identified here were comparative (that is to say, had more than one arm), of which 30 administered an intervention to the control arm (Figure 5.4A). If not used to assess drug efficacy in pooled analyses, such study data could still help in addressing questions around antischistosomal trial methodology, as well as in modelling or in validating models [159,164].

Additionally, studies retained in this scoping review did not necessarily aim at assessing drug efficacy in the first place. Some primarily aimed at characterising immune responses after treatment (15 studies), or at identifying biomarkers of infection or new diagnostic tests (14 studies). Many of these studies with alternate

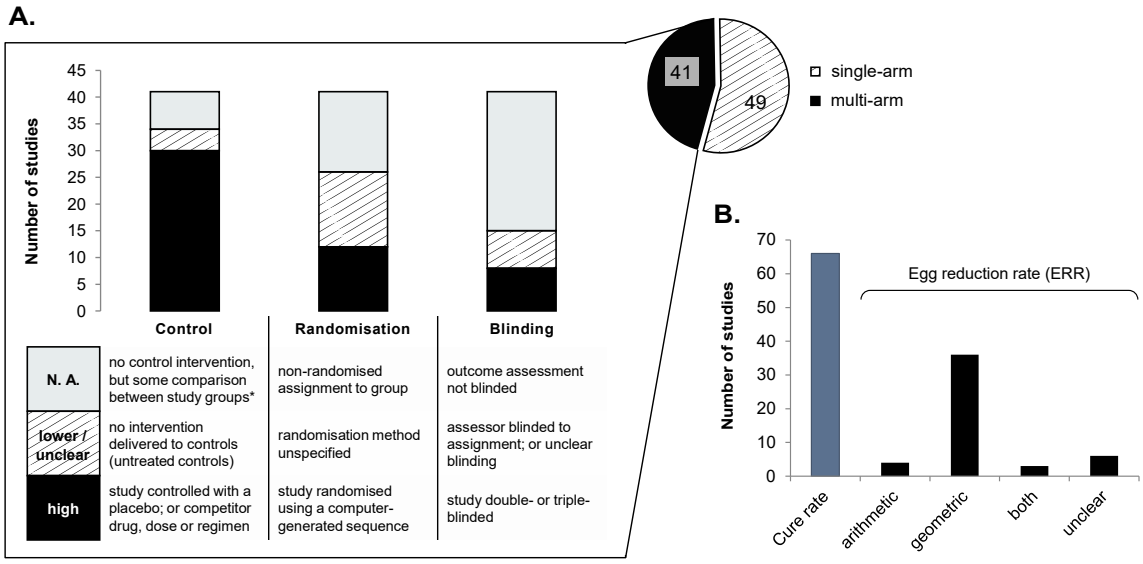


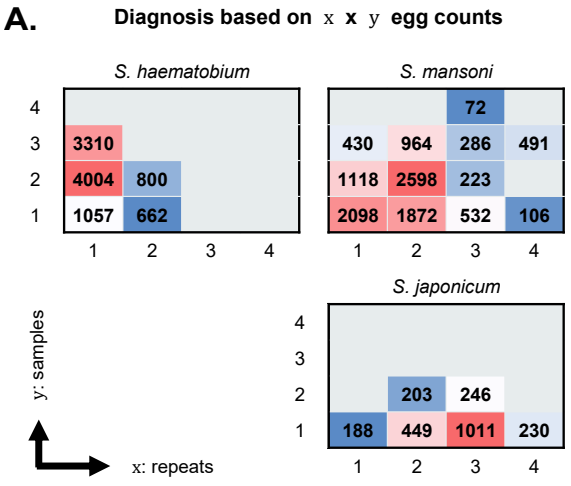
Figure 5.4: **Study design and outcome measures as reported in eligible articles and conference abstracts**. A. The quality of the three design elements (control, randomisation and blinding) was assessed against expected 'gold standards' of drug efficacy trials and graded as high, lower/unclear, or NA (not applicable). *For example, comparison of response to the same treatment delivered in different endemicity contexts. B. Count of studies providing a drug efficacy measure as cure rate and Egg Reduction Rate (ERR). All studies reporting an ERR also reported a cure rate, except one [178].

objectives did not readily report an efficacy outcome (cure rate or ERR) for the administered antischistosomal, hence their ineligibility for aggregated data meta-analyses but potential for use in IPD ones. A cure rate was provided in 66 of the 90 studies, making this the most commonly reported measure of antischistosomal efficacy (Figure 5.4B).

Detecting and quantifying *Schistosoma* eggs in excreta (urine or stool, depending on the infective species and form of the disease it causes) was the primary diagnostic test for all cohorts. Knowledge of participants' egg counts allows computing the population's MEC for each timepoint with a diagnosis, and eventually the corresponding ERR (as I explained in methods). Seven studies used an arithmetic mean to de-

rive the population MEC, but 39 preferred to use a geometric mean (Figure 5.4B). With a mix of arithmetic and geometric ERR co-existing in the literature, published population outcomes cannot be readily compared.

In addition, different laboratory techniques may be used to process excreta and prepare microscopy slides ahead of counting eggs, and variable numbers of independent samples per participant and/or slides per sample may be prepared (Figure 5.5). Those parameters impact on the diagnostic test sensitivity and on the population ERR eventually calculated [159]. To optimise comparability across studies, IPD should be shared down to the level of raw egg counts [133], which often corresponds to several measurements for one single participant and diagnostic timepoint (up to 12 in the present sample – Figure 5.5A).



B. Stool preparation technique

	S. mansoni	*S. japonicum*
Kato-Katz		
41.7, 42 or 42.7 mg	5045	717
50 mg	33	205
modified	215	
unspecified	3995	1260
Kato		
25 mg	154	
41.7 or 43.7 mg	104	
50 mg	155	
unspecified	892	
Katz		
unspecified	128	
Other		
concentration/hatching	9	145
unspecified	60	

Figure 5.5: **Number of participants assessed for infection with *S. haematobium*, *S. mansoni* or *S. japonicum* by diagnostic approach**. A. For each of y samples collected per participant, x microscopy slides were prepared to number eggs in urine (*S. haematobium*) or stools (*S. mansoni* or *S. japonicum*). B. The technique for processing urine samples was standard (10 mL urine filtration), but those used to prepare stool samples varied. Values are participants' counts.

Finally, the chosen time of outcome assessment introduces some heterogeneity that IPD meta-analyses can also not fully address, though this can be controlled for in statistical analyses [164]. Moreover, within this sample of studies, about 8,144 participants were followed-up at a comparable three to four weeks post-treatment.

5.2.3 Homogeneity in data management tools

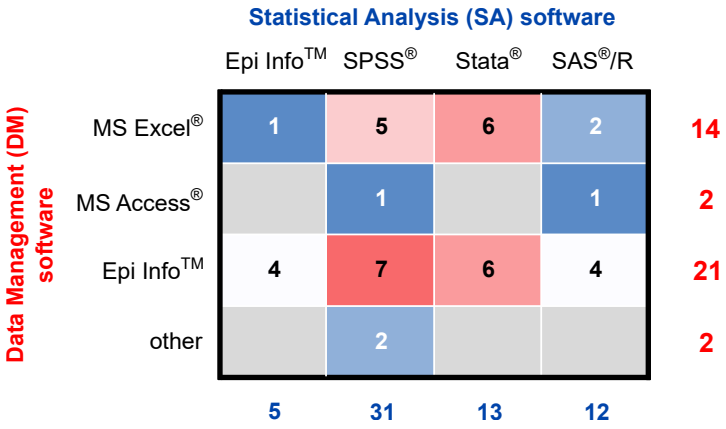


Table 5.1: **Data management (DM) and statistics software used in antischistosomal trials**. Numbers are counts of articles reporting the use of the corresponding combination of DM and statistics software. Totals of trials using a given statistical analysis (SA) package include studies which specified the SA but not the DM software. No article provided information regarding the DM but not the SA software.

While this information was not frequently reported (only 39 studies out of 90), there appeared to be limited variability in terms of the data management software used to process study data (Table 5.1). Indeed, 21 studies reported the use of Epi Info^{TM} (or EpiData), and 14 studies reported the use of MS Excel®, thus leaving very few observations of other software. Statistical analysis packages were more frequently stated (61 studies), and suggested a preference for SPSS® among schistosomiasis researchers, as this software was used in 50.8% of cases.

5.3 Meta-analysis

As a reminder, the hypothesis behind this meta-analysis was primarily tested as an illustration of the potential limitations of aggregated data meta-analyses. Strictly speaking, the meta-analysis fails to achieve its stated 'clinical and public health aim' (see below); however, this 'negative result' positively demonstrates the anticipated added benefit of sharing IPD.

5.3.1 Tested hypothesis: clinical and public health aim

The efficacy of praziquantel to treat urinary and intestinal schistosomiasis has been demonstrated in numerous trials, including several RCTs synthesised in Cochrane meta-analyses [137, 139, 140]. Altogether, evidence for the use of praziquantel as an antischistosomal is therefore strong. Yet, some studies performed since 2000 and identified in the scoping review reported unexpectedly low measures of efficacy, including three studies with cure rates below 40% [179–181] and another 6 studies with (uncorrected) ERRs strictly below 80% [182–187].

As praziquantel gets distributed more frequently and widely for preventive chemotherapy [144, 188, 189], it is important to understand the reasons behind those apparently poor responses to treatment. This meta-analysis aimed to investigate how the efficacy of PZQ40 (the recommended preventive chemotherapy regimen [142]) varies across places and over time, in light of the year when preventive chemotherapy was first introduced as a schistosomiasis control strategy in the corresponding

area. Increased drug pressure within targeted SAC in countries that have rolled out preventive chemotherapy for a while could indeed lead to a decline in PZQ40 efficacy, and there is agreement for the need to monitor efficacy of this drug to ensure success of this control strategy [154, 155].

5.3.2 Spatiotemporal coverage of included groups

The meta-analysis set out to identify spatiotemporal trends (if any) in PZQ40 efficacy among SAC, hence data underlying the model should not aggregate participants with mixed spatiotemporal characteristics. All except one article [190] reported on a single cohort, comprising participants who received PZQ40 on the same day (or PZQ40 together with placebo for one arm [186], here considered as equivalent to treatment with PZQ40). The study by Augusto *et al.* [190] recruited two independent cohorts in different seasons. Consequently, each cohort corresponds to a unique time of study, as desired. In contrast, some studies recruited their cohort(s) across distinct sites, to which specific GPS coordinates could be attributed with variable precision: overall, I am confident recruitment sites lie within a 100-km radius of the proposed point GPS data for at least 80% of the locations mapped. Whenever studies reported site-specific outcome measures, the study population was accordingly broken down into non-overlapping groups (one per site with unique GPS data).

From the 32 articles eligible for the meta-analysis (Figure 5.1), I identified 37 groups of schoolchildren (SAC) for whom a PZQ40 efficacy measure was reported

(or possible to derive from other published information) as cure rate and/or ERR. Of these 37 groups, two were recruited outside Sub-Saharan Africa (Philippines and Brazil) and/or assessed for *S. japonicum* infections [191]. Those two groups were excluded on the basis of their singularity, but I retained the other groups from the same multi-country study, recruited in Mauritania and Tanzania [191].

Eventually, 20 independent groups remained, corresponding to a total of 3,714 individual SAC treated for urinary schistosomiasis (caused by *S. haematobium*); plus 15 groups totalling 3,015 SAC treated for intestinal schistosomiasis (caused by *S. mansoni*).

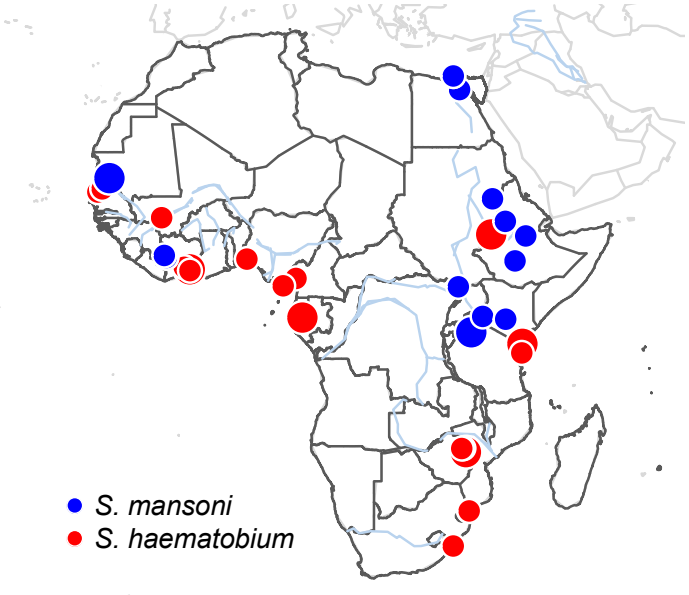


Figure 5.6: **Locations for which at least one measure of praziquantel 40 mg/kg efficacy among schoolchildren was found in the literature**. Red dots are for sites where *S. haematobium* studies were performed, blue ones are for *S. mansoni*. The dot's diameter reflects the precision of attributed coordinates for that site (over or less than 100 km confidence in matched GPS data).

Those 35 groups were recruited between 2000 and 2013, at sites in 27 geographical areas and across 16 countries (Figure 5.6). Most studies remained the unique one conducted in their area over the concerned period, but five geographical areas hosted several independent studies: Agboville [192,193] and Man [194–196] districts in Ivory Coast, Metobas district in Egypt [181,185], Murewha district in Zimbabwe [197,198], and the Trarza region in Mauritania [191, 199].

5.3.3 Range of reported praziquantel efficacy

When revisiting the original articles to extract complementary data for the meta-analysis, I found that variability in methodological and reporting choices was greater than previously uncovered. Not only ERRs may be derived from arithmetic or geometric MECs; but also the post-treatment geometric MECs may account for the whole study population or only uncured participants, and be calculated using diverse formulas. Similarly to Zwang and Olliaro [167], I applied 'corrections' in an effort to enhance the comparability of efficacy measures reported as ERRs, using the mathematical procedure described in methods. These corrections markedly reduced the amplitude of observed ERRs (Table 5.2 and Figure 5.7). Uncorrected ERRs were below the tentative 90% threshold of optimal efficacy proposed by WHO [143] for about a third (*S. haematobium*) and over half (*S. mansoni*) of the assessed groups. Among values that could be corrected, only one remained below this threshold. Of note, the 90% cut-off point suggested by WHO applies to arithmetic ERRs, and it is not fully transferable to geometric ones, which largely prevail here.

Reference	N	CR	MEC_b	MEC_f	formula	ERR_r	ERR_c
S. haematobium						82.7 (63.9, 91.9)	90.9 (82.3, 95.3)
Augusto, 2009 [190]	287	69.7	23.2	15.6	GM_0	32.8	94.5
Ojurongbe, 2014 [182]	245	49.4	9.8	4.2	GM_{unk}	57.1	NA
Augusto, 2009 [190]	269	98.2	23.5	7.3	GM_0	68.9	95.6
Inyang-Etoh, 2009 [186]	86	73.2	54.5	11.9	?	78.3	NA
De Clercq, 2002 [179]	88	30	8.7	1.7	GM_{1b}	80.5	NA
Tchuente, 2004 [200]	515	42.9	17.6	1.5	GM_1	91.4	NA
Keiser, 2014 [180]	21	33	45.9	2.7	GM_{0b}	94.1	96.9
Midzi, 2008 [201]	624	88.5	21.6	1.2	GM_{unk}	94.4	NA
Sissoko, 2009 [202]	389	53	41.2	1.8	GM_{unk}	95.6	NA
Mutapi, 2011 [198]	355	67	27	1.1	AM	95.9	NA
Keiser, 2010 [192]	26	88	32	1.1	GM_{0b}	96.6	96.7
De Clercq, 2002 [179]	45	76	10.1	0.3	GM_{1b}	97	NA
Stete, 2012 [193]	85	72.9	84.8	1.7	AM	98	NA
Mutapi, 2011 [197]	77	100	62	0	AM	100	NA
Guidi, 2010 [203]	178	46	-	-	NA	-	-
Ouldabdallahi, 2013 [199]	74	67.5	-	-	NA	-	-
Jinabhai, 2001 [204]	18	72.2	-	-	NA	-	-
Borrmann, 2001 [205]	89	73	37	-	GM_0	-	-
Mekonnen, 2013 [206]	85	87.1	-	-	NA	-	-
Kihara, 2009 [207]	158	89.9	-	-	NA	-	-
S. mansoni						75.4 (69.2-80.4)	94.8 (87.1-97.9)
Botros, 2005 [183]	32	62.5	38	13.8	GM_0	63.7	92.3
Erko, 2012 [184]	144	73.6	268	85	GM_0	68.3	96.8
Barakat, 2011 [185]	588	78.8	123.9	35.5	GM_0	71.4	98.9
Barakat, 2015 [181]	73	39.6	298.7	69.6	GM_0	76.7	87.6
Reta, 2013 [187]	187	82.9	258	53	GM_{unk}	79.5	NA
Obonyo, 2010 [208]	101	65	257.9	41.1	GM_0	84.1	98.4
Olliaro*, 2011 [191]	91	91.4	278.3	1.1	AM/GM_1	89.2	NA
Olliaro*, 2011 [191]	119	86.6	813.5	1.3	AM/GM_1	91.9	NA
Degu, 2002 [209]	148	94	-	-	?	97	NA
Scherrer, 2009 [195]	42	97.6	46.6	1.1	GM_1	97.7	NA
Namwanje, 2011 [210]	139	-	-	-	NA	99	NA
Mohamed, 2009 [211]	46	100	157	0	GM_{unk}	100	NA
Raso, 2004 [194]	27	59.3	-	-	NA	-	-
Adoubryn, 2012 [196]	124	78.8	-	-	NA	-	-
Kihara, 2007 [212]	1154	92.6	6.8	-	GM_1	-	-

*Reported MEC_b is arithmetic, while MEC_f and ERR are geometric.
AM = $mean(x_i)$; GM_1 = $antilog(mean(log(x_i + 1)))$ and GM_{1b} = $antilog(mean(log(x_i + 1)) - 1)$, $x_i \geqslant 0$.
GM_0 and GM_{0b} = $antilog(mean(log(x_i)))$, $x_i > 0$; GM_{unk}: unknown geometric mean calculation.

Table 5.2: **Reported and corrected measures of praziquantel efficacy**. See next page for complete caption.

Table 5.2 (caption): **Reported and corrected measures of praziquantel efficacy**. Corrections enabled homogenisation of MECs and ERRs, taking GM_1 as the standard. For each species, pooled means and associated 95% confidence intervals were calculated, using Huber-White sandwich estimators [173]. The pooled ERR_c jointly take account of corrected and raw data; raw values were kept for ERRs that did not require correction or that could not be corrected. Of note, the publication year indicated in the reference does not necessarily reflect the study year. N: Number of participants; CR: Cure rate; $MEC_{b/f}$: baseline and follow-up MEC, respectively; $ERR_{r/c}$: raw (as reported) and corrected ERR, respectively; NA: Not applicable.

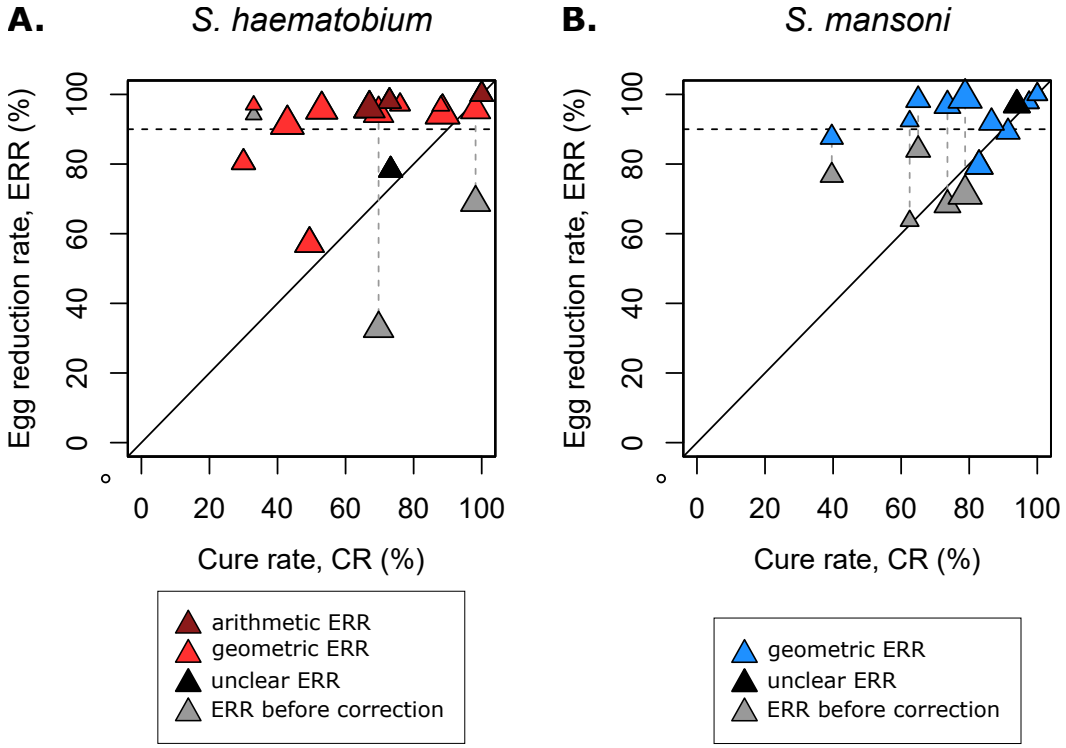


Figure 5.7: **Cure rates (CRs) *versus* egg reduction rates (ERRs) before and after correction**. Vertical dotted lines link uncorrected ERRs to their corrected value (the correction procedure is detailed in section 4.2.2). The horizontal dotted line underlines the suggested WHO efficacy threshold. Triangles are proportional to the natural logarithm of the sample size from which the estimate is derived.

From here onwards, MEC and ERR values presented and discussed correspond to corrected ones (unless otherwise stated).

5.3.4 Descriptive analyses prior to statistical modelling

CRs and ERRs in light of heterogeneity in study design and procedures

Participants of the studies considered in this meta-analysis were followed up at varied post-treatment timepoints within the retained 14-60 days timeframe, and diagnosed using varied numbers of samples and test repeats (Figure 5.8). Similarly to the impact illustrated above of comparing non-strictly equivalent MECs and ERRs, differences in procedures limit the comparability of outcomes from one study to the next, by inducing some variability between efficacy estimates which has no clinical or epidemiological relevance. For example, there appeared to be an excess of low cure rates at early timepoints of outcome assessment (three weeks) for *S. haematobium* infections, but the concerned groups also showed high ERRs (Figure 5.8).

No meaningful correction exists to homogenise MECs or ERRs taken at different timepoints; hence it is preferable to control for the effect of outcome assessment time by introducing it as a covariate in models (as performed in [164] with IPD). Aggregated data meta-analysis cannot account for other sources of variations that are due to differences in procedures (number of tested samples and test repeats, impacting on diagnostic test sensitivity): further covariates could in principle be considered in models, but with such a small sample (20 and 15 groups with available data at most for *S. haematobium* and *S. mansoni*), this would result in model overfitting [154].

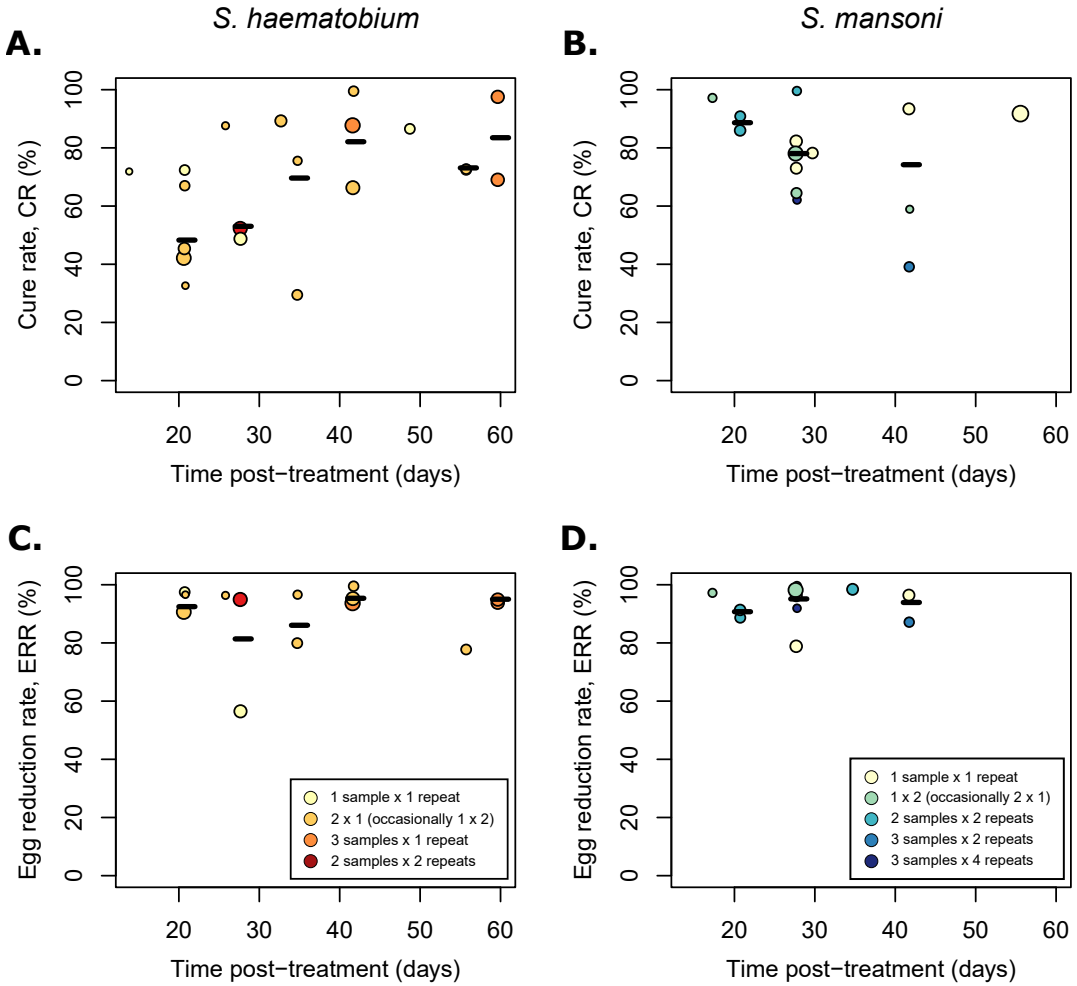


Figure 5.8: **Cure rates (CR) and egg reduction rates (ERR) by time of outcome assessment**. Time expressed from treatment day = 0. Black segments are means (weighted by sample size) of CR/ERR for each week, where applicable (more than one CR/ERR observation for the corresponding weekly period). Dots are colour-coded based on the number of samples and repeats used to assess treatment outcomes in the corresponding group. Dots are proportional to the natural logarithm of the sample size.

CRs and ERRs in light of clinical presentation and epidemiological setting

Because antischistosomal treatment efficacy can vary with initial intensity of infection, some authors report cure rates and ERRs separately by subgroup of baseline infection intensity. In other words, they consider participants with light, moderate (if applicable) and high intensities of infection at baseline as two to three separate study populations. For *S. mansoni* infections, the WHO defines infection intensities as light if an individual shows less than 100 eggs in their stool upon diagnosis by microscopy; moderate for 100 to 399 eggs observed; and high for 400 eggs and above [143]. Only two categories are defined for *S. haematobium* infections, with less than 50 eggs observed in urine samples considered as a light infection, and 50 or above considered as high [143]. Four studies [179, 182, 185, 200] reported a population (group) cure rate/ERR as well as the corresponding outcome measure(s) by infection intensity subgroup. One study [181] reported outcomes by subgroup only, thus the necessity to compute a weighted average to obtain a population-level, aggregated measure comparable to those reported in the majority of studies.

Figures 5.9 and 5.10 illustrate that, for the available data, cure rates and ERRs are not closely correlated with the initial MEC, nor with the prevalence of infection found in the study area. Although no striking effect is observed and in view of other (unaccounted) sources of dissimilarities across aggregated cure rates and ERRs, I still tested both covariates in statistical models. The baseline MEC may indeed impact on an individual's outcome, while the prevalence is a geographical characteristic that may account for some of the site-to-site variation in efficacy.

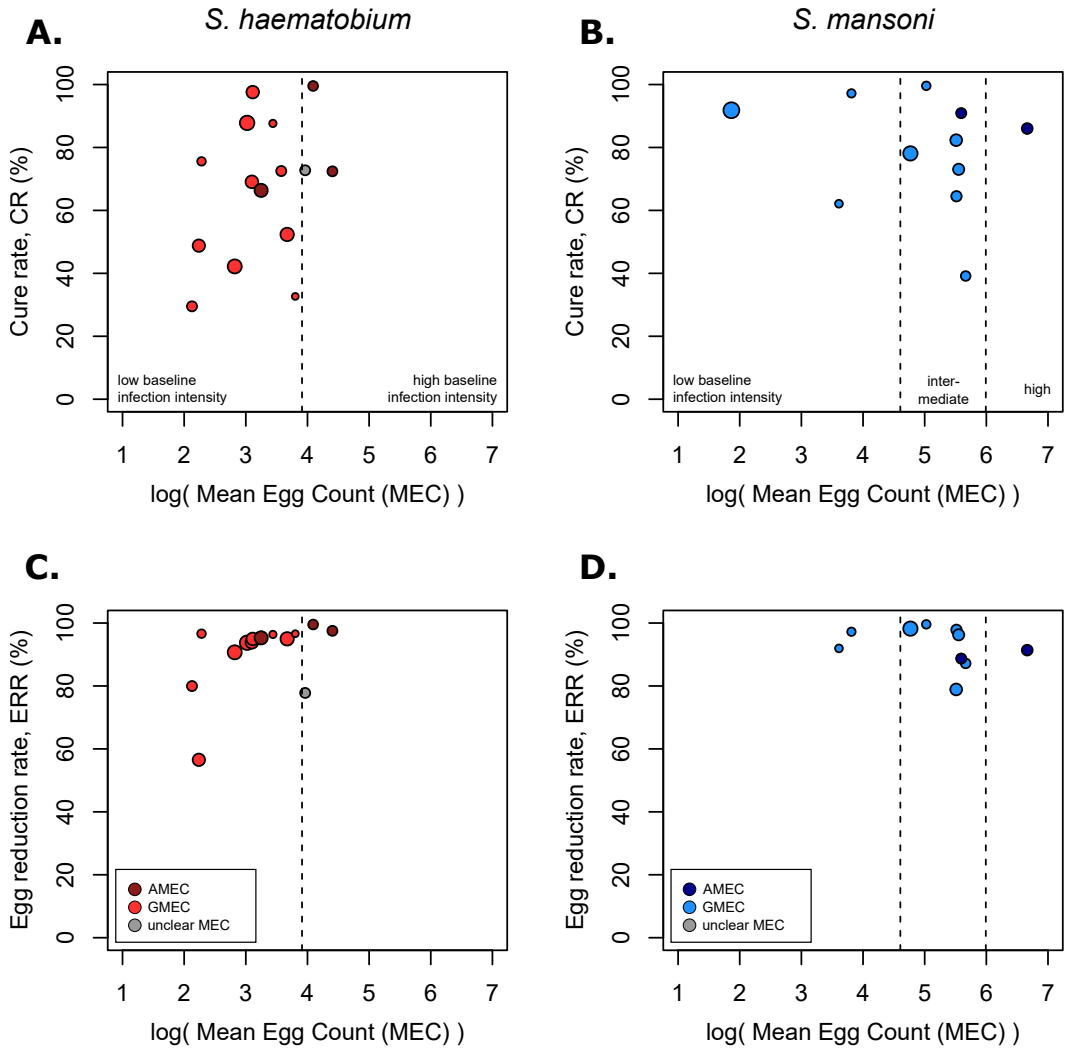


Figure 5.9: **Observed cure rates (CR) and egg reduction rates (ERR) in relation to participants' baseline infection intensity**. Vertical lines highlight thresholds of infection intensity groups defined by WHO at the individual level [143]. Mean egg counts (MEC) shown and the corresponding type (arithmetic, AMEC; geometric, GMEC; or unclear) are those reported at baseline and may differ from the ERR type. Dots are proportional to the natural logarithm of the sample size.

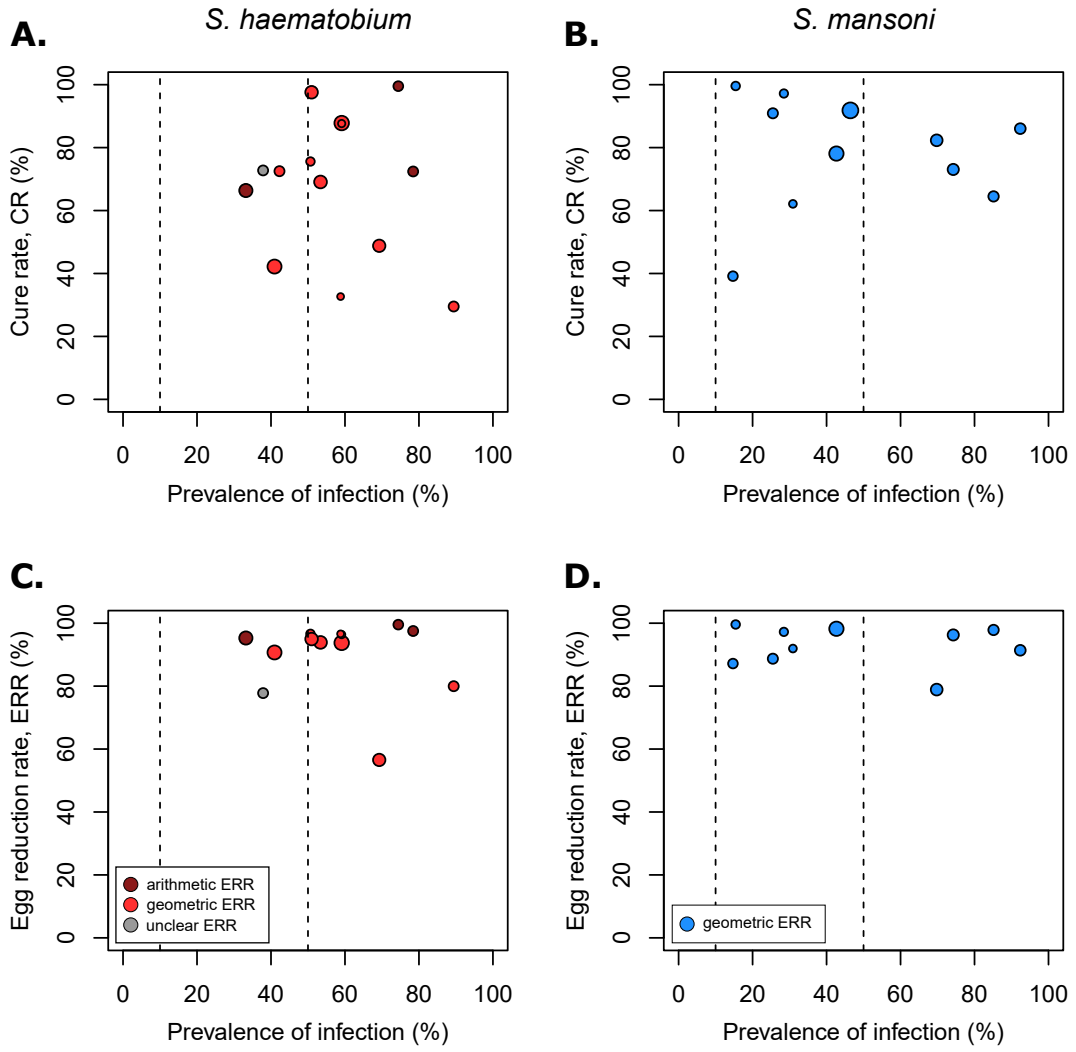


Figure 5.10: **Observed cure rates (CR) and egg reduction rates (ERR) in relation to the prevalence of infection in the study population**. Vertical lines highlight thresholds of prevalence defined by WHO to determine communities requiring preventive chemotherapy [143]. Dots are proportional to the natural logarithm of the sample size.

5.3.5 Univariate modelling from aggregated data

I fitted univariate regression models to the data using CR and ERR for each species as response variables, and using the variables of interest for the hypothesis (study year, and start year of preventive chemotherapy in the country) as well as the factors discussed above (prevalence, baseline MEC and post-treatment day of follow-up) as explanatory variables. A statistically significant effect of the study year was uncovered only for *S. mansoni* CR (Table 5.3). This effect is however driven by one particularly low observation (39.62% cure rate reported in Egypt in 2013 [181]); and the significance is lost when this outlier is ignored. The only other significant effect (or very close to significance) at the traditional 0.05-level is the one of outcome assessment time on CR for both species, but the associated relative risks are very close to 1, hence this covariate is not really impactful.

No multivariate modelling (fitting all covariates jointly to CR and ERR for each species) was performed, as the absence of significant or relevant effects in univariate models suggests that no more would come out of a multivariate one. In addition, the great heterogeneity in MECs and ERRs, which could only be corrected partly, suggests that modelling from aggregated data using those variables is far from ideal: the observations being compared are strictly speaking not comparable. Nonetheless, there remains a suspicious trend of seemingly decreasing PZQ40 efficacy over time when plotting the data and fitting the regression line from univariate modelling ($ERR \sim Study\ Year$ – Figure 5.11), which warrants closer investigation with reliable data such as homogenised IPD.

Variable (range of values)	**RR** (95% CI)	**Robust p-value**	**Missing**
S. haematobium CR (30, 100)			0
Study year (2000, 2011)	0.983 (0.817, 1.183)	0.866	3
PC start year (1980, 2013)	1.026 (0.998, 1.054)	0.138	2
Prevalence (18.9, 92.7)	1.012 (0.983, 1.042)	0.458	2
Baseline MEC (8.7, 84.8)	1.009 (0.955, 1.066)	0.765	5
Day of follow-up (14, 60)	1.034 (1.010, 1.060)	0.052	0
S. mansoni CR (39.62, 100)			1
Study year (2000, 2013)	0.844 (0.776, 0.919)	0.017	2
PC start year (1980, 2013)	1.025 (0.988, 1.063)	0.256	0
Prevalence (15.27, 93.01)	0.999 (0.955, 1.044)	0.960	1
Baseline MEC (6.83, 813.5)	0.999 (0.997, 1.001)	0.319	4
Day of follow-up (17.5, 56)	1.026 (1.009, 1.043)	0.040	0
S. haematobium ERR (57.14, 100)			6
Study year (2000, 2011)	0.847 (0.706, 1.016)	0.148	3
PC start year (1980, 2013)	0.973 (0.873, 1.085)	0.649	2
Prevalence (18.9, 92.7)	0.974 (0.930, 1.021)	0.334	2
Baseline MEC (8.7, 84.8)	1.054 (0.876, 1.268)	0.609	5
Day of follow-up (14, 60)	1.020 (0.963, 1.081)	0.532	0
S. mansoni ERR (79.46, 100)			3
Study year (2000, 2013)	0.875 (0.675, 1.136)	0.373	2
PC start year (1980, 2013)	0.966 (0.860, 1.084)	0.585	0
Prevalence (15.27, 93.01)	0.989 (0.939, 1.042)	0.707	1
Baseline MEC (6.83, 813.5)	0.999 (0.995, 1.002)	0.431	4
Day of follow-up (17.5, 56)	1.020 (0.827, 1.257)	0.864	0

Table 5.3: **Summary results of univariate statistical regression models fitted to praziquantel efficacy data**. Robust p-values were calculated using Huber-White sandwich estimators of the standard error [173], to account for the effects of potential correlation (dependency) among efficacies derived from the same study. Missing: Number of groups for which the corresponding covariate value was missing, out of 20 (14) and 15 (12) for *S. haematobium* CR (ERR) and *S. mansoni* CR (ERR), respectively. RR: Relative risk; CI: Confidence interval; PC: Preventive chemotherapy; MEC: Mean egg count.

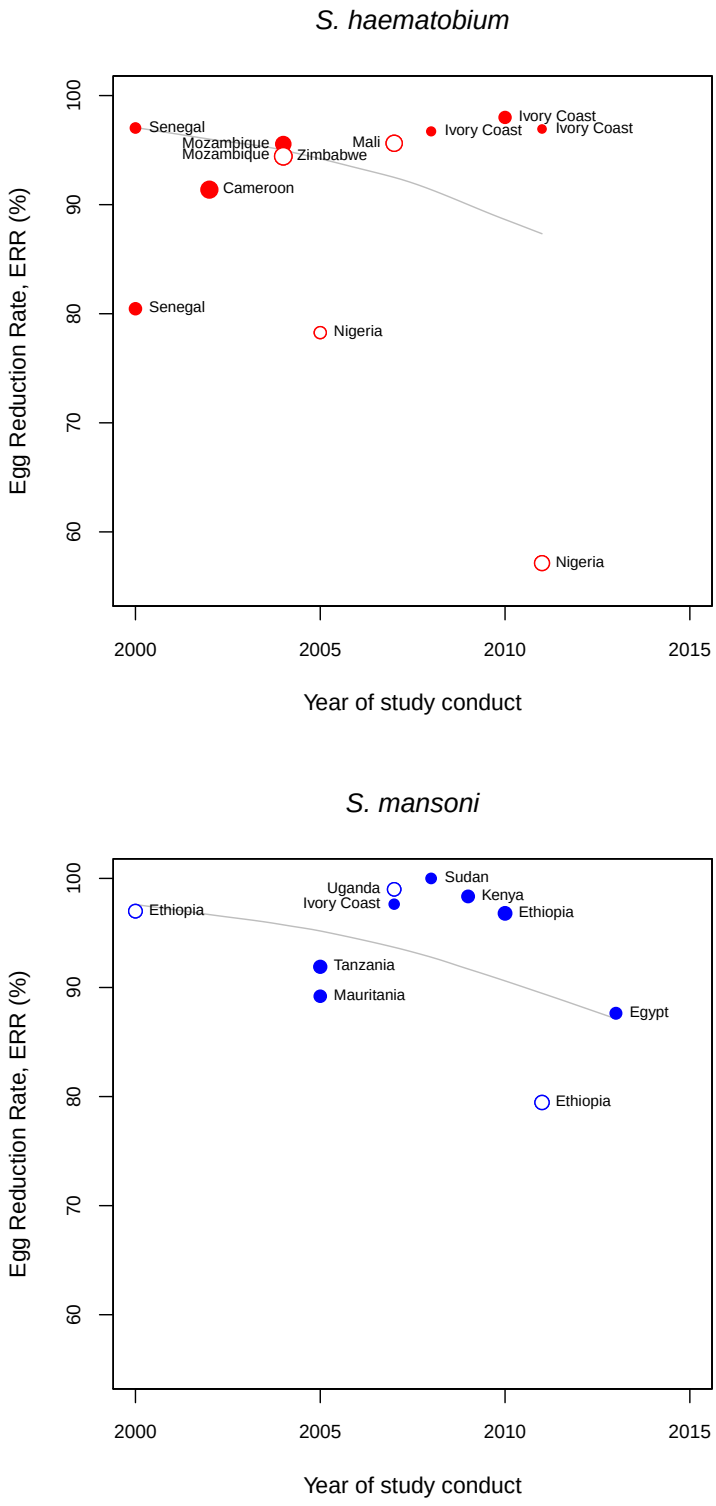


Figure 5.11: **Temporal trends in efficacy of praziquantel 40 mg/kg against schistosomiasis among African schoolchildren, reported as egg reduction rate (ERR)**. Lines are fitted values for the univariate model ($ERR \sim Study\ Year$). Plain dots are for ERR data which did not require correction, or which could be corrected; empty dots are for unclear ERR data which could not be corrected. Dots are proportional to the natural logarithm of the sample size from which the ERR estimate is derived.

Chapter 6

Discussion: Significance for antischistosomal efficacy data management and sharing

An examination of data management practices using published articles can only be partial and indirect; yet, the results presented in Chapter 5 already give some insights regarding usual data management tools and procedures in the schistosomiasis research community. In addition, when first starting to perform an aggregated data meta-analysis on the set of published studies, I did not anticipate encountering so many difficulties and sources of variation among reported population-level outcomes, which greatly limited the comparability of currently accessible therapeutic efficacy data across settings.

The data management implications of those important insights are summarised in Figure 6.1, and detailed in the remainder of Chapter 6. Of note, I discussed further strengths and limitations of the scoping review in Julé *et al.* [166] (published article reproduced in Appendix D), hence I primarily focus here on the data management lessons, most relevant within the context of this thesis. Lessons particularly concern the 'Analysing data' step of the lifecycle proposed by the UK Data Archive [52], as the landscaping exercise mostly highlighted how schistosomiasis investigators typically analyse and report their data. These analytical choices create challenges in re-using antischistosomal efficacy data in its aggregated form, and make the opportunities and value of creating an individual-level participant database clear.

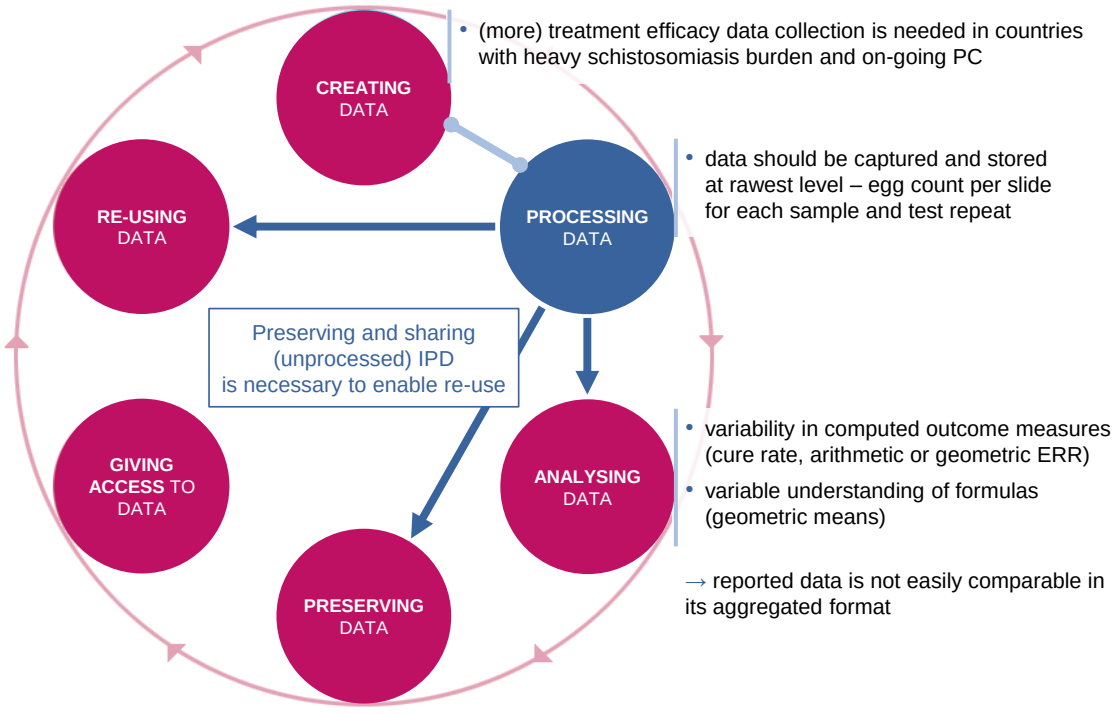


Figure 6.1: **Lessons from the schistosomiasis quantitative piece mapped onto the Research Data Lifecycle** (adapted from the UK Data Archive [52]). IPD: Individual Participant Data; PC: Preventive Chemotherapy; ERR: Egg Reduction Rate.

6.1 Interpreting guidelines: variability in study methods and data analysis

The landscaping exercise confirmed and extended the problem of variability in schistosomiasis clinical research methods and data processing approaches previously discussed by Danso-Appiah *et al.* [137]. This heterogeneity unnecessarily hampers the comparability of antischistosomal efficacy measures from study to study. WHO recently produced a detailed protocol [141] to promote consistent study implementation and data analysis, and to help make data summaries easier to re-analyse and compare. However, guidelines may be poorly followed, for example due to personal preferences: Saarnak *et al.* thus report the challenges of agreeing on a standard protocol for a multi-partner study on schistosomiasis, for which variations in the diagnostic method remained in practice, despite initial agreement on standard procedures [132]. Furthermore, global recommendations can change over time; and it will only be possible to apply newly recommended guidelines to retrospective data with access to appropriate IPD.

Provided that data can be shared down to the level of initial raw egg counts for each participant, a global database would allow pooling of data from studies with consistent diagnostic approaches; or, as shown in Olliaro *et al.* [159], to analyse the entire database using the WHO standard operating procedures (one single test of one single sample [141]). Preserving diagnostic test data at the rawest level is therefore essential, but it is unclear whether investigators usually keep this detailed information or only the average egg count for all slides and samples associated with

one participant's diagnosis. This emphasises the need to raise awareness about the significance of preserving collected data in this most valuable format, but also the importance of clarifying what granularity of IPD is relevant for each disease.

While some of the variability in antischistosomal trials is intrinsic to study design (for example, chosen time of follow-up for post-treatment outcome assessment), much of it relates to late data manipulation steps, analysis and reporting choices across articles. This is true for the main source of heterogeneity that was noted at the meta-analysis stage; namely, the heterogeneity in calculations performed to derive ERRs and MECs, beyond the choice of an arithmetic or geometric mean. Such variability would be straightforward to address given access to IPD, by recalculating these metrics in a homogeneous manner. In addition, IPD would enable the exploration and comparison of alternative ways of expressing efficacy [159, 164].

The criteria for inclusion in the aggregated data meta-analysis reflected two major considerations: (i) to focus on the current target group of preventive chemotherapy campaigns (mostly SAC in Africa), and (ii) to optimise data comparability by limiting sources of methodological variations between included studies. Access to IPD would certainly mitigate the latter constraint. Applying the inclusion criteria to aggregated data left a total of 35 independent groups with a reported efficacy measure. When breaking down the data by species of interest, whether *S. haematobium* or *S. mansoni*, this meant the analysis relied on 3,714 and 3,015 participants, respectively for each species. Those are low numbers in light of the 68.5 million SAC having received preventive chemotherapy with praziquantel in 2016 alone [144], and in comparison to the 20,517 participants identified in the scoping review.

Eleven studies (enrolling over 1,600 schistosomiasis patients altogether) with CR and/or ERR data had to be excluded from this meta-analysis on the ground that they did not report these efficacy measures within the SAC subgroup; although the age-range of participants they enrolled suggested this target age-group was covered. Access to IPD would enable selection of eligible individuals from each such study, with potential to increase the sample size for more meaningful pooled analyses. In addition, this would allow an analytical focus on the 'real' WHO-defined 5 to 14 year-old target group [162] (and whichever age groupings of interest, see [167] for further points on this aspect), instead of the more loosely defined 5-19 years old assessed here. Of note, this would not change the fact that there is a discrepancy between guidelines and field interpretation of who should be considered as SAC for preventive chemotherapy, hence analyses restricted to the WHO target group may not necessarily be the most informative.

6.2 Scope and directions for the development a data-sharing platform

The landscaping exercise enabled the identification of over 20,000 IPD from 90 schistosomiasis treatment efficacy studies published between 2001 and 2015. A major challenge in estimating the size of a disease-specific IPD platform through a review of the literature is that not all identified data will eventually contribute to the planned database: some data may be unavailable, for reasons ranging from im-

possibility to contact authors up to reluctance or inability on their part to share their datasets [28, 160]. On the other hand, studies published since the search was performed and before 2001 constitute additional sources of relevant IPD, along with studies missed by the systematic search and unpublished ones.

Moreover, local variations in the preferred participant follow-up time led to the exclusion of potentially interesting studies. When consulted on the development of a schistosomiasis treatment efficacy data platform [133], some investigators suggested extending the boundary of acceptable time of first assessment to three months, instead of the two months (60 days) used as cut-off for inclusion in the present scoping review. A hundred and four articles were excluded at screening on the ground of the outcome assessment occurring too late, many of which had performed it at three months post-treatment. The purpose of intended meta-analyses should guide the choice of cut-off timepoint for inclusion, regardless of these analyses being performed on aggregated or individual-level data. If data pooled within a shared database is to be used for estimating long-term effects of treatment on infection intensity in a given population (rather than being strictly restricted to drug efficacy monitoring), then no restriction on follow-up timepoints need apply.

Nonetheless, the benefits of gathering more data and enabling more analyses must be balanced with the additional costs of curating and maintaining those [213]. The present landscaping exercise suggests that the following variables should be considered in the structure of the pooled database: study country and site(s) (ideally, down to exact GPS coordinates), *Schistosoma* species under study, study period, diagnostic method (number of samples and test repeats, sample preparation

technique), detailed regimen (drug name, dose, formulation and treatment repeats, if applicable) and timepoint(s) of follow-up [166]. Any study dataset with those variables, if properly prepared and documented, should not be too burdensome to include in the existing database, regardless of exact methodological choices. This finite set of variables to cover in the individual-level database structure thus supports its technical feasibility.

6.3 *De novo* data collection and continuous efficacy monitoring needs

As stated above, the total of independent outcome efficacy measures underlying the meta-analysis was low, with a cure rate known for 35 groups and an ERR for 26 (Table 5.2). For some groups, data was also missing for key variables such as the year of study, thus further decreasing sample sizes in regression models (Table 5.3). Pooled analyses of antischistosomal efficacy trial datasets have illustrated the additional power offered by IPD over their population-based counterparts [159, 164]. In the absence of a comprehensive database, these analyses however remained limited to a small number of trials and IPD. Furthermore, and importantly when investigating spatiotemporal trends, multiple observations of praziquantel efficacy within one same year and/or same country (let alone site) were rare across the meta-analysis dataset I could compile. In that context of sparse information (and even using IPD), it is not possible to address key questions on regional variations in praziquantel ef-

ficacy, nor on long-term changes that may have arisen since the introduction and scale-up of preventive chemotherapy.

In addition, although 20,000 (and likely more) IPD is significant and suggests a substantial database could be created, this amount of data for all schistosomiasis treatment studies published in 2001-2015 is relatively low, and reflects the lack of investment in clinical research on this NTD [214]. African countries with most data available (Egypt, Uganda, Kenya, Niger and Zimbabwe – Figure 5.3) do not necessarily match those with highest prevalence or incidence of schistosomiasis, as evidenced by the Global Burden of Disease 2013 study [138]. Kenya is indeed one of the 10 most affected countries (both in percentage or total number of cases across the country [138]) and one with the largest amount of antischistosomal efficacy data from clinical trials. In contrast, other countries like Eritrea, Madagascar or Benin are among the most affected [138] and have a running control programme [162], but no identifiable trial data, as can be judged from the predominantly English literature. A shared database would help identify areas where data is lacking, and the collation of IPD should be performed in parallel with *de novo* data collection where necessary, and when resources allow. Countries with a heavy schistosomiasis burden and on-going control programme but no trial data may also be the most at need of clinical research and data management capacity strengthening.

Altogether, these observations are suggestive of the need to exhaustively collate past data; but also to collect new data and ensure that relevant information accompanies it, including temporal and spatial information, as well as other information to be refined in view of future analytical needs.

Finally, the meta-analysis suggested that transparent reporting of data manipulations may be the key area for data management capacity strengthening needs. The paucity of reporting on the exact calculation performed to obtain MECs and ERRs made the proposed correction difficult to implement; and IPD may be difficult to pool together if datasets are equally ambiguously documented. Some ERRs were left uncorrected, which makes it unclear whether ERR observations underlying regression models (a mixed of corrected and uncorrected ERRs) are comparable. As a reminder, the absence of correction was justified by the fact that ERRs were reported in the literature in the same way as the 'standard' defined for this analysis; or unjustified but unfortunately impossible to address. For this reason and as already emphasised, none of the decreasing trends in PZQ40 efficacy suggested from the modelling exercise is of great significance. They nonetheless underline the need for continued monitoring of praziquantel efficacy to ensure the success of preventive chemotherapy is not jeopardised [154, 155].

6.4 Conclusion on the schistosomiasis research landscaping exercise

The present systematic review and meta-analysis support the interest of developing an IPD platform to better monitor antischistosomal efficacy. More than 20,000 IPD could contribute to extensive meta-analyses using a common analytical framework, hence addressing much of the heterogeneity that limits the potentiality and strength

of evidence derived from aggregated data. Access and use of these IPD could improve our knowledge of how effective treatments are in the overall population as well as in subsets (age groups, coinfected individuals, etc.), across locations and over time, as the identified studies enrolled an interestingly varied set of participants. In addition, IPD may be used in analyses with a methodological focus, for example investigating how to analyse data and report outcomes in a meaningful way [159,164]. Altogether, findings from this first quantitative piece establish the value of investing resources in the creation of a schistosomiasis data-sharing platform. Not only would the platform be valuable for its potential scientific outcomes; but also it is becoming a necessity, as funders and publishers are pushing forward the data-sharing agenda and implementing requirements to make research data more widely accessible [12,18–24].

With this in mind, schistosomiasis stakeholders, researchers and policymakers were invited to discuss their views and concerns relating to the establishment of a schistosomiasis data-sharing platform at a WHO-convened meeting in September 2015, where results from this scoping review were presented [133]. From the viewpoint of scientific and ultimately health benefits of this work, the attendees suggested considering a joint schistosomiasis and STHs platform, owing to the overlap in geographical spread and study methods of both diseases [133]. A similar scoping review performed on these helminthic infections confirmed there is abundance of data to include in the database (about 35,000 IPD [157]). There was general agreement for the establishment of a data platform, provided that appropriate measures are in place regarding the data governance and access rights, and to ensure recognition of initial authors and producers of the data. This stakeholders' support suggests

the platform is in principle feasible, but some practical issues must be addressed to ensure full uptake and success of the endeavour.

As a final consideration, data management and analysis capacity-building was highlighted during the workshop as a key need to be addressed when forming a data-sharing network. In their multi-country schistosomiasis study, Saarnak *et al.* also noticed limitations in the data management capacity of their LMIC partners, many of whom had never used a "digital database" beforehand [132]. Those present at the WHO meeting who worked in less-resourced settings accordingly described capacity strengthening in data management and statistical analysis as essential. Without capacity development, some of them feared that 'their' data may be exploited by others with greater capacity in those areas, especially in statistics. This request for capacity strengthening opportunities aside of data-sharing activities echoed what has been reported and promoted in the literature [2, 5, 33–36], and reinforced the relevance of focussing on data management practices as part of this thesis. The box below summarises the first lessons regarding clinical research data management practices that can be taken away from this landscaping exercise, and to which I will come back in later chapters, as we learn further lessons from this mixed-methods investigation.

KEY MESSAGES ON CLINICAL RESEARCH DATA MANAGEMENT FROM THE SCHISTOSOMIASIS PIECE

- Variations in the methods underlying similar therapeutic efficacy studies hamper the comparability of efficacy data, especially (but not only) when it is in an aggregated form
- Reporting choices further hamper the direct comparability of aggregated data, especially if underlying analytical approaches and data manipulations are not explicitly reported
- Recommended Standard Operating Procedures (SOPs) can help to harmonise study data, but appear insufficient to address the heterogeneity in methods and reporting choices: researchers may interpret SOPs differently, and recommendations change over time
- For the reasons above, study data should be preserved in its rawest form, which is most likely to enable future and/or pooled analyses (including in light of updated SOPs)
- Databases of IPD should consider the granularity of data that is most appropriate to collate for each disease: this may involve collating multiple data points for one single individual-level measure
- Epi Info™ and MS Excel® are widely used for data management in the schistosomiasis research community, although the latter is not a data management software

Chapter 7

Exploration of data management strategies and outcomes in clinical research on malaria

The exploratory analysis reported in this chapter was designed to pursue the quantitative characterisation of data management practices initiated with the schistosomiasis work, while addressing some constraints of focussing on that 'neglected' disease. Sections 7.1 to 7.4 respectively cover the introduction, methods, results and discussion regarding this follow-up, quantitative piece of work. Complementary analyses are also presented in Appendix B.

7.1 Introduction to the second quantitative piece

As part of the systematic review and aggregated data meta-analysis presented above, I explored the typical data management tools (software) and processes (data aggregation and statistical analysis choices) in the schistosomiasis research community. This exploration relied on concise information found in published articles, and did not allow an in-depth investigation of data management strategies. Efforts to create a schistosomiasis and STHs data platform are now ongoing under the auspice of IDDO. The collation of source datasets and documentation from antischistosomal trials should provide further insights into investigators' data management practices; however, this data gathering and curation endeavour is lengthy. To pursue my characterisation of data management in clinical research on infectious diseases of poverty, I thus redirected the focus of my doctoral work to another disease with more varied and readily accessible data sources to examine (malaria). The choice of malaria was primarily driven by this immediate availability of datasets and associated documentation through WWARN, underpinned by a large amount of research and actors focussing on that high-profile disease, which jointly provide an opportunity to explore data management practices among various clinical research teams.

7.1.1 Malaria: background and relevance as 'case' disease

Malaria is a parasitic disease of major significance worldwide: the latest WHO World Malaria Report estimates there were 216 million cases of the disease in 2016, across 91 countries [215]. Two species (*Plasmodium falciparum* and *P. vivax*) are

responsible for most of the human disease burden, and *P. falciparum* accounts for the majority of malaria cases and deaths in Africa [215].

Treatment, control and research efforts faced with antimalarial resistance

Plasmodium has a complex lifecycle involving multiple biological forms, and is transmitted to humans by *Anopheles* mosquitoes [216]. Infection with *Plasmodium* induces fever and chills in naïve individuals, and can rapidly progress to severe and fatal forms of malaria if left untreated [216]. A number of antimalarial medicines has been developed and used in endemic areas; however, the repeated acquisition of resistance against effective compounds by the parasite (and especially, by *P. falciparum*) jeopardises our ability to treat and control malaria [60]. WHO now recommends **Artemisinin-based Combination Therapies** (**ACTs** – joint administration of artemisinin and other antimalarials from a different therapeutic class) as the primary therapeutic option [217]. **Long-Lasting Insecticide-Treated Nets** (**LLITNs**) and insecticide residual spraying remain the most effective interventions in decreasing the malaria burden in tropical countries [218, 219].

In the absence of alternative drugs (although a few are in the Research & Development pipeline [220, 221]), it is essential to monitor and seek to preserve the efficacy of current antimalarials. WHO recommends regular surveillance efficacy trials in endemic areas [50], and has developed tools to support malaria national control programmes in that data collection, management and analysis work [49].

Variety of actors and practices in the malaria research community

In contrast with schistosomiasis and other so-called NTDs whose profile was only recently raised [222, 223], the fight against malaria has attracted attention from a varied set of policymakers, practitioners and researchers for a longer time [224]. Furthermore, owing to the repeated need for surveillance efficacy studies and trials testing new candidate drugs, many more clinical trials of antimalarials than of antischistosomals have been performed (Figure 7.1).

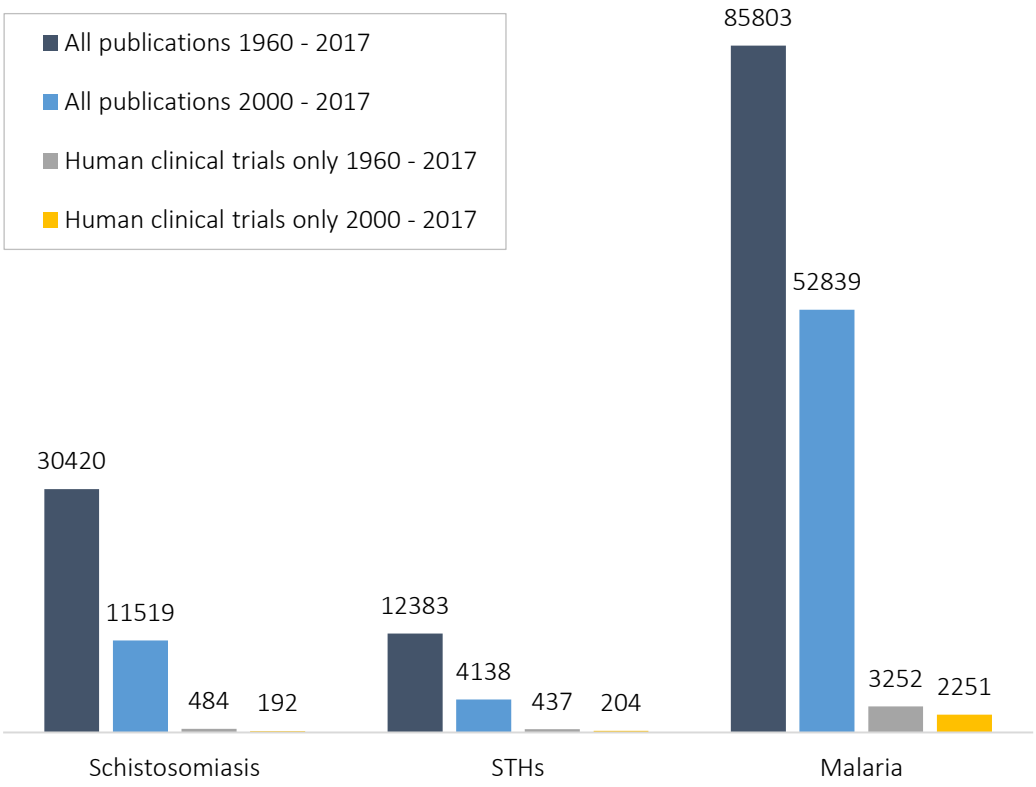


Figure 7.1: **Number of PubMed entries on schistosomiasis, soil-transmitted helminthiases (STHs) and malaria, for all publications and clinical trials only** (April 2018). Schistosomiasis search: ("schistosomiasis" OR "schistosoma mansoni" OR "schistosoma haematobium"). STHs search: ("soil transmitted helmin*" OR "soil transmitted nematod*" OR "geohelminth*" OR "ascariasis" OR "roundworm" OR "round worm" OR "roundworm" OR "trichuriasis" OR "whipworm" OR "whip worm" OR "ancylostomiasis" OR "necatoriasis" OR "hookworm infections"). Malaria search: "malaria".

The Infectious Diseases Data Observatory (IDDO) is a scientifically independent, multidisciplinary coalition that brings together members of the global infectious diseases community across the public health, research and humanitarian sectors to collaborate in the generation, analysis and application of data to improve patient outcomes [225]. IDDO is building upon the pioneering work of WWARN for malaria [213], which is now incorporated as a part of the IDDO portfolio. WWARN connects over 380 partners across the globe, ranging from private companies to academic research institutes and other key players in the global health field. This network provides a unique opportunity to explore the data management practices of numerous investigators interested in a given infectious disease of poverty, and operating within different organisational contexts.

Experience from WWARN data curators in gathering trial data from malaria research teams suggests great variability in the format and structure of databases, and in the associated data description. The lack of standardisation and occasionally sub-standard documentation of shared trial data greatly increases the cost and efforts required ahead of data re-use. The creation (and ultimately, the sharing) of self-explanatory, well-documented datasets could fast-track the resource-intensive, late curation process [6, 7, 25]; and strongly relies on the implementation of a robust DMS during (and beyond) the clinical trial lifecycle [10]. The pitfalls of (inadvertent) poor documentation or audit-trails of research activities and datasets have been illustrated by Buckheit and Donoho [226] and others [12], who reported the challenges of reproducing their own results after some time, thus highlighting the importance of continuous good data management for the viability of research data.

Moving forward with the schistosomiasis (and other) data platforms, it is important to clarify the reasons behind the variability in shared data and documentation, and to understand the extent to which these variations are due to critical differences in initial data management practices, or to the sharing behaviour of data contributors. Drawing lessons from WWARN's experience would also help to identify the remaining areas in need of improvement regarding data management and sharing within the malaria community itself.

7.1.2 Specific objectives and hypotheses tested

As a reminder and as graphically explained on Figure 7.2, a data management strategy corresponds to the 'theoretical plan' laid out ahead of the study, which should describe the personnel, tools and processes constituting the DMS (in the 'complex system' meaning of the term, as introduced in Chapter 2). Data management outcomes are the data files, including the dataset and the supporting documentation enabling its comprehension and re-use. Of note, the study documentation is a DMS outcome, but also an indicator of the strategy, as documents should precisely describe the intended procedures and plans.

In this second quantitative piece, I systematically examined the DMS outcomes in view to achieve the following objectives:

I. To characterise the variability in DMS outcomes across antimalarial trials, with emphasis on the comprehensiveness of supporting data documentation;

II. To characterise the variability in DMS strategies across antimalarial trials, as inferred from the procedures specified in accessible study protocols and/or DMPs, and from datasets and meta-data;

III. To uncover potential relationships between study characteristics and data management strategies and outcomes.

Achieving the first two (descriptive) objectives necessitated that I develop a framework for the systematic characterisation and comparison of data management strategies and outcomes across a large array of trials. Building on this quantitative characterisation, I searched for patterns and trends in data management indicators, and challenged the hypothesis that DMS vary with different attributes of the study and its actors. I focussed the latter correlation analysis on three major attributes.

First, I hypothesised that data management systems vary with the year when a study occurred, as different tools and guidelines become available over time to handle clinical trial data. Then, the parties involved (sponsor, funder and/or investigator) are suspected to determine data management strategies and outcomes, because different organisations and individuals likely have different procedures and preferences. Lastly, if taking a 'risk proportionate' approach to data management, the quality assurance process should be commensurate with the potential impact of an error in the data on study conclusions [227]. I therefore sought to test whether the study 'complexity' influences data management strategies, considering that larger, complicated trials may call for more robust systems to mitigate greater threats to data quality. I considered two sub-areas of complexity for a clinical study: one

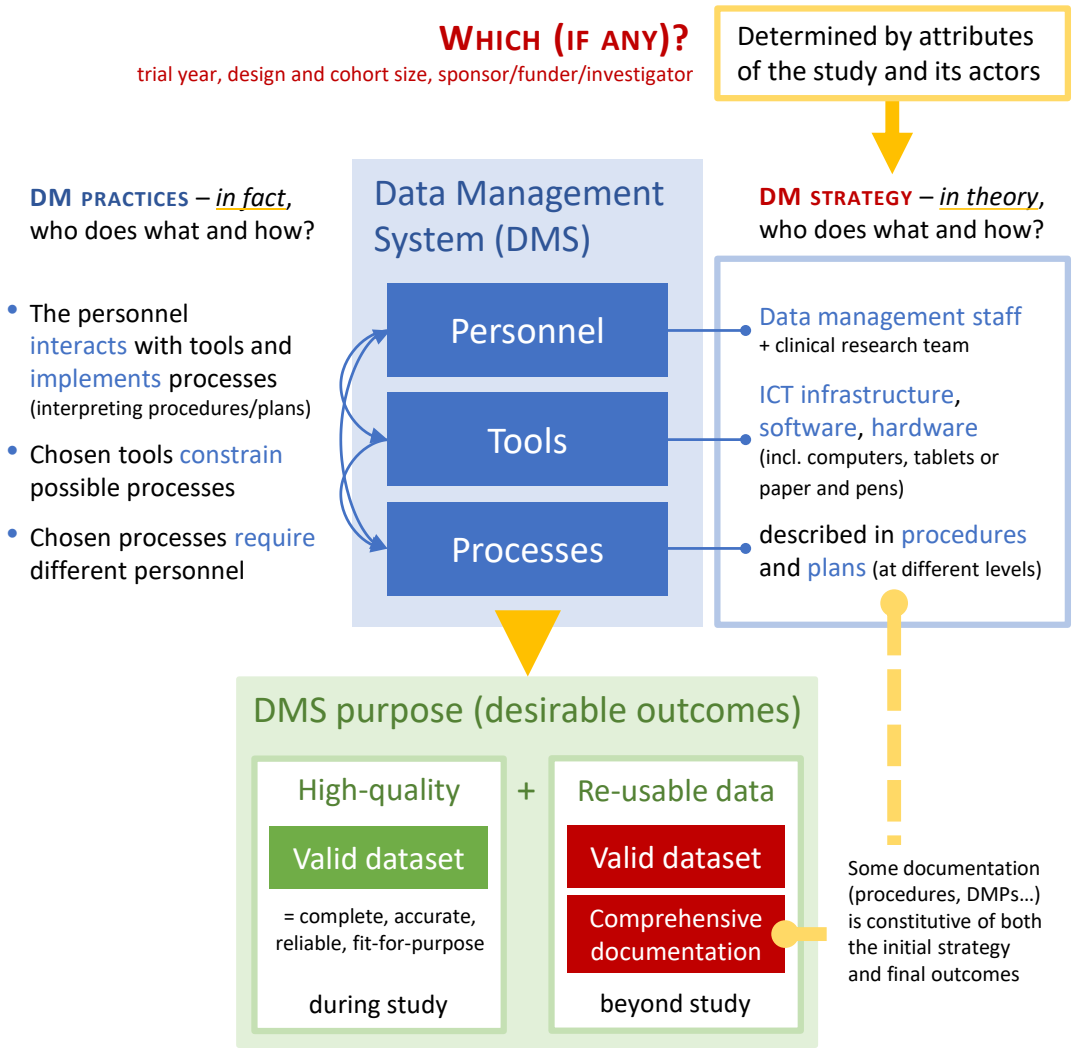


Figure 7.2: **Key components and outcomes of a Data Management System (DMS)**. The present quantitative analysis particularly investigates elements highlighted in red. Shared datasets and documentation serve to characterise data management strategies, and to search for correlations between the latter and attributes of the study and actors involved in its conduct.

related to the exact trial design (number of assignment arms, use of randomisation and/or blinding) and objectives (efficacy, safety, pharmacology); the other related to the size (total of participants enrolled, number of recruitment sites and countries).

7.2 Proposed quantitative methods to explore clinical research data management systems

This section briefly outlines the methods underlying the systematic exploration of clinical trial data shared with WWARN, to characterise and compare DMS within the malaria community. Appendix B further details the methodology, and especially, the iterative approach I followed to develop a 'theoretical' analytical framework (Appendix B.1).

7.2.1 Selection of antimalarial trials from the WWARN repository (sampling)

I pre-screened the WWARN repository for eligible studies in November 2016, at which time it contained 577 entries. Among those, I selected all treatment efficacy trials, and excluded other study designs. In case of duplicate WWARN entries for the same trial, I retained only the entry with the most documented or largest dataset (that is to say, largest number of participants). Studies recently shared and still awaiting curation were excluded, so that notes and reports from WWARN data curators could be considered in the analysis. Between January and April 2017, I contacted the contributors of 251 trials that remained after this preliminary screening. Contributors received a request (followed by a reminder) to opt out, or to provide additional information regarding their trial. In particular, I asked for a study protocol and for a dictionary describing the variables found in the contributed

dataset, as missing for the corresponding trial. Appendix B.2 reproduces an example of the communication addressed, and discusses the responses received from WWARN contributors contacted during this project.

After closing all communications, I re-examined eligible trials in view of additional documentation provided by contributors, or lack thereof. I checked for the availability of at least one 'key document', which I defined as a protocol, a report or an article enabling minimal understanding of the study and its corresponding dataset. A number of trials without any such documentation were excluded from the analysis (as presented and quantified in results below).

7.2.2 Data extraction

For each included trial, I extracted data from documentation related to the study (published article, study protocol and/or detailed report) and from documentation related to the dataset itself, as available (variables' dictionary, dataset annotations, database programming specification files, **Case Report Form** (**CRF**), data curation notes, etc.). I looked for study 'meta-data' within the data-related documentation, where meta-data is here defined as any structured description of the variables observed in the trial dataset [1, 54]. In parallel, I examined study protocols and reports for the presence of a DMP or equivalent data management section, as well as for the presence of a **Statistical Analysis Plan** (**SAP**). Of note, protocol sections on data management are rarely as comprehensive as DMPs should be, although guid-

ance regarding DMPs' expected contents is sparse and heterogeneous [116, 228, 229]. Here, I nonetheless refer to such small data management sections as DMPs, in light of the impossibility of accessing more exhaustive plans. Published articles were normally excluded from the assessment of DMP and/or SAP, as most publications contain very similar and concise methods sections, outlining their statistical (but not their data management) procedures. Such homogeneous and brief statements do not allow exploration of variability in DMP contents, as was my intention in the present piece of quantitative work.

Extracted information mainly concerned the study complexity (for example, information regarding the number of treatment arms, use of randomisation and blinding, cohort size and number of recruitment sites), and the study time and place. I also collected information regarding the study contributors (investigators, authors, sponsors and/or funders), with a particular interest in the type (public, private or non-profit) and country (local or foreign) of the diverse organisations involved in conducting the trial. Finally, extracted information covered indicators of initial data management choices (software used, dataset files and formats, etc.). Appendix B.3 comprehensively presents the variables and data I initially considered, and jointly explains why several of these variables were transformed (re-coded), or excluded altogether from statistical analyses. I made such modifications to account for correlations between variables reflecting a similar theme, as well as to minimise the overall number of covariates and statistical testing.

7.2.3 Statistical analyses and response variables

For statistical analyses, I considered three outcome (response) variables suggestive of initial data management strategies (Table 7.1). I defined these variables drawing on data management experts' viewpoints on the desirable attributes of study and data documentation (as reported in Appendix B.1); as well as on the work performed by data managers and clinical researchers from The Global Health Network, comparing the suitability of different software for clinical research data management [230].

Var.	Possible values	Description
DMP score (Y_{DMP})	**Low** – score 1 (or 0, observed for one trial) **Medium** – score 2 or 3 **High** – score 4	1 point per desirable feature of a DMP, namely the specification of: (i) data collection or capture methods (ii) data verification and QC procedures (iii) data security and protection measures (iv) data backup, storage and sharing strategies
Dictionary score (Y_{dic})	**Low** – score 1 (or 0, observed for one trial) **Medium** – score 2 or 3 **High** – score 4	1 point per desirable feature of a dictionary: (i) overall completeness, i.e. coverage of all var. found in dataset (ii) specification of name + brief description for each var. (iii) mention of possible values or levels for each var. (iv) mention of var. type or class (numeric, categorical…)
Software score (Y_{sw})	**Low** – Microsoft Excel® **Medium** – Epi Info™ or Microsoft Access® **High** – other (especially, SAS®)	Respectively, for trials reporting the use of: • MS Excel® for DM (regardless of shared data file and reported statistics software) • Epi Info™ for DM (or for statistics, if DM software unspecified); or MS Access® for DM (or MS Access® data file shared) • other software for DM; or SAS® for any purpose (including, SAS® data file shared with WWARN)

Table 7.1: **Response variables suggestive of initial data management strategies**. All outcome (response) variables are categorical variables with three ordered levels (low, medium, high). DM: Data Management; DMP: Data Management Plan (or procedures within study protocol); MS: Microsoft; QC: Quality Control; Var.: Variable(s).

For studies with available documents (DMPs), a first score (Y_{DMP}) was attributed based on the level of detail when describing the data management procedures, without intention to further 'grade' the appropriateness of described practices for the corresponding trial. A second score (Y_{dic}) reflects the level of detail and completeness in the information describing variables found in shared dataset. A third score (Y_{sw}) accounts for the fitness-for-purpose and appropriate functionalities of the software underlying data management, in line with the system's potentialities for data quality control checks and audit-trailing.

To uncover potential associations between study attributes and data management strategies, I ran preliminary bivariate analyses in R (version 3.4.1) [170], using **Analysis of Variance** (**ANOVA**), Chi-2 and Fisher's exact tests. To identify most influential factors, I then planned to run multivariate **Proportional Odds Logistic Regressions** (**POLR**), independently on each of the three response variables, and including as covariates all (and only) the variables having shown an influence in preliminary analyses (significant effect at the 0.05-level in ANOVA, Chi-2 and Fisher's exact tests).

7.3 Results: Variations in malaria research data management

In this section, I first describe the sample of antimalarial trials that contributed to the analysis, as well as the documentation that I could access to clarify the data management strategies underlying those trials. Using mostly descriptive statistics, I then assess correlations between study attributes and the three data management indicators defined above in Table 7.1.

7.3.1 Included trials and temporal trends in retrieved data

When first screening the WWARN repository for relevant studies, I identified 251 clinical trials with curated data. After I requested further documentation from data contributors, 167 of these studies met all inclusion criteria (Figure 7.3). Eligible trials included 108 studies performed in Africa by various institutions. In contrast, the 55 studies conducted in Asia largely came from investigators affiliated with Oxford overseas units. Some of these units have recently centralised their research data management and sharing systems [98]; hence, the data and documents they share is less likely to reflect the strategies and outcomes of the initial study DMS, which I seek to characterise here. Due to this rather homogeneous sample in Asia, I focussed the analysis on the African continent, likely to reflect a larger distribution of practices.

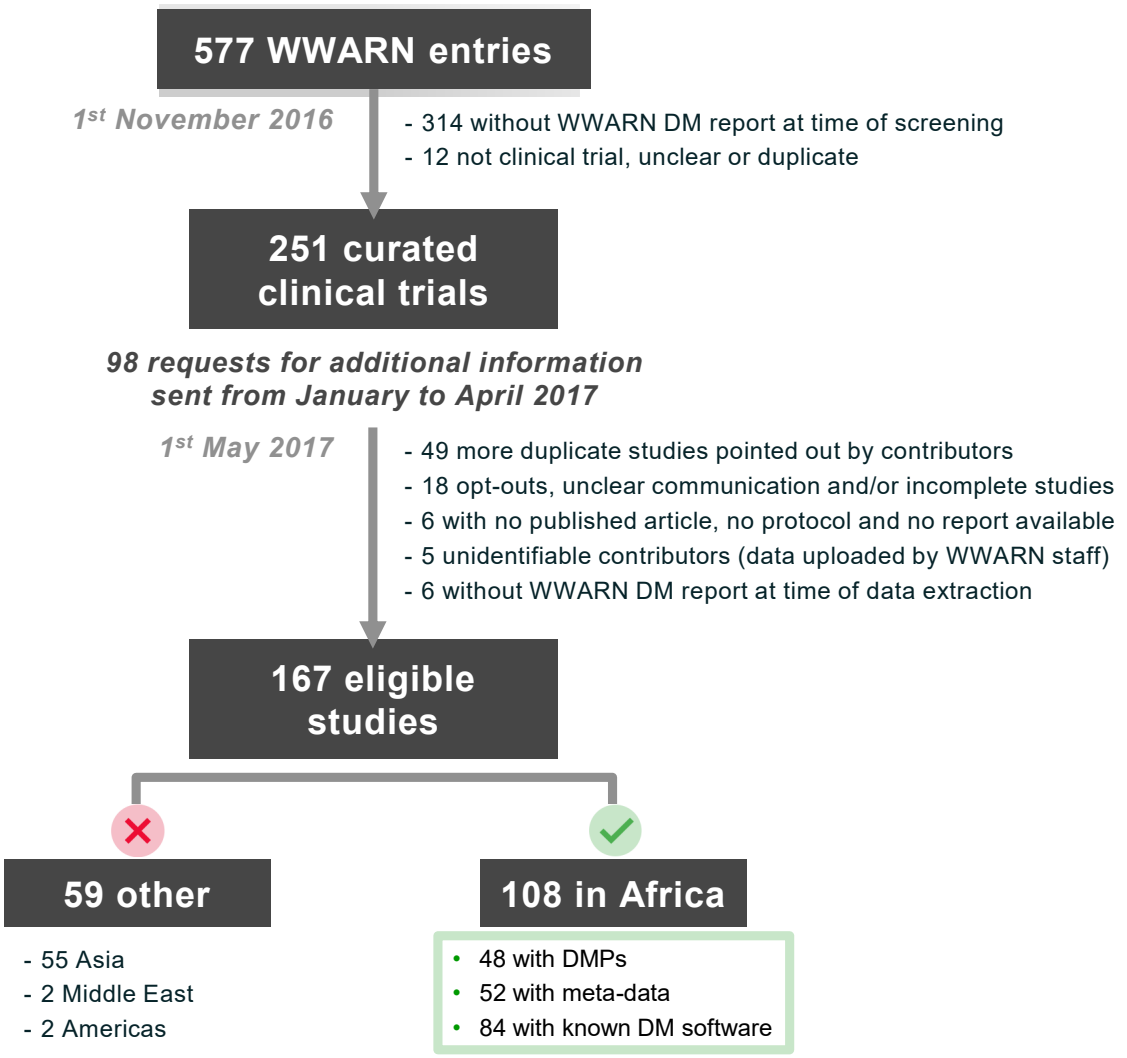


Figure 7.3: **Flow of studies screened within the WWARN repository, from eligibility up to inclusion**. Unique WWARN entries were retained if: (i) the associated study is a clinical trial; (ii) the dataset has been curated (that is to say, a WWARN Data Management (DM) report exists for the study); and (iii) the study data is documented by at least a published article, a study protocol or a detailed report. DMPs: Data Management 'Plans' (in fact, DM sections within protocols or reports); 'Meta-data': structured description of variables observed in the trial dataset (whether as direct annotations or separate documents).

Accessibility and existence of meta-data, DMPs and SAPs over time

The considered trials occurred between 1993 and 2015 (Figure 7.4). Most were published (93.5%), and the article was the only source of information available for 26 of the 108 studies (24.1%). Notably, of the 59 studies whose contributor responded to communications, 10 remained without accessible protocol or report, and 20 without any information describing variables in the shared dataset. Respondents rarely specified the reason for not sharing requested documents, although five acknowledged that data dictionaries did not exist for the corresponding dataset.

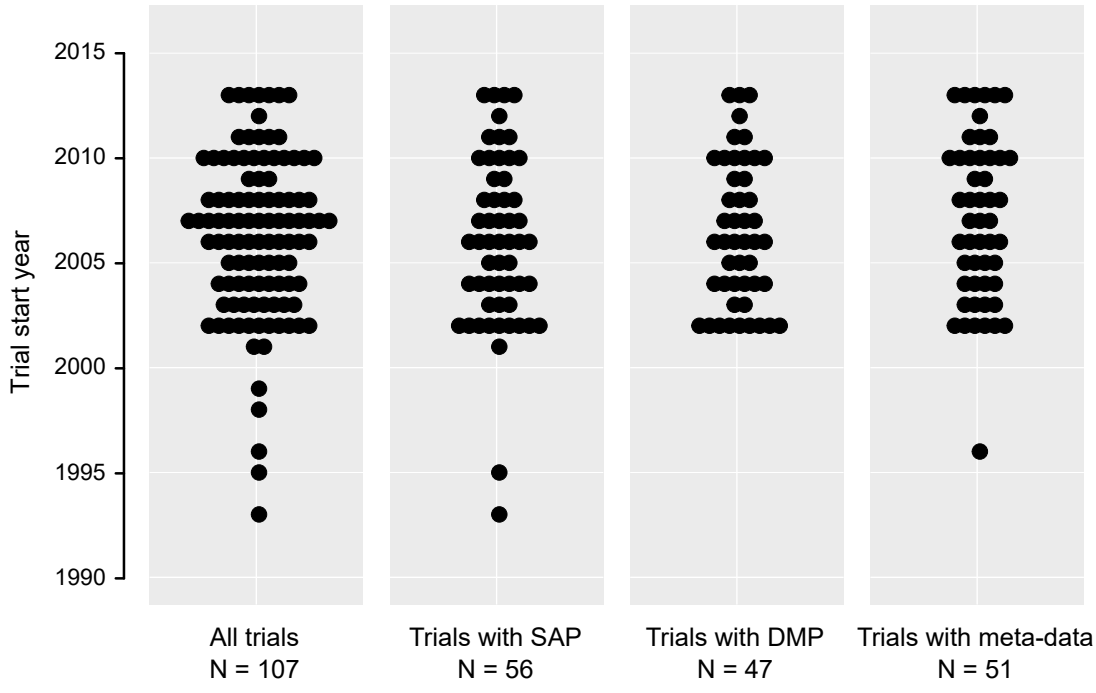


Figure 7.4: **Time distributions of the overall sample, and of sub-samples with different documents available**. Each dot represents one trial in the overall sample ($N = 107$); the trial start year was unknown for one study, hence the reduced sample sizes.

Data documentation fitting the purpose of a variables' dictionary ('meta-data') was available for 52 studies (Figure 7.3), and mostly as a separate file (37 studies or 71.2% of those with meta-data). Alternately, meta-data consisted in dataset annotations, programming specification files, annotated CRFs or similar data descriptions. An additional 11 studies had less standardised forms of data documentation, insufficiently detailed or structured to fit the purpose of a variables' dictionary.

I identified data management sections within the protocol, report (or in one case, article) of 48 studies. As a comparison, similar statistical analysis sections existed in 55 protocols or reports, and the contributors of two more trials shared comprehensive SAPs as standalone documents. Although there were overall few studies that started before 2002, it is noticeable that none of these had a DMP, and only one (started in 1996) had meta-data (Figure 7.4).

Data management tools and database formats

The data digitisation method was specified for 58.3% of the trials, and relied on single- (23 trials) or double-entry (38), or on scanning of paper forms (2). No study reported the use of direct electronic capture, possibly due to the newness of this data collection approach. There was variety overall in the software used for data management (known for 62 studies) or statistical analyses (known for 89 studies), but no clear temporal trend in the choice of software programme (Figure 7.5).

Among trials reporting the software used for data management, Epi Info^{TM} was the most frequently cited (used in 31 trials or 50% of those reporting this

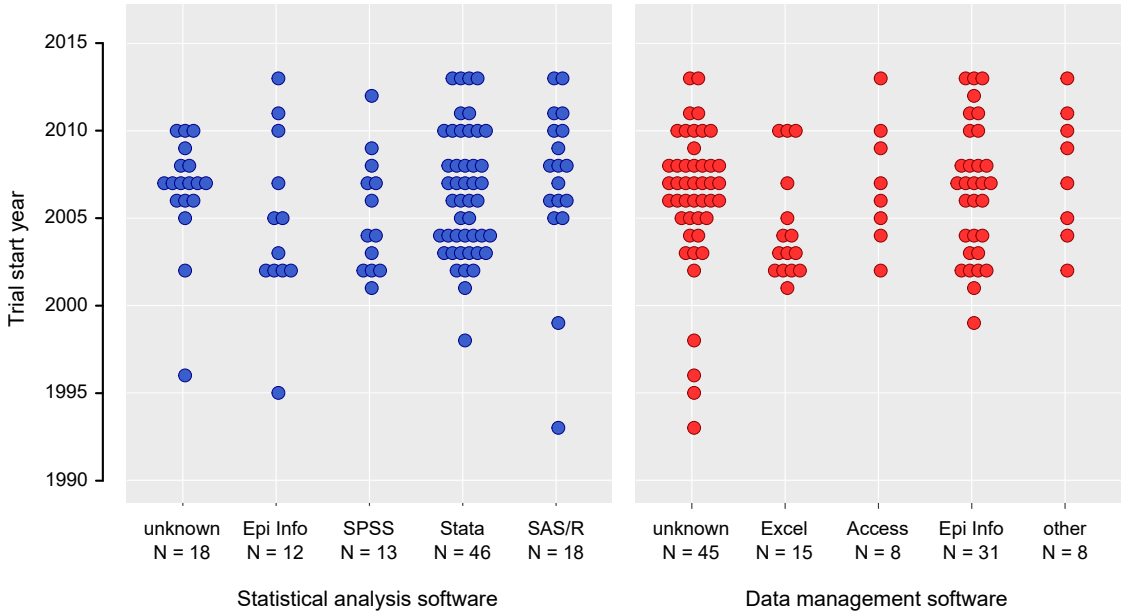


Figure 7.5: **Time distributions of trials by software used for data management and statistical analyses**. No significant difference in the year means and distributions by software – ANOVA outcomes: F-statistic = 0.952 (p-value = 0.437) and 1.114 (0.354) for the statistical analysis and data management software, respectively. 'Other' software cited for data management include OpenClinica, REDCap, SAS®, Stata® and SPSS®.

information), followed by MS Excel® (15 trials or 24.2%) – despite the fact that the latter software is not recommended for research data management owing to the difficulty in tracking changes [230], with inherent risks of mistakes (see for example, the 'horror stories' documented by the European Spreadsheet Risks Interest Group in and out of research settings [231]).

Stata® was the preferred statistical analysis software, with 46 trials reporting its use for that purpose (Figure 7.5). Importantly, only three of the 16 studies relying on SAS® for statistical analyses also specified the data management software they had used (SAS®, MS Access® and Epi Info^{TM}, respectively). SAS® is generally suggestive of a trial sponsored by the pharmaceutical industry. The private sector

frequently relies on a specialised unit or contractor to handle data during the study lifecycle, as explained by some data contributors from pharmaceutical companies whom I approached to request access to DMPs. In these cases, it is likely that data management rests on specialist software; thus the coding of the 'data management software suitability' variable (Y_{sw}) in consideration of the reported statistics software as well (as specified in methods, Table 7.1).

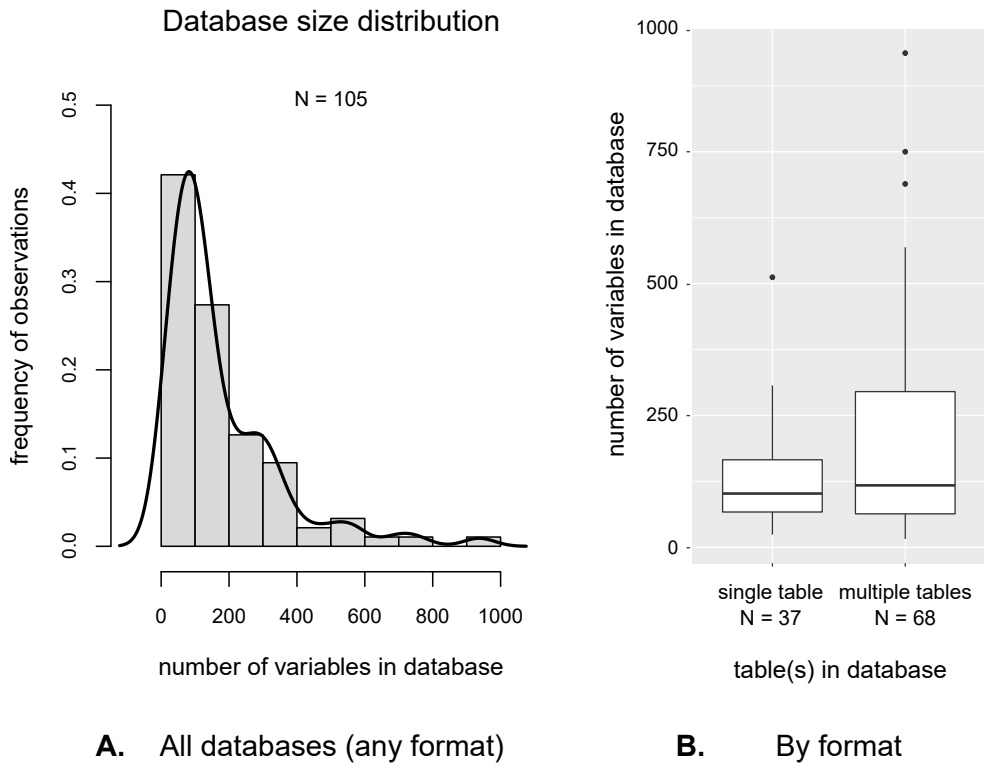


Figure 7.6: **Structure of shared databases**. N: sample size for the corresponding subsample. Three of the raw databases could not be opened in their original format, thus preventing assessment of their initial structure.

In terms of structure and format, most databases (63.0%) comprised more than one data table, but the total number of variables in databases varied greatly across

trials (Figure 7.6). This suggests different approaches when breaking down and organising the rather similar data collected for the purpose of these antimalarial efficacy trials. However, it is unclear whether contributors share a database that reflects the 'raw' locked trial database; or the final analysis dataset (or something in-between), which likely has a different structure and contains variables derived from source data for analytical purposes. Observed file extensions suggest that analysis datasets are the ones most frequently provided (as 53 trials shared a .sav or .dta file, suggestive of a SPSS® or Stata® dataset), although 39 databases were .xls (MS Excel®) files, which could be exported from various initial software.

Eventually, cross-referencing information from all data sources enabled me to infer the software used for data management in 84 studies. DMP, dictionary and software scores were then computed for each trial, as available. Two-by-two correlation tests between these three data management indicators highlighted a statistically significant correlation between the DMP and software scores (Fisher's exact test, p-value = 0.013), but no correlation with the dictionary score.

7.3.2 Correlations (or lack thereof) between study attributes and data management indicators

In light of attributes that could be 'quantified', I looked for potential correlations of data management indicators with the funding source(s), number of treatment arms, safety objective for the study, and cohort size. Appendix B.3 specifies other attributes of the study and contributors whose influence on data management strate-

gies and outcomes I intended to explore, but which I abandoned owing to subjectivity in interpretation of the data extracted from study documentation, or owing to inter-correlations between covariates.

Data management strategies and outcomes across funding bodies

Altogether, the 108 trials took place in 30 distinct countries, most of which were in Western (49 studies), Eastern (46) and Middle (22) Africa (macro-regions as defined by the United Nations [232]). Regional categories are non-exclusive, considering that 12 of the 15 multi-country studies were also multi-regional. Furthermore, the study region correlated with the funding type, as 16 of the 20 studies involving one private partner (whether as the single sponsor or as part of a public-private partnership) occurred in at least one country from the Western Africa group.

Overall, antimalarial trials in this sample were mostly funded by the public sector (54.6%), followed by the non-profit (36.1%) and private (18.5%) ones. Categories are not mutually exclusive, as some trials reported several funding sources. **Médecins Sans Frontières** (**MSF**) and/or Epicentre funded 15 trials (13.9% of the 108 studies), making this organisation the most frequently observed source of support, ahead of the WHO and associated bodies (for example, TDR) with 12 trials (11.1%). Of note, I initially set out to collect information regarding the sponsor, funder(s), and affiliation of the principal investigator(s) separately. In practice, these three types of study actors were often difficult to distinguish; hence, I focussed on funding sources, which were the most transparently reported piece of information.

	Public	Non-profit	Private	mixed
DMP score (N = 43), Fisher's exact test p-value = 0.0927				
low	4.7% (2)	9.3% (4)	0% (0)	4.7% (2)
medium	11.6% (5)	14.0% (6)	11.6% (5)	7.0% (3)
high	23.3% (10)	2.3% (1)	4.7% (2)	7.0% (3)
Dictionary score (N = 49), Fisher's exact test p-value = 0.2183				
low	12.2% (6)	4.1% (2)	2.0% (1)	2.0% (1)
medium	18.4% (9)	10.2% (5)	0% (0)	10.2% (5)
high	8.2% (4)	16.3% (8)	6.1% (3)	10.2% (5)
Software (DM) (N = 75), Fisher's exact test p-value = 0.1213				
MS Excel®	4.0% (3)	6.7% (5)	1.3% (1)	4.0% (3)
MS Access® or Epi Info™	30.7% (23)	8.0% (6)	6.7% (5)	8.0% (6)
other	8.0% (6)	8.0% (6)	6.7% (5)	8.0% (6)

Table 7.2: **Observations of DMP, dictionary and software scores by funding type**. The funder was unknown, or of unclear type, for 10 studies, thus the reduced sub-sample sizes. Numbers in brackets are counts of studies. Funding was categorised as: (i) public, for studies funded by a governmental body, such as a Health Ministry or National Research Council; (ii) non-profit, for studies funded by a charity or private foundation not seeking to generate profit; or (iii) private, for studies funded (and often, sponsored) by the pharmaceutical industry.

There was no statistically significant correlation between the type of funding and any of the data management scores (Table 7.2). Nevertheless, the public sector frequently had high DMP scores. In practice, publicly funded studies may have detailed data management sections in their shared protocol (as suggested by the present data), but possibly no separate DMP. In contrast and as suggested when communicating with data contributors from pharmaceutical companies, data management units or contractors are likely to hold detailed DMPs for studies sponsored by the private sector, and these documents are unaccounted for here.

Data management strategies and outcomes by trial design and cohort size

The 108 trials included turned out to be quite homogeneous in their design and objectives. The number of treatment arms was one of the few 'complexity' indicators showing variability across the sample, but it did not correlate with scores on any of the three data management indicators (Table 7.3).

	Treatment arms			Clinical data	
	1	2	3+	efficacy	eff. + safety
DMP score	(N = 46), p-value = 0.368			(N = 48), p-value = 0.010	
low	0%	15.2%	6.5%	12.5%	8.3%
medium	8.7%	28.3%	6.5%	12.5%	33.3%
high	2.2%	19.6%	13.0%	2.1%	31.2%
Dictionary score	(N = 47), p-value = 0.471			(N = 52), p-value = 0.708	
low	2.1%	10.6%	4.3%	5.8%	13.5%
medium	10.6%	17.0%	8.5%	13.5%	25.0%
high	4.3%	34.0%	8.5%	19.2%	23.1%
Software (DM)	(N = 80), p-value = 0.215			(N = 84), p-value = 0.012	
MS Excel®	3.8%	8.8%	5.0%	11.9%	6.0%
MS Access® or Epi Info™	5.0%	28.7%	20.0%	13.1%	39.3%
other	5.0%	20.0%	3.8%	8.3%	21.4%

Table 7.3: **Observations of DMP, dictionary and software scores by trial design and objectives**. P-values are from Fisher's exact tests for the corresponding sub-table.

In contrast, studies with safety data had significantly higher DMP scores and used software other than MS Excel® for data management (Table 7.3). This is possibly because the collection of safety data requires more thought regarding data management planning and analysis. Safety-related information indeed comes in

more varied forms, including qualitative data that must be appropriately coded and interpreted, and rapidly reported in case of severe adverse event.

In terms of 'size', trials with the lower DMP score or relying on MS Excel® for data management seemed more likely to have smaller cohorts (Figure 7.7). In contrast, single- and multiple-site studies were remarkably similar in terms of data management indicators (Appendix B.3), which further emphasises a likely size effect related to the number of participants rather than to the number of sites.

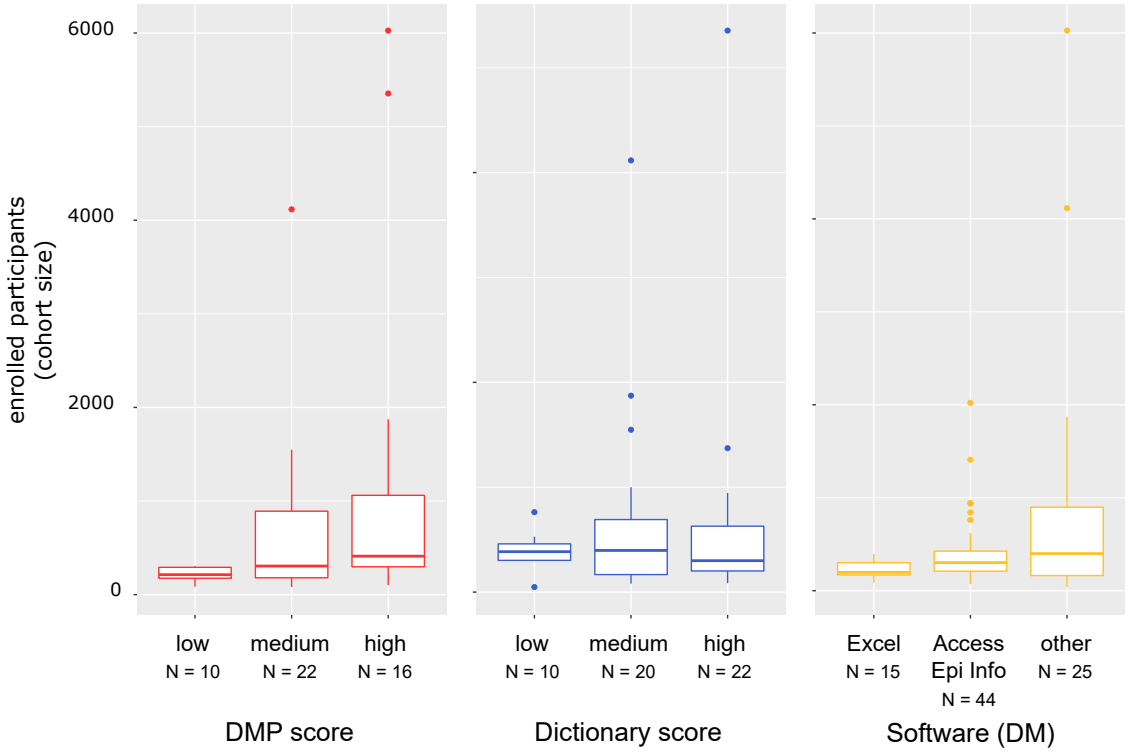


Figure 7.7: **DMP, dictionary and software scores by cohort size**. ANOVA outcomes: F-statistic = 2.379 (p-value = 0.104), 0.339 (0.714) and 3.869 (0.025) for the DMP, dictionary and software scores, respectively.

As a final note, in the absence of 'recent' DMPs and meta-data altogether (Figure 7.4), I did not investigate temporal trends in their contents. Furthermore, owing to the rare statistically significant effects uncovered in preliminary analyses, I did not run multivariate POLR on any of the data management indicators.

7.4 Discussion: Lessons from the exploratory analysis of WWARN data

This analysis systematically depicted the raw data and accompanying documents that are shared on a data platform (WWARN) with a focussed scope (antimalarial trials). Through this work, I characterised and partly quantified the variability in shared data. Despite the well-defined type of studies collated by WWARN, incoming datasets are far from homogeneous in terms of structure and format, and in terms of readability and comprehensiveness of descriptive documentation. I identified only few correlations between study attributes and data management strategies and outcomes, in part due to the limitations of the methodological framework (discussed in section 7.4.2), and in part because other factors I did not capture in this quantitative analysis must be at play; for example, the perceptions of those involved in data management. The findings reported here still point towards some areas in need of improvement regarding data management (and sharing), and suggest a certain randomness of 'who does what and why' when it comes to implementing a DMS.

7.4.1 Data management aspects in need of greater attention

This exploration of data management procedures, plans and meta-data surrounding shared datasets reflects and expands previous findings by others, who surveyed researchers to gain an understanding of their data management and sharing practices. In particular, Federer *et al.* [4] assessed the training needs of biomedical researchers in the USA regarding data management, and the gaps in knowledge they identified largely align with the gaps that transpire from the examination of data found in the WWARN repository. Among nine aspects of data literacy explored in Federer *et al.*'s study, respondents self-assessed as least competent with 'ontology' (understood as the selection and use of relevant standards or coding dictionaries to describe study data) and with 'data management plans' (including the selection of appropriate file formats, and the design of a plan for storage and preservation) [4]. Similarly and as detailed below, my findings suggest that training for malaria investigators (and others) is needed on the expected contents of variable dictionaries and DMPs, but also on the choice of appropriate software for clinical trial data management.

Variable dictionaries (or meta-data fulfilling that role)

The present analysis highlighted the variability in documents fulfilling the purpose of a data dictionary, which is to provide a detailed explanation and description of each variable in a dataset. I found meta-data for 52 trials (48.1% of the total sample), and over half of these documents scored only low (10 or 19.2%) or medium (20 or 38.5%) when assessed against desirable features of variable dictionaries. Moreover,

in response to my request for additional documentation, many contributors made enquiries about what I meant by dictionary. While poor phrasing from my part cannot be ruled out (see Appendix B.2 for a sample request), this difficulty with communicating the need for a data dictionary suggests a lack of awareness from investigators regarding the role and relevance of this document.

Of note, I found no correlation between software and dictionary scores, although more elaborate data management programmes (and statistical packages) enable dataset annotation. Style *et al.* thus report importing data to Stata®, in order to annotate data from a RCT in the field of maternal and child health [93]. The examination of raw datasets from the WWARN repository suggests an irregular use of such software functionalities to produce dictionaries embedded within datasets of antimalarial trials.

When surveying biomedical researchers in the pre-clinical and clinical sciences about their data-sharing practices, Federer *et al.* [7] found that most respondents (72%) claimed to include additional material when sharing research datasets. However, only about a third stated that they provide a data dictionary or codebook: 31.2% of scientific (pre-clinical) and 38.5% of clinical researchers, which aligns with my observation that 34.3% of the considered studies have a separate dictionary. Interestingly, in Federer *et al.*'s survey, the fraction of clinical researchers (46.2%) considering that their datasets did not require supporting information was almost twice as high as the fraction of scientific researchers (25.8%) with this same belief. The claim by clinical researchers that no supplemental documentation was necessary is particularly interesting, because none of the surveyed investigators considered that

their raw data was in a shape that made it ready to share, and they all reported spending some extra time preparing their data for the specific purpose of sharing [7].

Together with mine, these findings emphasise the need to raise awareness among clinical research staff regarding the importance of data documentation overall, and of variable dictionaries in particular. Beyond awareness raising, Tenopir *et al.* found a need to develop user-friendly tools to assist researchers in creating meta-data: researchers whom they asked about their data management and sharing practices felt most dissatisfied about existing tools, with no improvement noted between a survey conducted in 2009-10 [233] and in 2013-14 [234].

Data management software

The 108 antimalarial trials relied on varied software for data management and analysis; and the manipulations performed on the shared data since database lock were rarely documented in a clear manner. The distinction between raw and derived variables could easily be mentioned in dictionaries, and it was sometimes documented in database programming specification files.

One of the few significant correlations between a data management indicator and a study attribute was between software score and cohort size (Figure 7.7). Among trials for which a data management software could be inferred ($N = 84$), all those reporting the use of MS Excel® had recruited at most 400 participants. Larger studies were more likely to rely on 'specialist' data management software such as SAS®, REDCap or OpenClinica. Of note and as can be judged from the sparsely

reported information, software programmes used in the schistosomiasis and malaria research community for data management are similar, with a majority reporting the use of Epi Info^{TM}, followed by MS Excel® in both cases. In contrast, malaria researchers seem to favour Stata® over SPSS® for statistical analyses, with the reverse preference in the schistosomiasis community (Chapter 5, Table 5.1).

From the sampled data, it appears that evolutions in existing tools (especially, software and data capture technology) do not directly translate into changes in chosen software for clinical trials (Figure 7.5). Reasons for failure to upgrade to more adequate systems are not evident from the quantitative analysis, but changing practices may require an upgrade of people's skills; provided that they received data management training in the first place, which past surveys and reports suggest is not often the case for investigators [3, 4, 7]. Developing the desirable competencies and appropriate workforce takes time, and the impact of a possible change in practices following recent technological advances may not be observable over the relatively short period sampled here (1993 to 2015, with a large excess of post-2002 studies).

As a reminder, the DMP and software scores were two correlated indicators of data management. Similar trends in these two scores suggest that clinical research teams with the competency and resources to use 'more elaborate' tools (requiring more programming skills, as well as a network and client-server IT infrastructure [230]) are more likely to also be aware of data management requirements and write more detailed DMPs. The joint increase in DMPs and software scoring may also be suggestive of involvement of data managers and computer scientists in this aspect of the trial, leading to better data management practices overall.

Data managements plans and procedures

Finally, data management processes were only briefly covered in the majority of examined study protocols or reports, as 66.7% of the scored DMPs were incomplete (low and medium scores) for the few attributes I assessed; and it is largely unclear whether more exhaustive, independent DMPs exist for the concerned studies. Furthermore, hardly any of the study documentation covered data management personnel, and if at all, the latter point often focussed on statistics needs, specifying that an epidemiologist was involved in handling the trial data.

The situation regarding data management planning may improve over time (older studies in this sample had no DMPs), especially as funders and scientific journal publishers make requests for such documents more frequent and explicit [12, 19–24, 235]. However, the lack of detail in DMPs I could access suggests investigators still need to be made aware of what data management sections should cover, as some USA-based researchers also stated in Federer *et al.*'s study [4].

7.4.2 Threats to the validity and generalisability of findings

Three main aspects critically discussed below limit the validity and generalisability of findings from this quantitative analysis.

Methodological limitations and constraints due to available information

Although I sought to develop a systematic and comprehensive framework of analysis, the variables eventually retained in the data extraction instrument were largely constrained by the (limited) information reported in the documents I could access, and by the possibility to 'quantitatively' capture it (Appendix B.1). In addition, coding the 'qualitative' contents of study and data documentation in a systematic manner still involved some level of interpretation and subjectivity. A single analyst (myself) undertook data extraction and document scoring, and these activities occurred over a contained period of six weeks to limit variability. Nevertheless, the documents analysed to derive data management scores came in disparate and hardly comparable forms. This was particularly the case of 'meta-data', which among others came as a table of variables and their meaning in a standalone file, as annotations of the dataset readable when appropriately opened in a statistical package, or as a programming specification file embedding a description of collected and derived variables. For these reasons, the derived data management indicators explored here, as well as the characteristics of the study and its actors, remain imperfect constructs of what they are meant to represent.

In parallel, difficulty identifying and characterising the relative contribution of sponsors, funders and investigators in the trials' design also prevented me from drawing strong conclusions about who does what in terms of data management. The reported analysis focussed on funding sources for practical reasons; but in practice, and as suggested by disclaimers on different documents and/or articles, funders

rarely play a concrete role in study design and implementation, unless they also act as the sponsor (as most likely for studies funded by the private sector). As this doctoral work is concerned with capacity strengthening in LMICs, a more interesting comparison would be across locally-led trials, to clarify the barriers that LMIC-based investigators face on the ground when establishing a DMS. The available information on sponsors and investigators did not allow for this level of analysis; and in fact authors and investigators tend to be interconnected (jointly working on multiple studies, as illustrated by the network in Appendix B.3), thus making it difficult to disentangle who suggested which data management tool or process initially.

Confounding effects and sample bias due to sharing behaviour

In this analysis, the 'sharing behaviour' of the data contributor confounds the ability to assess data management practices. At least four reasons (listed below) may explain the absence of specific documents for a trial in the WWARN repository. In the absence of response from most contributors to questions seeking to clarify why they would not or could not share documentation, it is difficult to separate data management from data sharing practices.

Firstly, missing documents may simply not exist, although this seems unlikely for study protocols. One respondent to my request for a protocol nonetheless specified there was none, other than the brief document that they had created *a posteriori*, upon request from WWARN data curators. Secondly, data contributors may have lost the study and data documentation. Thirdly, the data contributor may not be

in a position to access or share the requested documents, as suggested by multiple cases of forwarding my requests to alternate parties. Data managers or statisticians sometimes referred me to principal investigators and/or sponsors to clarify the right to share; and conversely, investigators referred me to other colleagues more likely to hold the requested documents. Fourthly, contributors who have ready access to study and data documentation may simply not share it, for lack of time or because they fail to realise its importance for subsequent data users. The latter justification suggests limited awareness regarding the difficulties in curating and managing research data for re-use, an issue that training in clinical research methodology and data management should address. The time and effort that data sharing demands, including "the amount of extra work to respond to members of the research community who ... want explanation" [6] have been identified as deterring factors from the sharing endeavour [236], and they also likely influence which documentation has been shared in response to my requests.

The reasons for missing documentation listed above are interesting as they point towards many challenges of clinical research data management in the long run and within teams; however, they are not all suggestive of a poor initial DMS, which is what this analysis sought to investigate. The investigators' sharing behaviour may bias the sample of accessible documentation, as those who share may be more confident about their data management strategies and outcomes. They may also be those accustomed to international collaborations, or otherwise more engaged with WWARN. For instance, the fact that MSF/Epicentre studies are highly represented and rather well documented in this sample may not only relate to Epicentre's good

practice of producing detailed study reports, but also to the past affiliation of the WWARN director with this organisation. For this reason, Epicentre investigators may be more engaged with WWARN (the former affiliation of one WWARN key member with their own organisation creating some level of trust), or more responsive to requests for supporting documentation (when sent by a former colleague).

Likewise, past research has shown that biomedical researchers find "assisting a known colleague" a strong motivation for sharing (invoked by 49% of respondents to a survey on data-sharing practices in the USA) [7], and even those who do not commonly share data in repositories are generally willing to share with individuals that they know or trust [7, 236]. Researchers in LMICs also find it important to have a personal connection with people to whom they send their raw research data, especially if this data is to be shared before full publication of results [37].

Representativeness of the trial sample (and sub-samples)

In terms of patients included in clinical trials of ACTs, WWARN estimates holding 80% of research data generated worldwide, with the remaining 20% impossible to retrieve because of loss or corruption, which has been estimated to concern a large fraction of the global research data [160, 237]. Although this suggests that studies held in the WWARN repository provide a good picture of clinical research on malaria, it is important to note that this estimate reflects the number of patients included in trials of ACTs rather than the number of trials on any antimalarial, which is the relevant 'population' and unit for the present analysis.

In line with the objective of performing IPD analyses, WWARN data-gathering efforts may focus on larger trials, and it is possible that many small-scale trials account for the estimated 20% missing. In addition, applying the selection criteria for this analysis further reduced the representativeness of the WWARN sample of trials. Here, I considered only 108 trials performed in Africa, a small fraction when compared with the over 3,252 publications on any antimalarial trial identified in a recent search (Figure 7.1). Although not formally tested and although I did not notice strong biases in the sub-samples (underlying the analyses by data management indicators), the latter may be even less representative of the malaria community as a whole, owing to their small size and dependence on access to information.

Overall, I cannot claim that the trial sample I analysed provides an accurate picture of DMS in clinical research on malaria (hence frequencies and amplitudes of effects should be considered as trends and relative measures); but the uncovered inadequacies are still of relevance to pinpoint areas in greatest need of improvement, as previously outlined in this discussion and summarised in the box below.

7.4.3 Conclusion on the exploration of data management strategies and outcomes in malaria research

The quantitative analysis of data management strategies and outcomes in the WWARN community suggests that, even among individuals and institutions engaged in IPD sharing, challenges remain concerning the provision of meaningful supporting documentation, and concerning initial data management processes and tools.

Different types of clinical studies (especially, trials with different designs and objectives) may require different DMS; but with the exception of a safety objective for the study and of the cohort size, I could not identify correlations between study attributes and higher scores on data management indicators. In fact, running safety studies may require greater expertise in clinical research, which could explain greater awareness of data management requirements and apparently 'better' practices. Further work is needed to clarify the 'human factors' influencing initial DM practices (ahead and regardless of sharing), and to identify other influential factors not considered in this 'reductionist' quantitative analysis. This required a shift in the methods, and is the focus of the next (qualitative) part of this thesis.

KEY MESSAGES FOR CLINICAL RESEARCH DATA MANAGEMENT FROM THE WWARN PIECE

- The size and structure of databases underlying similar therapeutic efficacy studies vary greatly across studies, suggesting heterogeneous approaches to data handling
- Antimalarial trials enrolling more participants and involving a clinical safety assessment are generally supported by more robust DMS
- In contrast, other study design elements do not influence data management strategies
- Documentation and meta-data accompanying clinical study data vary in comprehensiveness and readability, which should be addressed to enable correct interpretation and easy curation of individual-level data
- Clinical researchers seem unevenly aware of the need for detailed DMPs and variables' dictionaries (in parallel, desirable contents of these are in need of clarification)
- In particular, the personnel involved is rarely described in protocols' data management sections, and emphasises statistical competencies rather than data management ones
- Epi Info™ and MS Excel® are widely used for data management in the malaria research community (outside of industry)

Part III

Qualitative work

Chapter 8

Introduction to the in-depth case studies

This chapter introduces the last piece of empirical research performed during my doctoral studies, relying on a qualitative, two-case study design. With this final piece, I investigate the perspectives of local researchers on clinical research data management practices in their organisation(s) and country. In this introductory chapter, I first justify the choice of design and I position the qualitative work within the wider context of my thesis (section 8.1). I then provide contextual information about the main institutes I visited (cases), and about the corresponding country and national research system (section 8.2). With these broad elements of context, I set the scene for presenting and interpreting the qualitative results in later chapters, where I will emphasise the data management aspects of research systems. Chapter 8 ends with a statement of the specific objectives pursued in this piece (section 8.3).

8.1 Motivations for a qualitative exploration of clinical research data management systems

The previous, quantitative pieces of research relied on indirect sources (published literature and shared data) to characterise data management practices and capacity-strengthening needs in LMICs. I examined documents out of their initial context, hence I could not fully reconstruct influences from the clinical study setting on data management choices. Performing qualitative research can be a way to complement partial answers provided by reductionist quantitative work [123, 238, 239]. Qualitative findings can illustrate, clarify or put "meat on the bones" of quantitative research outputs, which would remain difficult to interpret and transfer to practice without this enhancement and contextualisation [123, 238, 239]. In this project, two in-depth qualitative case studies enabled me to expand the picture of clinical research DMS strengths and weaknesses provided by quantitative analyses.

8.2 Case study sites and their contrasting national environments

The qualitative case studies and work reported throughout Part III took place in Dakar, Senegal (February 2017) and in Brazzaville, Republic of the Congo (May 2017). The rationale behind the selection of those sites is given in methods (Chapter 9, section 9.1). The retained case study countries contrast in terms of political

stability and development indexes (Table 8.1) [240, 241], and accordingly offer contrasting environments for clinical research. In this section, I describe the two sites I visited, in terms of their general organisation and of the clinical research work they typically conduct. This description focusses on the institutes that hosted me and facilitated my fieldwork, which constituted the 'cases' under study *per se*. This contextual information originates from background research performed before going to the field (notably, examination of institutions' websites and research outputs), then complemented and validated by participants' perspectives gathered on-site.

Indicator	Senegal	Republic of the Congo
Population (total inhabitants)	15,411,614	5,125,821
World Bank income group	low	lower-middle
GDP (billion USD)	14.684	7.834
Statistical capacity score	75.6	50
Life expectancy at birth (years)	67.1	64.6

Table 8.1: **Comparison of Senegal and the Republic of the Congo across selected indicators defined by the World Bank** [242]. Data is for year 2016 [243, 244]. The statistical capacity score is calculated on a scale from 0 to 100, by aggregating scores across 25 criteria representative of a country's ability to produce national (demographic) data [245, 246]. GDP: Gross Domestic Product. USD: United States Dollar, value current as of 2 February 2018.

8.2.1 'UCAD Parasitology' and its long history of medical education and research in West Africa

The first case study explored data management practices underlying malaria studies performed by (or in close collaboration with) the Department of Parasitology and

Mycology at the **University Cheikh Anta Diop** (**UCAD**) in Dakar – hereafter referred to as 'UCAD Parasitology'. When its School of Medicine opened in 1918, Dakar became one of the first places in francophone Africa where it was possible to study medicine [247]. In line with this well-established tradition of education in medicine (and shortly after, in other fields as well [248]), UCAD is a popular university attracting individuals from West Africa and beyond, including some of the Congolese participants I interviewed in Brazzaville. The current head of UCAD Parasitology has acted as the lead of WWARN's **West Africa Regional Centre** (**WARC**) since 2016 [249]. WARC has offices on the UCAD campus, where the few local WWARN staff also support UCAD Parasitology's activities. Its strong involvement with WWARN and associated exposure to data sharing make UCAD Parasitology an institution of choice to investigate clinical research DMS in LMICs.

Common research activities at the Parasitology Department include assisting the **Malaria National Control Programme** (Programme National de Lutte contre le Paludisme, **PNLP**) in evaluating antimalarial efficacy or the overall impact of **Seasonal Malaria Chemoprevention** (**SMC**). For countries with seasonal outbreaks of malaria like Senegal, WHO recommends this strategy of preventive mass drug distribution to at-risk children [250]. Senegal started implementing SMC in rural areas in 2012. Evaluation of malaria control campaigns relies on cross-sectional surveys to monitor key malaria-related and behavioural indicators (morbidity and mortality, use of LLITNs, prevalence of resistance markers among parasite populations in targeted areas, etc.); and on case-control studies testing the hypothesis that having received SMC reduces the risk of infection with *Plasmodium*. Researchers affiliated

with UCAD Parasitology are also involved in clinical trials. Indeed, at least eight trials included in the quantitative analysis presented in Chapter 7 had an investigator affiliated with UCAD, making this organisation one of the most represented in the sample after MSF/Epicentre.

8.2.2 'The Foundation' (FCRM) as a pioneer and catalyst for clinical research in Middle Africa

The second case study revolved around data management practices among clinical researchers in the Republic of the Congo. The **Congolese Foundation for Medical Research** (Foundation Congolaise pour la Recherche Médicale, **FCRM**) facilitated my visit and interviews in the country. This foundation was created in 2008 [251], with a mission to promote knowledge exchange and experience sharing among researchers in the Congo and in the region (Middle Africa), through the conduct of clinical studies and training of students, and through capacity-building initiatives aimed at health and research support staff. Most clinical studies performed in the Congo remain observational and cross-sectional in design, investigating biological factors associated with different disease outcomes. Clinical trials are usually restricted to therapeutic efficacy studies (phase IV) recommended and sponsored by WHO, to monitor continued efficacy of antimalarials [50, 252]; however, the FCRM is planning some earlier-phase trials. This pioneering role of the FCRM in a country with a recent history of performing clinical research makes it an interesting contrasting case to UCAD Parasitology.

8.3 Specific objectives of the qualitative piece

This qualitative research piece importantly contributed to the achievement of all three aims set for this thesis (Chapter 1, section 1.2). More specifically and as depicted in Figure 8.1, I here pursued the following objectives:

I. Describe the resources currently accessible for the setup and implementation of a clinical research DMS in selected LMIC settings, and understand the reasons for their availability or scarcity;

II. Describe local data management practices on-the-ground, and understand how and why they may differ from institutional policies;

III. Highlight involved staff's perceptions, knowledge and experiences of data management, as well as their perceptions of challenges and facilitating factors in the performance of their data-related activities;

IV. Explore the limitations and strengths of DMS in different LMICs (and propose likely causes for these), along with the associated constraints and opportunities for strengthening data management capacity in such settings;

V. Complement and re-examine the quantitative findings, by clarifying potential relationships (and interrogate their causality) between study characteristics and local choices in terms of data management practices.

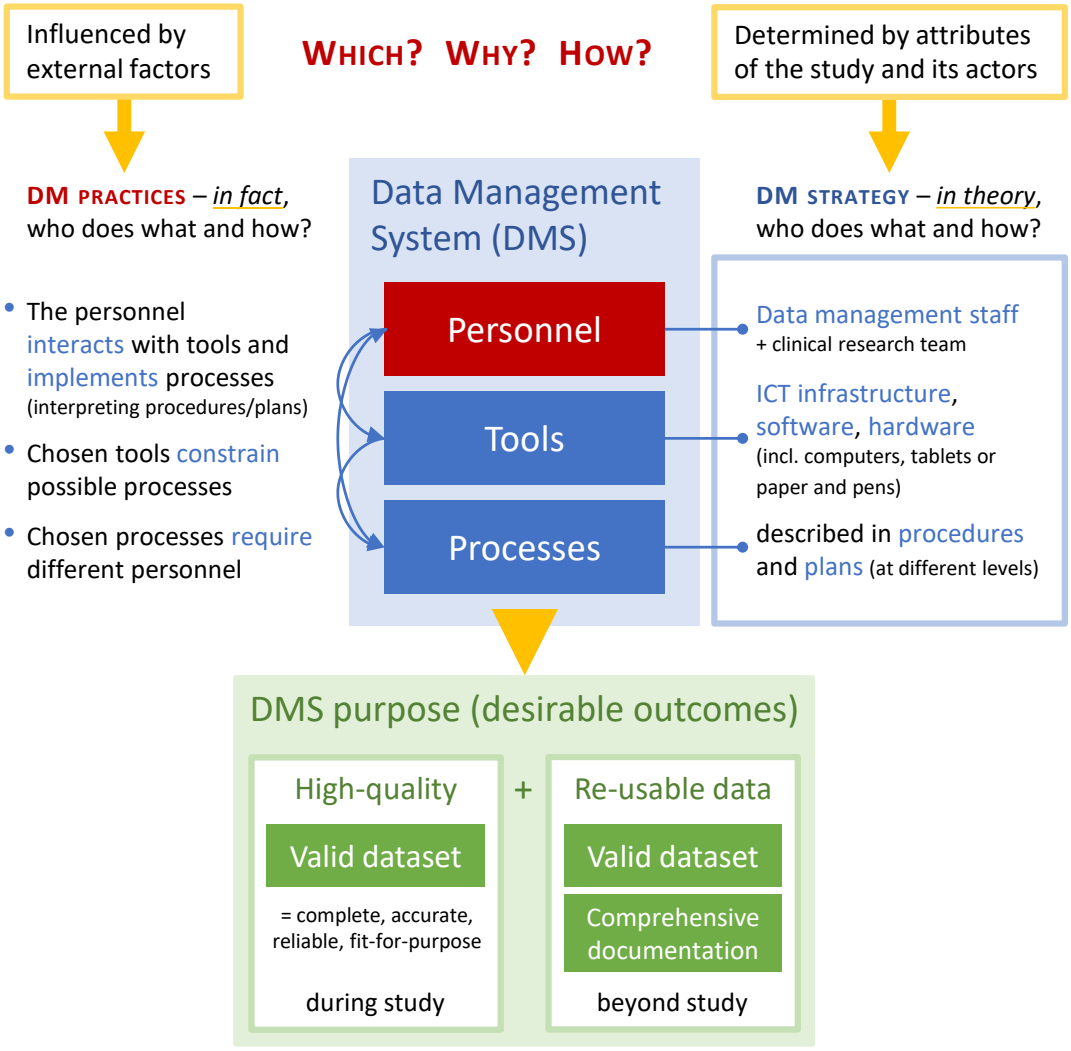


Figure 8.1: **Key components and outcomes of a Data Management System (DMS)**. The present qualitative piece of work particularly investigates elements highlighted in red. Interviewing and observing the personnel involved in data management and clinical research enabled the characterisation of data management practices, and the identification of external factors influencing these.

Chapter 9

Qualitative methods

This chapter describes the overall study design and criteria applied when selecting cases and participants. It also specifies the protocol I followed to generate and analyse data regarding the two case studies, and the adaptations I made in practice, in keeping with an iterative approach (typical of qualitative research [253]).

9.1 Sampling of cases and participants

Case studies are a design of choice when seeking to investigate a phenomenon in great depth, and within its real, 'natural' context [254]. A multiple (rather than single) case study design has value to compare DMS in different LMIC contexts, and to question the transferability of findings from one to the next. In line with the scope and timeframe of this mixed-methods project, I opted for a two-case study design, and restricted variability between cases and participants by focussing on

their experiences of conducting malaria research. I selected two research institutes as focus cases for in-depth investigation, and recruited participants within and around those institutions, as further outlined below.

9.1.1 Selection of case research institutes

Study sites were in part chosen purposively as contrasting cases, and in part through convenience sampling [254, 255]. For convenience, I drew on established WWARN partnerships and on the prior experience of my supervisors, whose long expertise in the malaria field permitted them to identify institutions that both matched my selection criteria, and where visits were likely to be productive. This pragmatic approach enabled me to secure important practical and logistical support from local co-investigators, facilitating access to ethical review and introductions to participants. Beyond convenience, study sites were chosen as contrasting cases: UCAD Parasitology is an African institution with a rather long history of performing clinical research locally, whereas the FCRM is a recent player. Additionally, the quantitative analysis highlighted that the WWARN repository contains few datasets from the Congo (regardless of institution) but varied data from Senegal (especially from UCAD); and it seemed important to clarify the reasons behind variable data management (and sharing) choices among UCAD researchers.

Other factors that influenced the selection of UCAD Parasitology and FCRM include the fact that they had explicitly approached WWARN in the past to partner in data management capacity-strengthening activities; and the fact that they were

both based in French-speaking countries. Data managers I met during the planning phase of my doctoral work (for example, at conferences) highlighted a particular gap in French resources and in networking opportunities for francophone data managers in Africa. They also suggested that national health information and reporting systems (hence possibly, clinical research DMS too) may differ in structural organisation between French- and English-speaking African countries, owing to diverging colonial influences. As I planned to explore only two cases, these perspectives underlined the value of limiting case variability by focussing on francophone African countries with a history of French colonialism. This selection criterion presented the added benefit of enabling me to perform interviews in my mother tongue.

9.1.2 Selection and recruitment of participants

Anyone working at the case institution (whether directly affiliated or occasionally collaborating with it) was *a priori* eligible to take part in an interview. Participants also had to have experience with some aspect of data management for a malaria-related study. Within this "sample universe" [255], I sought to recruit individuals in various roles and levels, to capture a large range of perspectives on clinical research data management. Staff who did not play a major role in data management aspects (namely, the planning, creation, processing and/or quality control of study data) were not eligible for in-depth, audio-recorded interviews.

Field facilitators identified and introduced me to most of the interviewed participants. Occasionally, interviewees suggested colleagues with whom I should talk

(snowball sampling [253,255]). After a few days on-site, I also became familiar with most of the permanent staff at host institutions, and I could accordingly arrange to meet and interview all key actors in the data management process. Eligible participants within case institutions were indeed relatively few, especially those working in a data management role. In contrast, when visiting institutions other than the 'case' ones, I only spent a short time there and rarely could ensure that I had met all data management staff. In these situations, I was more reliant on field facilitators, whose support however proved essential in obtaining appointments and in quickly building trust with individuals whom I met and interviewed on the spot.

9.2 Qualitative data collection procedures

The same set of methods served to generate qualitative data at both study sites, but the significance of one method over another varied according to case specificities.

9.2.1 Protocol adaptations in light of the situation at sites

Before my field visits, I anticipated conducting a mix of in-depth interviews and focus group discussions, with staff primarily involved in: (i) data collection or entry (for example, data clerks and research nurses); (ii) data management and quality assurance (data managers, computer scientists and statisticians); and (iii) study oversight (investigators and trial managers). Individual interviews allow capturing

of personal experiences and opinions, including on some potentially sensitive issues [253]. Group discussions would enable investigating how individuals (in a given clinical research role) may share or disagree with certain viewpoints, as well as under what terms a consensus may be reached [253]. Due to limited numbers of clinical research staff in a data management role found at sites, I however abandoned the idea of organising focus groups.

In contrast, the case study methodology unexpectedly evolved into an 'ethnographic approach'. Ethnographies are particularly useful to study interactions and group dynamics within teams or wider organisations [253, 256]. The present study does not constitute an ethnography *sensu stricto*, as the time I spent in the field (three to five weeks at each site) was limited compared to what this methodology generally involves. Nonetheless, I became immersed in the research teams whose practices I was interested in understanding; and the informal conversations I had with local colleagues became an essential source of information to shed light on DMS, alongside my daily observations.

9.2.2 In-depth, semi-structured interviews

Semi-structured interviews rely on open-ended questions encouraging participants to discuss personal experiences and viewpoints. Questions are 'loosely defined', in the sense that the interviewer has key topics they wish to address, but interviews may unfold differently from one to the next, to allow following up with interesting points raised by interviewees [253]. Appendix C.1 covers the topic guide and ten-

tative questions I used for all interviews, regardless of site and participant's role. Briefly, I invited participants to describe their background and current role, as well as the type of research projects they usually conduct or implement. I then prompted them to discuss their daily activities and perspectives on each step of the clinical research data flow (planning, collection and entry, processing and validation, analysis, preservation). I particularly encouraged them to depict the challenges and successes experienced. To conclude, I also asked for their views on training opportunities and capacity-strengthening needs regarding clinical research and data management.

Interviews varied in length, depending on the participants' experience with data management and on their answers' depth. In general, interviews lasted about an hour and took place in participants' offices, some of which were shared workspaces and did not allow for a private discussion. Most in-depth interviews were audio-recorded, but some participants preferred that I only took notes. This request often came from participants who seemed less familiar with my study, such as those interviewed at external institutions I visited only once.

9.2.3 Informal conversations and participant observations (ethnographic approach)

While in Dakar, I engaged in different activities with my study participants, including the delivery of a seminar and introductory course on data management and sharing (on request), aimed at postgraduate students from UCAD Parasitology. Data managers proved particularly willing to help me and to share their experience

throughout the duration of my visit; and I spent extensive time with them in and out of the office and work hours. Moreover, I had met one of them previously (at a TDR workshop we jointly attended), which further facilitated open conversations. I captured valuable information from my daily observations, impressions and informal interactions with participants by keeping a detailed journal while on-site.

Similar observations and informal conversations contributed to my understanding of DMS in the Congo. However, this second case study was less conducive to an ethnographic approach focussed on a single institution and its internal dynamics, mostly because individuals involved with FRCM jointly held responsibilities in other organisations (as I will describe in section 10.1). I thus obtained the richest data in the Congo through varied individual, in-depth interviews.

9.2.4 Ethical considerations and informed consent

The present research received ethical clearance from the **Oxford Tropical Research Ethics Committee** (**OxTREC** reference 5120-16), as well as from local institutional ethics committees (references 00000020/MSAS/DPRS/CNERS in Senegal and 012/CIE/FCRM/2017 in the Congo). In line with approved procedures, participants' informed consent was obtained before each in-depth interview (whether audio-recorded or not), and documented in writing through signature of a consent form. I explained my research project to the participants I approached, based on the information sheet accompanying the consent form (reproduced in Appendix C.1). No written consent procedure was used in the case of informal conversations and

observations, especially because they happened in a rather spontaneous fashion; but verbal consent from individuals involved was sought before taking notes related to information they shared on these less formal occasions. In one case where information seemed potentially sensitive, I sent my notes to the participant for review and editing before incorporating them in the analysis.

9.3 Qualitative data processing and analysis

9.3.1 Data preparation

Qualitative data came in two shapes: notes on my observations and informal conversations, and audio-records of in-depth interviews. I regularly revisited notes when analysing interview data and writing up results, to check the validity of emerging ideas and findings. I fully transcribed interviews myself, and I anonymised a subset of transcripts for translation by a contractor, to enable discussions with my supervisors. I performed analyses using the French (original) interview transcripts. Quotes reported in results are my own translations, as some meaning sometimes got lost in the translation proposed by the contractor, who was unaware of the context and tone of participants. Throughout Part III, italicised quotes (including within the main text) are direct quotations from participants.

9.3.2 Analytical strategy

Key themes discussed by participants and ideas they triggered were reviewed throughout the whole study lifecycle (from collection to write-up). The analysis thus was an iterative and continuous process, seeking to keep my incoming data and research questions in perspective [257]. I also had to produce preliminary reports and presentations for colleagues in Africa and Oxford. However, the 'formal', focussed analysis of interview transcripts only took place after I completed both site visits. By then, I had gained a strong command of the data, as I became very familiar with it when repeatedly listening to it throughout the interview transcription process.

In-depth interviews were analysed using a thematic analysis approach, and considering both cases jointly. I defined an initial qualitative coding scheme following familiarisation with the data. This scheme evolved as I analysed interviews and became immersed in the data, with the creation of new codes when new concepts arose, and the re-shaping and re-organisation of existing codes [258]. I regularly discussed findings with my supervisor who has expertise in qualitative research. This iterative coding was also influenced by my readings of literature on health systems, whose analytical frameworks drawing attention on social components of complex systems seemed informative for the present case studies, and will be discussed in Chapter 11.

The use of NVivo 10 [259] facilitated the data coding process. The final set of qualitative codes (listed in Appendix C.2) highlighted the key elements at stake in the functioning of local DMS, and formed the basis for the write-up of this qualitative piece.

Chapter 10

Results:

Challenges and opportunities in data management at research institutes in Africa

This chapter depicts data management practices at clinical research institutes in Senegal and in the Congo, as described by participants. I first present the individuals I interacted with and interviewed while in the field, in terms of their job titles, roles and affiliations (section 10.1). Information gained on-site from the participants' identification and recruitment constitutes a finding in itself, as it significantly deepened my understanding of 'who does what' in terms of data management in the two LMICs. Therefore, presenting participants directly leads to a discussion of the

personnel involved in handling data at different clinical research institutes, and of the varied training they received to carry out their assigned tasks (section 10.2).

Following this presentation of participants and existing HR, I describe data management processes at the two case study sites (section 10.3), in light of the challenges and opportunities provided by organisations and the available material (ICT infrastructure, hardware and software). I then outline key enabling factors and barriers to change in clinical research and data management practices in both countries, in line with participants' perceptions (section 10.4). Finally, I discuss some strengths and weaknesses of current managerial processes at the two sites, and of career development pathways for clinical research staff in the corresponding LMIC (section 10.5).

10.1 Participants recruited and institutes visited

In Dakar, Senegal

In line with my interest in clinical research data management practices, I spent most of my time in Dakar at the 'UCAD data centre'. The Parasitology Department informally established this data management unit in recent years; but data managers based there may jointly work with researchers from other UCAD departments, or with external contractors. In addition, partners from UCAD Parasitology occasionally use the data centre's workspace for their own data entry activities, thus making this unit a UCAD structure with rather fluid boundaries.

In line with the procedure reported above (section 9.2.2), I recruited eight investigators and research support staff from UCAD for audio-recorded, in-depth interviews (Table 10.1). An additional three participants agreed to give an interview, but preferred it not to be audio-recorded: two were data entry staff working at the data centre on recurring short contracts, and one was a postdoctoral researcher in entomology on a **Malaria Research Capacity Development** (**MARCAD**) grant. While on-site, I also encountered five Professors or Assistant Professors working on malaria at UCAD Parasitology, most of whom engaged in informal conversations or were otherwise involved with this case study. In particular, one Assistant Professor acted as field facilitator by introducing me to potential interviewees, and by providing background information on their role and organisation, thus helping to corroborate or question my early findings.

I visited three institutions outside of UCAD itself: the **Institute for Research and Development** (Institut de Recherche et Développement, **IRD**), the Dakar regional office of the **National Agency for Statistics and Demography** (Agence Nationale de Statistique et Démographie, **ANSD**) and the **Tuberculosis National Programme** (Programme National de la Tuberculose, **PNT**). I interviewed (without audio-record) one to four individuals in each place. The IRD was particularly relevant as the institution formerly hosting the 'data centre', and where many UCAD data managers explained having "*forged their experience*" before switching over to UCAD Parasitology. One UCAD data manager concurrently worked at the PNT, hence a visit there to gain an understanding of the different contexts within which a clinical research data manager may be simultaneously evolving.

Table 10.1: **Participants recruited for audio-recorded interviews at the University Cheikh Anta Diop (UCAD) in Dakar, Senegal**.

ID	Job title (and roles, if different)	Primary affiliation	Comments
DM11	Data Manager	UCAD data centre	Mostly working with PI81*
DM12	Data Manager	UCAD data centre	Mostly working with PI11; Repeatedly hired by WWARN and Tuberculosis National Programme
IT11	IT Technician	UCAD Public Health, ISED	Primarily affiliated with Public Health department, but available to support researchers from the Faculty of Medicine, Pharmacy and Odontostomatology (FMPOS) as a whole
SA11	MARCAD Programme Manager (Data Manager/ Statistician)	UCAD Parasitology	Employed to manage Malaria Research Capacity Development (MARCAD) programme; involved in data management and statistical analyses for PI11
RS11	Medical Doctor and PhD Research student	UCAD Parasitology	Working with PI11 on Seasonal Malaria Chemoprevention evaluation for Malaria National Control Programme; supervising laboratory activities (biological samples)
RS12	Research assistant	UCAD Parasitology	Epidemiology background (Masters level), assisting with data collection and validation in PI11's team
PI11	Medical Doctor and Professor in Parasitology (Principal Investigator)	University of Thiès	Recently appointed to new university; research activities still based at UCAD
PI12	Medical Doctor and Assistant Professor in Public Health (Principal Investigator)	UCAD Public Health, ISED	UCAD Public Health/ISED representative to Malaria National Control Programme, specialised in environmental health

*PI81 was a Medical Doctor and Professor with strong expertise in parasitology and epidemiology, who engaged in informal conversations.

Participant's identifiers (ID) reflect their main area of expertise regardless of title or role, to facilitate comparison across case study sites of perspectives from individuals with similar proficiency – DM: Data Management; IT: Information Technology; SA: Statistical Analyses; RS: Research (junior level); PI: Research (senior level).

Other abbreviations – UCAD Parasitology: Department of Parasitology and Mycology (Service de Parasitologie et Mycologie); UCAD Public Health: Department of Preventative Medicine and Public Health (Service de Médecine Préventive et de Santé Publique); ISED: Institute of Health & Development (Institut de Santé et Développement).

In Brazzaville, Republic of the Congo

The FCRM facilitated my fieldwork in Brazzaville. At the time of my visit, the FCRM had two major workspaces in this city: a laboratory and research office on the campus of the Faculty for Health Sciences at the University Marien Ngouabi (where I spent most of my time), as well as an administration and data storage place on the campus of **WHO's Regional Office for Africa** (**WHO-AFRO**). In line with its independent status of foundation and activities of supporting others in performing clinical research, individuals working at the FCRM often belonged to other organisations too. Moreover, many qualified Congolese investigators or statisticians frequently act as consultants for national and international bodies, including after they officially retired. Owing to this 'mobility' of researchers across organisations, individual practices became at least as important to understand as institutional ones, and I extended the anticipated case boundaries beyond the FCRM solely.

Tables 10.2 and 10.3 below list participants recruited for in-depth, audio-recorded interviews in Brazzaville, as well as the main institution(s) they worked with. I could visit many of these institutions when interviewing participants at their primary workplace. Informal conversations and observations involved another three participants usually based at the FCRM laboratory, including an Assistant Professor and two PhD students working on malaria. Of note, another few FCRM biologists worked on HIV/AIDS and diarrhoeal viruses, and were therefore not eligible for inclusion in my malaria-focussed investigation of data management practices.

In both cases, individuals in a job role restricted to data management activities were rather few. This situation limited the number of interviewees I could recruit. More importantly, this rarity of eligible participants reflects the overall HR for clinical research data management available in the two countries, which I further describe in the next section.

Table 10.2: **Participants recruited for audio-recorded interviews in Brazzaville, Republic of the Congo** (Part I).

ID	Job title (and roles, if different)	Primary affiliation	Comments
PI21	Biologist (Principal Investigator)	CERVE (retired)	Worked (and still does) as research consultant, especially with PNLP or WHO (for antimalarial efficacy trials)
PI22	Biologist and Lecturer (Principal Investigator)	LNSP + University	Research activities with FCRM and LNSP (head of Parasitology Service); teaching responsibilities at the University Marien Ngouabi's Faculty of Sciences and Techniques
PI23	Biologist and Lecturer (Principal Investigator)	LNSP + University	Research activities at LNSP; teaching responsibilities at the University Marien Ngouabi's Faculty of Sciences and Techniques
PI24	Biologist (Principal Investigator)	LNSP	Research activities at LNSP (head of Molecular and Parasitological Epidemiology Service, within the Direction of Preventive Medicine and Clinical Trials)
PI25	Pharmacologist and Lecturer (Principal Investigator)	IRSSA + University	Research activities at IRSSA (pharmacology)
PI26	Biochemist and Lecturer (Principal Investigator)	IRSEN + University	Research activities at IRSEN (biochemistry); currently advisor at Ministry of Scientific Research and Innovation
PI27	Paediatrician (Principal Investigator)	Private practice	Formerly Minister – Ministry of Health and Population, and Ministry of Higher Education; worked in both hospitals (Brazzaville General Hospital and University Hospital)
PI28	Paediatrician (Principal Investigator)	Hospital	Epidemiology expertise, also works as research consultant for international organisations

Table 10.3: **Participants recruited for audio-recorded interviews in Brazzaville, Republic of the Congo** (Part II).

ID	Job title (and roles, if different)	Primary affiliation	Comments
AM21	Project, Operations and Administrative Manager	FCRM	Primarily performing administrative duties at FCRM, also involved in ethics and quality assurance for clinical trials
AM22	Medical Doctor and PNLP Programme Manager	PNLP	National Malaria Control Programme (Programme National de Lutte contre le Paludisme, PNLP) manager
SA21	Statistician and Data Manager	FCRM	Provides support to FCRM researchers and multiple others (for example, students in PI26's laboratory)
SA22	Microbiologist and Lecturer in Biostatistics	Hospital + University	Retired; worked mostly as a microbiologist at one hospital in Brazzaville, as a lecturer at the University Marien Ngouabi's Faculty of Health Sciences, and as consultant in biostatistics for international organisations/others

Participant's identifiers (ID) reflect their main area of expertise regardless of title or role – AM: Project, operations, and administrative management; SA: Statistical Analyses; PI: Research (senior level).

Other abbreviations – CERVE: Centre for Vegetal Resources Studies (Centre d'Etudes sur les Ressources Végétales); FCRM: Congolese Foundation for Medical Research (Fondation Congolaise pour la Recherche Médicale); IRSSA: National Institute for Research in Health Sciences (Institut National de Recherche en Sciences de la Santé); IRSEN: National Institute for Research in Exact and Natural Sciences (Institut de Recherche en Sciences Exactes et Naturelles); National Public Health Laboratory (Laboratoire National de Santé Publique).

10.2 Establishing a permanent, institutional data management workforce

In this section, I describe the barriers and enabling factors that participants mentioned in relation to the establishment of a data management workforce within their institution and country. These factors mainly include the availability of workers with specific skills in data management or statistics, and institutional policies impacting on the ability to retain competent staff in clinical research teams.

10.2.1 In-country human resources to draw on

Participants highlighted the key role of data managers, biostatisticians and data entry clerks in properly handling clinical research data. In practice, the background and expertise of those hired in that function differed between Senegal and the Congo. There is general recognition in the field of clinical research capacity development that the role of clinical research data manager is in need of greater professionalisation (as reviewed in Chapter 11's discussion). Findings from the two countries I visited support that view, and suggest it also concerns the role of data entry staff.

Computer scientists, statisticians or epidemiologists may fulfil the role of clinical research data manager

Investigators and data managers at UCAD described that in Senegal, finding individuals with expertise in data management was not an issue. These individuals

perceived as 'expert' data managers were all skilled computer scientists. As an illustration, within UCAD Parasitology, the two data managers and the programme manager (doubling up as data manager and statistician) had primarily trained as engineers in **Information Technology** (**IT**). They described their IT background as a useful "*starting point*" (in the words of DM11); however, they saw further specialisation as a requirement to fulfil the responsibilities of a data manager and/or statistician in practice.

In parallel, the investigator at UCAD Public Health explained that in his department, "*almost*" everyone could execute data management tasks thanks to their background in epidemiology, which he associated with expertise in data "*analysis, collection and interpretation*".

This conflation between data management and analysis activities, and the allocation of data management tasks to individuals with a statistics or epidemiology background, was typical in the Congo. Two Congolese individuals (SA21 and SA22) were introduced to me as data managers by other participants, although they had received training in mathematics and statistics, and in epidemiology and public health, respectively. Accordingly, they described being involved in statistical analyses as much as (if not more than) in data management.

Moreover, participants I interviewed at FCRM and elsewhere in the Congo were not aware of any specialist clinical research data manager within the country. The FCRM statistician/data manager, whose statistics training was orientated towards

economics, explained that many of his mathematician fellows had pursued careers in business rather than in clinical research. He further explained that in the Congo, graduates with an IT background might become data managers (as I found in Senegal); however, they were more likely to move into the field of economics than health, and again to acquire their expertise through specialist training in statistics (made available for example by the World Bank) rather than in data management.

Overall, Congolese investigators felt that finding statisticians who could support them with both data management and routine data analyses is not an issue. They however considered that biostatisticians able to conduct "*complex*" analyses are relatively rare, and the two Congolese statisticians whom I interviewed indeed presented their data analysis competency as "*unique*" in the country.

Data entry clerks may professionally specialise in data entry, or perform this activity on the side

At UCAD, the two data entry clerks whom I interviewed performed data entry as a full-time job, following initial training in the area of project management and accountancy (for lack of a better translation). In line with malaria seasonality in Senegal, which is the main research focus within UCAD Parasitology, clinical research data entry is not required throughout the year. Consequently, investigators recurrently employ data entry clerks on short-term contracts. The latter group of staff explained that they work with other organisations in the meantime, including outside of academia and health research.

One Senegalese data manager thus described that the issue in accessing highly competent data entry staff is one of competition and timing. Less experienced (and less competent) data entry clerks are occasionally hired to fill in for their more qualified counterparts, which has consequences for the accuracy of entered data:

> "*the problem is to find a qualified, available data entry clerk ... we've noticed that some of them actually have a lot of experience, and so pay a lot of attention during entry. And there are others ... [who] make a lot of mistakes. And others, who are not accessible because they are currently working elsewhere*" – DM12, Data Manager (UCAD data centre)

Other individuals acting as data managers at UCAD Parasitology had a clear preference for EDC, and as a result did not discuss the challenges of hiring data entry clerks. Nonetheless, one of them considered it preferable for data entry clerks to have a background in IT, because this helps to grasp the logic behind edit-checks and questionnaires' structure, and to trouble-shoot the data entry application.

Data entry staff in Senegal did not voice a need for greater familiarity in IT, but a strong wish to become involved in data collection, which related to the perceived threat of EDC to their current jobs (further discussed in section 10.3). They considered that their experience with data entry has prepared them well for data collection, as they are acutely aware of the necessity to fill forms in a transparent and legible manner. Indeed, Senegalese data entry clerks said that deciphering poor handwriting of medical jargon is one of the key challenges of their work, which they address using tools with word spelling predictive capability (for example, Google).

In the Congo, the data manager at FCRM and several investigators from other institutions took a very different approach, describing data entry as not being a formalised or professional role. Instead, this task is largely performed by students or young people who have yet to establish a career, and whose lack of data entry experience often results in high rates of mistakes and significant amounts of time spent on validating data entries:

> "*we don't have people really equipped for this type of job [data entry]. Generally, we take the students who are at the Faculty of Medicine ... It takes a lot of [quality] control time ... when we check, we have a lot of [mistakes]. It's like you're attempting to redo the same job ... that they already had to do*" – SA21, Statistician and Data Manager (FCRM)

In fact, among Congolese investigators, students appeared to offer a key opportunity to address HR shortages in clinical research within and beyond data entry and management. Biologists at the **National Public Health Laboratory** (Laboratoire National de Santé Publique, **LNSP**) and independent clinical investigators (paediatricians) suggested this situation is mutually beneficial, as it gives students valuable work experience:

> "*I was lucky because, almost all my Masters [students] who come for short durations have been involved with the project ... so that's what gave me an advantage, to get them involved directly in [data] entry. Otherwise, I couldn't*" – PI23, Biologist and Lecturer (LNSP and University Marien Ngouabi)

> "*in terms of human resources, we don't have too many. So we take advantage of the arrival of these youngsters, to whom we provide a complete training*" – PI24, Biologist (LNSP)

The last strategy to support data entry at the LNSP was the use of general secretaries, to enter data themselves or to oversee students' data entry work.

10.2.2 Constraints and opportunities in staff recruitment imposed by organisations and project cycles

Organisational policies were raised as one area influencing the establishment of a sustainable research workforce, with particular implications for data management. UCAD Parasitology is embedded within a relatively rigid structure, imposed by a long-standing university. The latter only recognises specific contracted staff roles (for example, researchers, lecturers, administrative personnel, laboratory/IT technicians), excluding the role of data manager. One data manager at UCAD Parasitology also mentioned a policy requiring that a certain proportion of fieldworkers are recruited amongst healthworkers at study sites, which certainly impedes the diversification of data entry clerks' portfolio of activities.

In that context, UCAD investigators hire data managers by drawing on project-specific needs and funding, whose short-term duration is a challenge for the continuity of data management teams. Coping strategies include repeatedly seeking new project funding, moving staff between projects, or merging data management with other roles:

> "*The difficulty is being able to keep them [data managers]. Because ... we are within a university structure ... [where] the level [role] of data manager rarely even exists ... we must recruit and pay them using external funds ... given to us by research projects ... of limited duration. So we can find them, but we can't keep them ... A lot of manoeuvring is necessary to try and pay them via other projects, to keep them for as long as possible*" – PI11, Medical Doctor and Professor in Parasitology (University of Thiès)

Figure 10.1: **The case of the MARCAD programme manager**: An example of "*manoeuvring*" to retain data management staff, and some associated benefits and drawbacks.

As part of the need to manage MARCAD funds and activities, investigators at UCAD Parasitology see an opportunity to hire certain cadres of essential administrative support staff – including, but not restricted to, a *"programme"* manager (SA11). As explained by this participant, such MARCAD administration staff may be expected *"to get involved at the same time in different projects"*, and potentially in different areas of activities too. For the programme manager, this means handling data management and statistical analyses for clinical research studies on top of his MARCAD administrative duties.

This coping strategy enabled him to become a permanent member of staff at UCAD Parasitology, rather than a consultant in statistics (as was formerly the case). He felt this situation has benefits for data quality, not least because colleagues now have more ready access to ask him questions throughout the study process. However, he noticed unanticipated drawbacks of having to manage MARCAD on top of statistical analyses, which others at UCAD Parasitology are used to task him with. Notably, this work overload induces delays in data validation and analysis, which in some occasions even made investigators wonder whether something had failed during data collection and processing.

The situation of the MARCAD programme manager at UCAD Parasitology best illustrates such "*manoeuvring*" towards having a functional, long-term clinical research support team at the departmental level. This participant explained he is involved in data collection, analysis and preservation (which he described as "*[his] real area of expertise*") as much as in project management (his "*main role*"). While overall, this dual role benefits the institutional DMS (see Figure 10.1 for a detailed account), the concerned individual also noted negative consequences (especially, delays) of his juggling with two different sets of responsibilities:

> "*I haven't finalised everything [data validation and statistical analyses] yet because of my [programme manager] work … [In the meantime] I've placed temporary data, so that no one says 'but wait a minute, maybe it [data collection, with EDC] did not work?'*" – SA11, Programme Manager (UCAD Parasitology)

In contrast, the FCRM has a more flexible organisational structure, made possible by its independent status as a foundation and its more recent creation. In addition, one of FCRM's core missions is precisely to provide clinical research support to investigators. Indeed, the **European Development Clinical Trials Partnership** (**EDCTP**) provides substantial funding to the FCRM and the University Marien Ngouabi, which is one centre of excellence within EDCTP's **Central Africa Network on Tuberculosis, HIV/AIDS and Malaria** (**CANTAM**) [260]. The Foundation thus has an administrative workforce fulfilling research support activities for FCRM and other CANTAM partners, which is 'built in' the organisation's structure:

> "*The Foundation is a story, I can say of heart, of people ... So us here, the administration ... [if] you are a PI – it's all very well, but you can't do everything! So we help you to keep going. You are a young researcher, you won't have to pay the manager, you choose one of these [CANTAM] institutions ... they are the ones who will do [administrative and project management]*" – AM21, Project, Operations and Administrative Manager (FCRM)

As a final consideration on the local workforce, participants at both sites described resorting to collaborations with other institutions in Europe or the USA to address some limitations of their institutional capacity to handle large volumes of data or complex analyses, and to face "*the urgency to produce a report*" (in the words of the Medical Doctor and PhD Research student at UCAD Parasitology).

10.3 Available material and its influence on data management procedures

In this section, I turn to the procedures and materials underlying clinical research data management in the two countries. Practices that participants described relate to the following stages of the clinical research data lifecycle: pre-study/planning, collection (creating data), processing, analysis, and preservation. These five broad steps of the study and data management process appeared most helpful when coding and charting the qualitative data, to make sense of procedures in place at the visited sites (see Appendix C.2 for a detailed mapping of interview data, and Tables 10.4 and 10.5 below for a summary).

10.3.1 Similarities in data management challenges and opportunities across institutes and countries

Here, I first lay out some common strengths and weaknesses of DMS in the two LMICs as perceived by participants, before turning to the most distinctive feature of each site's DMS; namely, the widespread use of spreadsheet software (MS Excel®) in the Congo (section 10.3.2), and the on-going switch from paper to EDC in Senegal (section 10.3.3).

Table 10.4: **Data management process at clinical research institutes in the Republic of the Congo and in Senegal** (Part I).

REPUBLIC OF THE CONGO	SENEGAL
Pre-study/Planning	
• Efficacy studies of antimalarials rely on existing data collection forms designed by the WHO • For observational studies in a clinical setting (hospital), research students develop data collection tools with their supervisors (clinical investigators) • National bodies (demography institute and LNSP) re-use and regularly update questionnaires for cross-sectional surveys, in consultation with PNLP (demography institute) or clinicians (LNSP) • When using MS Excel®, the master database is copied and pasted on the researcher's personal computer; when using Epi Info™, a data entry mask is prepared	• The design of data collection forms (whether they end up on paper or on an electronic device) is a collaborative process, involving most clinical research staff (including data managers and fieldworkers) – in recent years, PIs learnt from past mistakes that involving data managers from scratch is critical • Experienced fieldworkers receive 48-72h study-specific training on data collection tool • Data managers create a database, using software according to data capture strategy (paper or EDC) and personal preferences – for example, MS Access®, SQL/MySQL (proprietary or open-source)
Collection	
• On paper forms, by students or individuals trained on study protocol and data collection procedure – training emphasises that "*missing data are not acceptable*" (SA22) • Forms are checked (especially for data completeness) ahead of data entry – by a dedicated supervisor/monitor in the field (before sending forms to a 'data centre'), or by investigator at central level (who contacts fieldworkers or participants themselves to address issues)	• On paper or EDC, by experienced fieldworkers recruited in Dakar and/or local healthworkers • Weekly to daily calls from investigators and research assistants to the local supervisors in the field + Monthly monitoring/supervision visits • On-site monitoring/supervision (before centralisation) involves data manager – critically important to "*understand the origin of errors*" and "*be able to bring solutions*" (DM11 and DM12)
Processing	
• Data entry on MS Excel®/Epi Info™ by students 'in charge' of the study, or by data entry clerks (secretaries, or students on short contracts) • Validation involves manually cross-checking paper forms against dataset contents, one by one – performed by PI or data entry supervisors (including, secretaries): or visual check of dataset contents to spot out-of-range/aberrant values (potentially with help of plots) – most often performed by statistician (data manager) • MS Excel® spreadsheets from different data entry clerks or students are copied and pasted back to a master-file – the FCRM data manager and head of laboratories meet about weekly to check they hold the same version	• If using paper, data entry on MS Access® by professional data entry clerks on short-term contracts, under supervision of data managers/ research assistants to help resolve potential issues encountered by data entry staff • Data entry validation written procedures exist: a percentage of the dataset from each data entry clerk is cross-checked, and acceptable error rates are defined to determine when to redo the whole data entry event • Database harmonisation is challenging when surveys happen recurrently and over a long time – villages' structure may change, participants may move, etc. hence requiring adaptation in database fields and identifiers

Table 10.5: **Data management process at clinical research institutes in the Republic of the Congo and in Senegal** (Part II).

REPUBLIC OF THE CONGO	SENEGAL
Analysis	
• Additional data verification/corrections may be needed at this stage • Investigators perform basic analyses with MS Excel® (or occasionally, Epi Info™), and delegate complex/multivariate analyses to expert statisticians (using other statistics packages, for example SPSS®)	• Additional data verification/corrections may be needed at this stage – mostly due to poor version control of datasets • Basic analyses performed by research assistant/student; other analyses performed by expert statistician (using R and Stata®) + delegated to international collaborators (if high volumes of data and associated analyses)
Preservation	
• Diversification of storage: everyone involved in a study holds a copy of the dataset + additional copies on USB sticks, hard drives or CDs/DVDs, and as attachment to e-mails to oneself	• Online, centralised storage (Cloud) • Storage on local + external servers, with automatic backup • Manual backup copies on hard drives

Procedures reported for the Congo are representative of those at the Congolese Foundation for Medical Research (FCRM) and at the National Public Health Laboratory (LNSP); those reported for Senegal illustrate practices at UCAD Parasitology. Organisations with different standard data flows and quality control strategies, for example relying on routine Health Information Systems (HIS), are beyond scope and not covered here, but mentioned in Appendix.
EDC: Electronic Data Capture; PI: Principal Investigator.

The strong perceived need for local servers

The physical resources (ICT infrastructure and hardware) available to support local DMS in Senegal and the Congo differed in their reliability and amount; but in both countries, a large proportion of participants found the availability of servers inadequate at the institutional and national levels.

UCAD Parasitology has its own internal network, but the data managers and statistician operating the IT material specified that more "*powerful*" and additional backup servers are needed. They saw investigators as "*not ready*" (that is to say, not

particularly willing nor incentivised, for example by funders) to invest in upgrading current IT systems, which jeopardises future management and preservation of research data. As put by one data manager: "*people won't be able to work anymore*" without greater investment in robust data management material at UCAD.

Although they regretted the absence of local (institutional or national) servers, Senegalese participants could rely on the internet; and they utilised servers from private providers (MS OneDrive®) and collaborating universities (in the USA or UK) to store and synchronise their study data between devices. Despite some reticence (discussed in section 10.4), investigators appreciated this solution for the practicality of access to up-to-date data from anywhere, including at conferences overseas.

Clinical research institutes in the Congo also suffered from a lack of robust computers to use as servers. In fact, Congolese participants hardly ever relied on internal networks, and used alternative (manual) systems to exchange and update data files between laptops. The FCRM data manager explained that this disconnected ICT infrastructure hampers the use of "*dynamic*" (client-server) data management software, such as **Structured Query Language** (**SQL**) databases and OpenClinica. This constraint also partly explains why MS Excel® prevails as a 'data management' software; a situation that the same data manager deplored. Furthermore, he explained that the internet coverage in the Congo is insufficiently reliable to address this gap by using remote servers accessible via the web.

In the absence of storage and backup servers, Congolese investigators sought to preserve data by creating multiple copies of important files and diversifying their

physical locations. They described sharing datasets "*widely*" among those involved in the study, sending datasets as attachments to an e-mail addressed to themselves, or copying datasets on **Universal Serial Bus** (**USB**) sticks and portable hard drives stored at home. Congolese investigators found it critical to hold copies in different physical spaces: as put by a biologist at the LNSP "*it can burn over there*" in their "*not politically stable*" country. Furthermore, several investigators found that sharing data with an international organisation like WHO (perceived as trustworthy and well-resourced) was a more sustainable way to preserve data than sharing with national bodies, where they explained infrastructure and procedures are lacking for that purpose.

Data management across far apart spaces: a bottleneck in the data flow

In both countries, participants directly involved in data management emphasised the burden of gathering data from the site (where collection happens) to the 'central level' (where data entry and further processing tasks happen). The "*return*" of paper forms from collection site(s) to the data management centre was unanimously described as "*always difficult*". Difficulties primarily relate to the time-consuming back-and-forth of forms to resolve queries (mentioned by Senegalese data management staff) or to the large amounts of missing data (mentioned by Congolese statisticians conducting data management). Even when forms reach adequate levels of accuracy and completeness, participants agreed that converting 'paper data' into 'numerical data' remains a burdensome process. Data entry was seen as time-consuming, tiring and error-prone, and therefore requiring careful validation.

The latter validation process seemed even more burdensome in the Congo, because data is often entered in software with few built-in data quality control functionalities (MS Excel®), and because hired data entry staff often have little experience. Congolese investigators described manually checking form by form that data on paper matches data in the dataset; a process they qualified as "*fastidious*", "*very, very, very hard*", and "*not very efficient*" though "*allowing correction of certain things*". Other frequent strategies implemented by Congolese statisticians to control accuracy of data entry included a visual examination of the dataset, potentially with the support of plots, to spot out-of-range or otherwise aberrant values.

In Senegal, project or data managers expressed greater confidence in the tools and personnel involved in data entry. Therefore, they only verified a pre-defined percentage of the data entered by each clerk against source forms or registries, according to a written procedure. Exceeding the acceptable error-rate may nonetheless require redoing the whole data entry, a situation perceived as difficult though necessary:

> "*I found [data validation] difficult ... [when we had to] redo the whole [data] entry ... People take a lot of time to capture [data], and after validation we see that there are a lot of errors, so we have to redo everything. As we have procedures we must abide by them (laugh)!*" – RS12, Research Assistant (UCAD Parasitology)

Of note, statisticians in both countries mentioned spotting further data inconsistencies during analysis (including the fact that "*some persons took drugs before being born*"), thus suggesting that data quality control remains imperfect. Cross-checking source data (again) at this late stage was seen as burdensome, because the voluminous amounts of source paper forms are often stored in a 'remote' location.

This was the case in Senegal, where UCAD Parasitology and the IRD share data entry, management and storage responsibilities across their premises (or used to do so); but this particularly applies to the FCRM, whose laboratory (where data processing happens) and administration (where data storage happens) are located in opposite sides of Brazzaville.

10.3.2 The case of the Congo: Limitations of DMS predominantly relying on spreadsheet software

Most interviewees in the Congo mentioned using MS Excel® to manage (and often, analyse) their study data. Investigators preferred this software because it is easier to use for the "*non-[data] savvy*" research staff, but also because it has "*prodigious capabilities*" that "*one often does not realise*". Furthermore, the WHO has developed a pre-programmed MS Excel® spreadsheet allowing data entry, validation and analysis for therapeutic efficacy studies of antimalarials (in its eighth version as of April 2018 [49]). Although it does not address all pitfalls of data management relying on spreadsheet software (not the least, a lack of traceability in data modifications [230]), this WHO tool appeared to fill a real gap for local investigators. Congolese participants from different institutions invariably mentioned it as the reference tool for antimalarial trials.

Independent clinical investigators working and conducting studies with their patients in the hospital setting favoured Epi Info^{TM}; which was also the preferred

option of the statistician/microbiologist who used to work at a hospital (SA22). Advantages of Epi Info^{TM} over MS Excel® cited by its Congolese users include the possibility to constrain data entry and limit errors at this critical, error-prone stage. This is achieved for instance by adding value range checks on numerical variables, or by imposing a choice between pre-specified options for factor variables. Of note, their availability in French is a likely additional driver for the use of both MS Excel® and Epi Info^{TM}, as some investigators were aware of other tools but stated they preferred those enabling translation.

Many challenges that Congolese researchers face in implementing robust DMS and in ensuring constant data quality relate to the limitations of MS Excel®. All participants mentioned similar issues, and they generally agreed these issues were challenging to address; however, not all MS Excel® users were aware of the fact that using different tools would resolve several challenges. Most strikingly, one investigator explained at length the burden of having to enter data and create a new spreadsheet table for each piece of analysis he performed, without realising database management software already exists to circumvent this exhausting approach.

Spreadsheets can partly be programmed to constrain or control data entry; a feature of the WHO tool for antimalarial therapeutic efficacy studies that participants appreciated. From interviewees' responses upon probing on their data validation strategies, it nonetheless seems that data entry checks with MS Excel® remain of limited scope and infrequent use. Coupled with a lack of focus or tiredness from

those entering data, this free-field entry system can result in many inadequate or incomprehensible entries in study datasets:

> "*Students were so tired [mostly, because of the repetitiveness of the data entry task], to the point that [they] put a lot of abbreviations, which say absolutely nothing when you don't know the meaning*" – PI23, Biologist and Lecturer (LNSP and University Marien Ngouabi)

Finally, version control of the master database is an additional challenge associated with the use of MS Excel® and with the absence of an automatically synchronised file on a server. Participants (other than the FCRM data manager) did not recognise this as an issue, but I found the recurring copy-and-paste events from one file to the next as a threat to data quality. Indeed, parts of datasets held on computers of different researchers may inadvertently get erased or lost in transfer when collating them within one (MS Excel®) file.

10.3.3 The case of Senegal: Barriers and enablers of transitioning from paper-based to EDC systems

A key feature of DMS at UCAD was the recent introduction of EDC (relying on Android technology and **Open Data Kit** – **ODK** [261, 262]), as well as the ongoing implementation of a shared data management and storage system. Some data management staff at UCAD Parasitology are driving the adoption of these technological advances (especially, one pro-active change agent whose key role I detail in section 10.5); but opinions differed regarding their pros and cons.

Two of the three Senegalese interviewees with strong data management expertise (DM11 and SA11) had a clear preference for EDC. The third one expressed a more moderated view, emphasising that "*one shouldn't reject one system over the other; because at any time, one [system] may become better*" (DM12). Investigators and data managers however agreed that EDC results in significant gains in time, particularly since data collection and entry occur simultaneously. They also saw EDC as putting data quality at less risk (Table 10.6).

Table 10.6: **Major advantages and limitations of EDC in terms of data quality, as perceived by Senegalese participants**. Advantages mostly relate to the bypass of the error-prone data entry process, while limitations primarily include risks of data loss (especially where the ICT infrastructure for data transfer is unreliable).

PERCEIVED ADVANTAGES	PERCEIVED LIMITATIONS
• Electronic questionnaires can have mandatory fields or other constraints to limit missing or erroneous data directly at the collection point. • Data collected with EDC reaches the central level almost in 'real-time' (if the ICT infrastructure allows immediate transmission from potentially remote sites). • Data managers and statisticians can promptly review the data and communicate their queries to the field; thus bypassing the lengthy, back-and-forth exchange of paper forms between study site(s) and the data centre to chase missing data and resolve other inconsistencies. • Direct capture of data in a numerical format prevents illegible handwriting. • Data validation and processing are faster, as collected data is in a similar format as needed for analysis.	• Fieldworkers are still barely familiar with EDC technology and may fail to save data they have collected. • Tablets may get damaged before data is transmitted to the central server, especially if collection happens in remote locations where data cannot be transferred in 'real-time'. • Data transmission may fail, or incomplete data may be received: one data manager (DM12) cited the example of a study where one variable disappeared altogether from the dataset.

Participants in Senegal expressed mixed views about the cost-benefit of EDC. One data manager and one investigator at UCAD Parasitology specified that paper-based collection is much more expensive, owing to the high cost of data entry.

Nonetheless, others saw the upfront investment necessary to set up an EDC-based DMS as a major barrier to its wider implementation:

> "*often [the barrier] is the cost ... if they [investigators] have to buy 20 tablets for a study – the investment cost, they prefer buying paper*" – DM11, Data Manager (UCAD Data Centre)

Furthermore, the absence of a server to host data at the institutional or national level means that researchers have to pay for private providers to transmit and host the study data, which one UCAD Parasitology investigator qualified as an "*extremely expensive*" service. He and others also questioned the reliability of this service in terms of data security and integrity during transfer.

More generally, participants saw the risk of data loss as the greatest weakness of DMS relying on EDC (Table 10.6). Until reliable technological tools and infrastructure becomes available, paper-based DMS appeared more robust and sustainable, by enabling to refer back to 'physical' source forms during and beyond the study lifecycle. In addition, most participants at UCAD Parasitology remained unclear about when to perform on-site monitoring of data captured electronically, and they considered it problematic to skip this usual quality control step.

Eventually, participants at UCAD agreed that the decision to use EDC or paper for one study often rests on the sponsor; thus suggesting some expectation for funders and sponsors to bear the cost:

> "*if the sponsor imposes tablets, we take [some]; if the sponsor has nothing we have to work on paper*" – PI12, Medical Doctor and Professor in Public Health (UCAD Public Health)

Beyond considerations directly relating to data quality and cost, EDC was debated because of the upgrading of skills and reshaping of clinical research roles that it entails, which I further discuss in the next section, together with other challenges and opportunities of working as a team in the context of change.

10.4 Balancing change and continuity in clinical research DMS

Participants in both countries described current data management practices while expressing the need for change in one area or another. In this section, I explain how some individuals acted as pioneers and key drivers of improved practices. These individuals are essential in triggering change, or sometimes, in their own words, in "*imposing*" it, despite the challenges that accompany modifications in long-engrained habits. These participants emphasised that to achieve effective change, tangible evidence of the benefits associated with new habits or systems must accrue in a timely fashion. In addition, strategies for continuity must be in place, so that lessons learnt are retained and organisations can progress.

10.4.1 The role of change agents in improving local practices

From participants' accounts, I identified two major factors playing an instrumental part in organisational change: (i) individuals whose values and personal traits push them to contest the *status quo*, and (ii) the snowball effect of their early successes,

which encourages others to follow suit. In both countries, I met individuals demonstrating a proactive behaviour towards change, and a strong will to improve theirs' and colleagues' practices; but Congolese participants seemed more cognisant of the key role of their local 'change agents' [263, 264] than Senegalese participants were.

All participants who had worked with the FCRM director perceived her as an essential driver of progress in the local clinical research landscape, hence the necessity to involve her in future initiatives to strengthen DMS in the Congo:

> "*That's her who really brought research innovation to the Congo, one must realise that ... The Professor who can ... perhaps shake up the Ministry and get this kind of proposals through, that's [her]*"
> – SA21, Statistician and Data Manager (FCRM)

The FCRM project manager thus described that under the director's lead, all members of the Foundation have taken a similar change agent role, hence contributing to the development of a clinical research "*culture*" within the Congo. She (AM21) further explained that their joint efforts are progressively taking effect, despite initial "*battling*" against officials' and individuals' resistance to change, "*fear of what people might say*", or lack of interest in a data-driven approach to policy-making.

One Congolese investigator outside of FCRM also expressed his strong desire to establish "*[his own] lab*", despite associated risks and (out-of-pocket) costs. Students' enthusiasm rewarded his bold move, which also changed the mindset of other investigators regarding the feasibility of certain techniques in their LMIC setting:

> "*I really took this decision [to start my own lab], let's say in a self-induced manner ... But that's when people understood – 'really, so we can do that?' I've said 'absolutely!' ... We can do that ... we*

> *can do anything*" – PI26, Biochemist and Lecturer (IRSEN and University Marien Ngouabi)

At least one worker at UCAD demonstrated a similar, proactive attitude towards change, which proved essential in improving systems and practices among his peers. This programme manager described the tenacity he needed, in order to convince his team to use a shared drive for storage of study datasets and documentation:

> "*I've done [tried] it twice, it's at the third attempt that it worked ... one day when we were all gathered ... I've forced people, I've said 'Come on, you have to create [the folder] ... on the spot, [while] I have everyone.' And [I] showed the example*" – SA11, Programme Manager (UCAD Parasitology)

The team's principal investigator eventually reaped unanticipated benefits from the system proposed by his manager and got really "*excited*" about it (in the words of the programme manager), thus becoming a success example for other group leaders. Such opportunities for others to see the advantages of new systems are even more necessary when people do not find faults in current systems, in which case one Senegalese investigator suggested they "*tend to forget*" to strive for improvement.

10.4.2 Working as a team in an evolving environment

The recent introduction of EDC and the move towards shared (centralised) storage spaces in Senegal create an opportunity for streamlining data collection and management. While change agents are facilitators of the adoption of these new

technologies, others' attitudes and the shift in team dynamics that they induce may act as a barrier to the upgrading of data management practices and systems.

For example, centralised DMS handled by project managers or computer scientists can be seen as a threat to the remit of some roles, inducing unwarranted loss in 'data ownership' rights. The programme manager proactively advocating for data centralisation thus suggested that investigators' scepticism regarding systems they cannot control hinders change in data management habits:

> "*It's like not being the owner. It's rather difficult to understand, ownership ... As [I] manage the data ... to do something, it is necessary to get back here [to the system I've put in place]. But the idea ... is [to] have a tool, or ... a facilitator of the overall process*" – SA11, Programme Manager (UCAD Parasitology)

Further challenges of technological change in Senegal relate to unanticipated 'competition' between staff with different skills and preferences for old or new approaches. Operating data collection systems relying on paper and on EDC requires different sets of competencies, and not everyone in current data management teams have both. The investigator interviewed at UCAD Public Health explained his highly positive experience of using ODK for a handful of studies, but the challenges of rolling out EDC more widely, owing to a need to first upgrade programming skills in his department. Data entry staff at UCAD Parasitology felt particularly 'threatened' by the increasing use of EDC, as this risks putting them out of work. Such fears underpinned data entry staff's drive to widen their skill set, to include for example field data collection. As already mentioned, UCAD data managers expressed mixed opinions about EDC: one individual with particularly strong expertise in pro-

gramming (including, on Android systems and with ODK) had a clear preference for EDC, to which he attributed only benefits, whereas his colleague seemed more familiar with and attached to the use of MS Access® and paper. The programme manager who stated having "*forced*" some investigators at UCAD Parasitology to move towards EDC most clearly described the way that tensions might arise within teams as a result of this innovation:

> "*When data collection ... [is] via electronic means ... this is rather my speciality. If we collect paper data and we need to enter it, it's [another data manager] ... I have taken away part of his activities; it wasn't electronic before. So ... you diminish someone's work, it's like I take him his job (laugh). It's not easy*" – SA11, Programme Manager (UCAD Parasitology)

No comparable 'competition' was mentioned in the Congo, possibly because DMS there are still in their infancy, and need to be concretely established before the challenges of change are experienced within clinical research teams.

10.4.3 Documenting to ensure continuity

To balance out the challenges of change, whether this change is due to a technological advance or to other factors (for example, staff turnover), and to ensure continuity in clinical research projects, participants in both countries emphasised the importance of written procedures or similar documentation of current practices, systems and datasets. Having a new member join in the team can be difficult in the absence of such archives, as expressed by project and data managers, and as experienced by some at the start of their current job.

Areas suggested by one UCAD data manager as most requiring attention concern the documentation of database development, as well as the breadth and depth of existing documents, so that a newcomer can "*understand from procedures only what [he has] to do, and how to do it*". Another Senegalese participant expressed concerns over the version control of procedures and datasets, noticing that some of his colleagues simultaneously hold two different versions. More generally, the Senegalese staff with data management expertise mentioned difficulties with keeping track of files, owing to a lack of transparency regarding study activities and other on-going projects. They agreed that relying on investigators to handle their study data and other files individually would undermine sustainability:

> "*There are different things that people have in their computer, sometimes old computers ... But the new one who comes, he takes a lot of time before he knows. Even the seniors who have the things, they don't remember very well what they have, what's the last version*" – SA11, Programme Manager (UCAD Parasitology)

In fact, this observation was instrumental in pushing for the establishment of a centralised DMS, incorporating a transparent audit-trail system. In this case, technological change was necessary to enable continuity in clinical research practices, and to mitigate the challenges of poor communication regarding data files' modifications. The limitations of relying on individuals for sustainable data handling also emphasises the significance of institutional policies, which remained rare in the visited institutions as far as data management is concerned.

10.5 Developing oneself, developing others: barriers and enablers of career development

So far, I have emphasised the available human and physical resources, as well as on the processes and institutional structures underlying DMS in Senegal and the Congo, which I discussed along with the barriers and facilitators of their establishment or enhancement. In section 10.4, I started to explore social interactions and human behaviour, whose impact on DMS I further present in this final results' section.

10.5.1 The 'art' of people management

Here, I focus on the many challenges and opportunities of becoming a leader, managing a team and allowing co-workers to grow in their role.

Good communication and interpersonal skills as essential, under-developed competencies for managers

Senior investigators at UCAD Parasitology saw training in leadership and management as important to help clinical researchers progress to high-level academic positions. While they often could not benefit from such training themselves, they are trying to make it available to the next generation of African researchers. As I could observe when attending the MARCAD research methodology course while in Dakar, such courses for doctoral students and postdoctoral fellows include a large amount of taught sessions on 'soft skills' (see [265] for a review and discussion of

what these skills involve). Interestingly, one data manager at UCAD also referred to principal investigators as "*businessmen*", suggesting that they do not make much use of their research skills anymore: in his view, senior investigators are largely involved in seeking funding and allocating tasks or resources, but they do not handle the day-to-day technical operations of clinical research (for example, by going to the field or monitoring data quality).

In contrast, the FRCM project manager emphasised that different roles in clinical research and administration should not be confused. According to her, investigators must rely on competent managers, and become aware of that necessity, "*because you can't do both, you don't have the same competencies*". Congolese researchers actually did not express a wish for training in leadership and management, although this topic was not at the core of my interview questions and may have been overlooked.

Many participants in both countries, and particularly data managers, still described having to line-manage a team or otherwise handle co-workers, for which they drew on self-learnt, everyday-life strategies rather than on formal training. A data manager in Senegal explained that data monitoring activities require him to solve issues arising in the field with data collection staff, using tactful communication. The FCRM data manager described a similar approach to address challenges in overseeing data entry staff:

> "*You have to sometimes be there all the time ... to oblige them to pay a bit more attention ... Having an exact procedure to handle a case, it's difficult ... We adapt on the spot: with that person, I must behave in such a way to not irritate them; but at the same time I also want to push [them] to work well*" – SA21, Statistician and Data Manager (FCRM)

This participant also mentioned that studies involving partners external to the FCRM itself are always more cumbersome, owing to the more complicated processes that are necessary to coordinate everyone's activities.

Beyond the strict context of managing a team, good relationships were described as essential to circumvent burdensome (and sometimes, dysfunctional) communication pathways. The research student at UCAD Parasitology explained that, when performing the SMC evaluation, it is the PNLP's responsibility to inform healthworkers of an upcoming visit from UCAD researchers. This requires cascading information from the central, national level down to local communities, via intermediate health districts' divisions and sub-divisions. When UCAD research teams arrive at selected sites and realise that healthworkers have not been informed of their arrival, the possibility to conduct the study rests on researchers' ability to convince key local players, which they usually achieve by "*playing on the friendly relationships that link [them with locals]*" (RS11).

Leading and retaining ownership while delegating responsibilities

Good managerial support appeared to contribute to job satisfaction. FCRM staff thus described having open discussions with and feeling inspired by their leader, an important aspect to counterbalance the occasional stress of their job and to motivate them to take responsibility for their team's activities:

> "*I laugh a lot, why – because also I have someone behind me, who is trustful, who boosts me – surely who is not easy [soft]. Because unfortunately, everything when she is not [here] relies on my head [shoulders], and so must be properly done. It's not the others that*

> *we see, unfortunately that's me. So I must ensure that the success that we want to have, is efficient [effective]*" – AM21, Project, Operations and Administrative Manager (FCRM)

Outside of FCRM, investigators in the Congo often oversee data management and quality control themselves, partly because no data manager is present to take over these duties. However, this retention of control may also relate to concerns about delegating. Along with other investigators, one biologist at the LNSP expressed "*anxiety*" about "*always depending on someone even for little things*", as well as a need to "*reassure [herself]*" by personally cross-checking all data entries, despite the great burden of data entry validation discussed above. One Congolese participant even suggested that clinical researchers' reliance on others for data-related tasks (and especially, for planning and analyses) was an unacceptable situation:

> "*It is necessary that, themselves [clinical investigators], they are able to carry out [methodologically-sound analyses] ... [But, the fact] that they are dependent on those computer scientists and others, that's not working. That's it, I've said it!*" – PI27, Paediatrician (Private Practice)

In Senegal, investigators at UCAD Parasitology entrusted data managers with guaranteeing data quality, and did not seem anxious about delegating data tasks. Both data managers however emphasised the importance for specialist clinical research data managers to be involved in wider aspects of studies, something they said investigators at UCAD Parasitology only recently realised and facilitated. In particular, the Senegalese data managers found it essential for them to contribute to study planning, and to spend some time in the field at the data collection site(s):

> "*They [investigators] saw there were some issues in the field, and they understood it is important ... for data managers to be involved also in the [study] setup ... and [in the field], because when you go ... there, you understand the origin of errors ... [in] the data you handle*" – DM12, Data Manager (UCAD Data Centre)

From their experience, failing to involve a clinical research data manager in the design of data collection tools has led to omission of important fields in the database, or to duplicated participants' identifiers on forms, which prevented their matching to biological samples. Of note, only the first issue could be fully resolved, when the concerned data manager was sent to supervise data collection on-site, at which point he noticed and addressed the mismatch between forms and database structure.

Interestingly, and despite strong feelings about the importance of their own wider and earlier involvement in studies, data managers did not recognise or act upon similar requests from data entry clerks. The latter group of staff nonetheless wished to become involved in studies beyond their assigned data entry tasks, on the basis that this would improve the quality of data collection and entry.

Acknowledging contributions and sharing credit

Beyond their apparent involvement and responsibilities in clinical studies, the two data managers at UCAD Parasitology expressed a wish for their role to be "*taken [more] seriously*". They seemed disappointed with the attitude of some investigators towards data management, illustrated by the limited time and resources that senior clinical researchers allocate to this key activity. Senegalese data managers

thus expressed a wish for greater recognition in terms of salary and academic roles (including publishing opportunities):

> "*it would also be necessary to give more consideration [to the data manager] ... in terms of publications ... It's not a small job to stay on a database, and to work on it over days and days. And then ... PIs and others who've served for something – funding maybe, will be put at the front light ... We can't do much, this is politics (laugh) ... But we also want publications. To exist, you have to publish*" – DM12, Data Manager (UCAD data centre)

Data managers at UCAD Parasitology also valued informal marks of appreciation and career development opportunities. For example, one of them explained he enjoyed working closely with investigators on data analysis or other aspects of the study, which he viewed as a chance to build his capacity, especially in statistics.

A different perspective was shared by a Senegalese investigator, who described an incident when he felt he had given 'too much' freedom to a data manager, following which the data manager 'refused' to hand the data over for analysis. These varied perspectives illustrate the tensions around sharing tasks and credit within clinical research teams. Such tensions seemed more acute in Senegal than in the Congo, possibly because data managers in the latter country are biostatisticians, who receive greater credit for their contribution to statistical analyses.

10.5.2 Barriers and enablers to training

Interestingly, most participants expressed a wish for additional training in biostatistics, regardless of role or country. The only notable exceptions were four individuals

who already had this specific expertise: the programme manager (SA11) and the Professor in Public Health (PI12) in Senegal, and the two statisticians in the Congo (SA21 and SA22). Senegalese data managers saw statistics as the last competency to master, to enable them to handle the whole data flow from creation to analysis. These data managers and computer scientists thus perceived training in statistics as a means to grow in their roles, potentially up to "*becoming an investigator [them]selves*". Junior researchers in Senegal also considered statistics proficiency as essential for them to become fully-fledged investigators. Likewise, Congolese investigators at later stages of their career found that a lack of expertise in clinical research methodology and statistics limited the visibility of their work at the international level, hence stalling their career progress.

In the remainder of this section, I report the challenges and opportunities that participants encountered in accessing training to develop their biostatistics or other key data management skills.

Work overload and time constraints from lecturers...

Participants with statistics expertise in both countries voiced difficulties related to training others in this area. Primarily, the shortage of biostatisticians results in a teaching work overload for the few able to deliver such training. One Professor at UCAD Parasitology explained he could not face the demand when he started to offer regular seminars in statistics to postgraduate students, and had to stop this initiative. Similarly, the Professor at UCAD Public Health explained that

researchers in his department are "*too busy*" to devote much time to teaching on top of their research activities and other commitments, further emphasising that "*it's crazy*". Accordingly, the junior researchers and one data manager regretted a lack of opportunities to attend short courses at UCAD, and they mentioned statistics and advanced data management methods as areas of particular interest to them.

In absence of expert lecturers, Congolese investigators with good exposure to statistics, but who do not consider themselves a specialist in the topic, are required to train students. They recognised this situation as sub-optimal, although the best they could do to address the gap. Beyond the case of statistics, the FCRM project manager explained that the Foundation receives many requests from the University Marien Ngouabi to train students on specific biological and clinical research methods. The small laboratories of the FCRM cannot currently accommodate this high demand, despite a strong will to help:

> "*it always pains us a lot, to be unable to welcome all these persons who want our support for training*" – AM21, Project, Operations and Administrative Manager (FCRM)

Highlighting a context of limited training opportunities, most investigators and both statisticians in the Congo explained that they received a significant part of their training outside of the country, usually in Senegal and France. Key areas on which to seek training abroad were research methodology and statistics (for investigators) and data management (for the FCRM statistician).

...and from students alike

In parallel, the work overload and limited time of potential 'students' (especially, researchers and practitioners wishing to receive any additional training) was frequently raised as a high barrier to their own development. In the Congo, several investigators and the PNLP programme manager insisted they "*wouldn't have time to go and sit*" to follow a course. These individuals suggested online courses would be easier to fit in their schedule, although they acknowledged these would still require freeing some time for learning. In addition, some lecturers underlined that aspiring learners may not stick to online training without face-to-face support from a tutor, who would be essential to help them to overcome difficulties in understanding course contents:

> "*You need a tutor ... I know that sometimes people are put off by difficulties ... I know [people] who sign up, who take distance-learning [or] online courses, but who never finished them*" – SA22, Microbiologist and Lecturer in Biostatistics (retired)

In both countries, blended learning (that is to say, the mixed use of eLearning and face-to-face sessions [266]) was suggested as a possible solution to address lecturers and students' time constraints, with particular potential for success with younger individuals. Indeed, the latter were seen as more accustomed to using the internet and new technologies to learn, as well as being less time-pressured than senior investigators. The Professor at UCAD Public Health described using a particular blended approach to cater for the needs of busy professionals, in which 'students' are given relevant documents to read in preparation for a flexible and short tutoring session, focussed on clarifying obscure points. This 'low-tech', one-to-one blended

learning approach appeared to mitigate the time constraints of both the tutor and trainee, but I could not gain an understanding of the learning outcomes.

Importantly, the key change agent in Senegal voiced a similarly pro-active approach to his own capacity development, and seemed to find the time for personal training aside of his dual statistician and programme manager role:

> "*[I have] a certain will to not sit down and get bored ... [if I don't train I'll become] bored, so I tell myself 'here, I don't know anything, I must address that. [The] work I've done in the past ... if I have to do it again, how could I do better, so that it's faster?' ... I like learning, until the head is tired!*" – SA11, Programme Manager (UCAD Parasitology)

This highly motivated participant however regretted the absence of certification for most online modules or other capacity development opportunities he seized.

Further generic barriers to online training included the difficulty in identifying existing opportunities. In fact, one Congolese investigator, who anticipated that useful resources would result from my research, recommended that I follow training in marketing to learn how to advertise these resulting resources at best. Importantly, participants in both countries preferred French material for training, whose limited availability they deplored. The same Congolese investigator as mentioned above explained how reading through English resources is a time-consuming activity requiring intense concentration, hence putting francophone researchers at a competitive disadvantage faced with their Anglophone counterparts:

> "*The problem is that in English ... I have a slightly delayed thinking compared to when I act [work] in French. It means that it requires me more energy, more time*" – PI25, Pharmacologist and Lecturer (IRSSA)

The need for better foundation courses before specialist modules

Compounding the barrier to capacity building due to lack of time for students and lecturers, the latter group described a mismatch between the student's level of knowledge at entry to courses and the amount of time allocated to teaching. Illustratively, in Senegal, the Professor in Public Health explained that some of his sessions on the use of software turned into introductory statistics courses. While "*this was not the purpose of the training*", he described such a digression as essential to enable individuals to understand the functionalities of the tool he was presenting. In the Congo, the microbiologist/lecturer in biostatistics was outraged by the amount of material on software, data management and statistics that the faculty expected him to cover with university students in a limited time, suggesting that basics should be better covered before introducing courses in advanced statistics and software:

> "*You can't do [cover] Epi Info and data analysis in 20 hours. It's unthinkable ... The informatics course has 40 hours, and in those 40 hours, you have to teach Windows ... Microsoft applications ... certain [software] like PDF ... you can't dare to add [statistics] ... So I said I couldn't carry on, they can find someone else!*" – SA22, Microbiologist and Lecturer in Biostatistics (retired)

Similarly, some Congolese investigators suggested that short modules and one-off initiatives are insufficient to fix the current capacity gap. In the words of a biologist at the LNSP, proper foundational, "*global*" and "*integrate[d]*" training, from "*how to create data*" up to "*how we uncover information that may be relevant to make a decision*" is required to develop the local capacity in health information and data management, as well as in statistics. Some Congolese academics (and in particular,

the biostatisticians) felt that a shift is necessary among medical doctors, from a culture of 'clinic and cure' to one of 'public health and prevention'. According to the microbiologist/biostatistician, clinical investigators need to "*become immersed*" in statistics rather than call statisticians for help at the time they need to publish with higher impact, to eventually obtain a professorship. In a similar line, the FCRM statistician voiced doubt over the usefulness of the newly created Masters degree in Public Health at the University Marien Ngouabi. He considered that this course was unlikely to address the gap in statistics and data management capacity in the Congo, describing the curriculum as suffering from a pitfall perceived as typical in the country of being more "*clinical*" than "*quantitative*".

As a final consideration and limitation of 'quick fixes', the FCRM project manager mentioned the lack of opportunities to conduct clinical research in the Congo as a barrier to local capacity strengthening, because people that the Foundation or others train may have to leave the country if they wish to make use of their newly-developed research skills.

Chapter 11

Discussion of qualitative findings

In Chapter 10, I presented the findings from my fieldwork and in-depth interviews, seeking to stay close to the perspectives expressed by participants. In Chapter 11, I now interpret and critically analyse participants' views, drawing on insights from the literature on clinical research data management and health systems. In particular, I illustrate how the Aragón framework [125, 267] helps to shed light on key elements and processes underlying the functioning of DMS at the two case study sites, and likely by extension, in other settings as well (with limitations to the transferability of findings discussed in section 11.3). I then discuss the lack of clarity around clinical research data management as a professional role, which some authors have already highlighted, and which my qualitative case studies evidenced as well. Finally, I close this chapter with considerations on the trustworthiness of such findings [268], derived from a few specific cases and qualitative methods.

11.1 Shedding light on data management systems using the Aragón framework

In this section, I clarify the origins and my own understanding of the Aragón framework, before reconsidering findings from my qualitative data in its light. In the original paper by Aragón, taking a 'hard' perspective on understanding (and influencing) the functioning of a system meant examining its constitutive parts with an engineer's mindset [125]. In contrast, Aragón described a 'soft' perspective as acknowledging "complexity and confusion" in organisational processes, which are as such not amenable to fixing and modifying in an engineering sense, but nonetheless critical to act upon when seeking to strengthen a system [125].

The **Collaboration for Health Policy and Systems Analysis in Africa** (**CHEPSAA**) adapted the Aragón framework for the analysis of health systems (for illustrations of the CHEPSAA interpretation of Aragón's work, see the case studies accompanying teaching resources on its website [269]). In this adaptation, CHEPSAA suggests that four hard components make up the 'engineerable' part of the system: Organisational hierarchy, HR establishment, Technology, and Finance. I understand 'organisational hierarchy' as the 'immaterial structure' or backbone of an organisation. In other words, this concept encompasses the roles, rules and constraints (whether these are official and written policies, or tacitly agreed), which remain regardless of the people belonging to the organisation at a given time. The Aragón framework shifted my thinking from the issue of access to resources (whether physical infrastructure, technological tools, funding or skilled staff) to the question

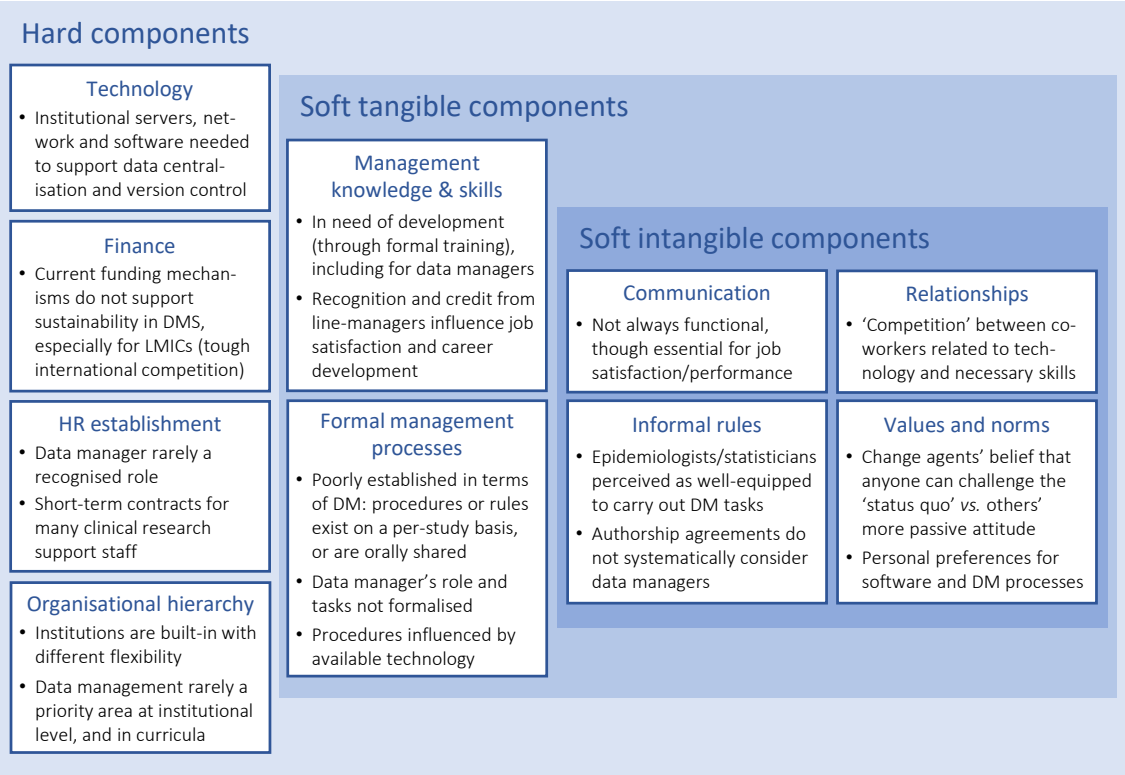


Figure 11.1: **The Aragón framework applied to clinical research DMS in Senegal and the Congo**. Terms in blue are components of the Aragón framework, as interpreted by CHEPSAA [269].

of how organisations are set up to use resources. Indeed, the organisational setup significantly affects local DMS in Senegal and the Congo, as I explained throughout Chapter 10 (especially, in sections 10.2 and 10.3) and discuss further below.

In CHEPSAA's representation of the Aragón framework, soft components are further divided into tangible and intangible ones. Though more difficult to engineer than the hard components of a system, the soft tangible components include managerial aspects (procedures, knowledge and skills), which can still be manipulated to some extent. The soft intangible components are at the core of an organisation and are much harder to influence, as they strongly depend on the informal ways in

which individuals interact with each other. Throughout my fieldwork, I uncovered impacts of soft components on the functioning of DMS (as outlined in sections 10.4 and 10.5). Those impactful soft components and their interconnections with hard components became further evident through the contrasting of the two cases, and through their examination in light of the Aragón framework.

Figure 11.1 presents the Aragón framework as it applies to my qualitative data, drawing on the graphical representation of Aragón's work first proposed by Elloker *et. al.* [270] and then taken further by CHEPSAA [269]. Of note, I substituted the terms 'hardware' and 'software' normally attributed to the Aragón framework by 'hard components' and 'soft components', to avoid confusion with the ICT and computer science meanings of these original terms, which are of great relevance to a thesis on data management.

11.1.1 Technology: The need for servers and role of change agents in upgrading data management practices

As expected in the case of a technology-intensive topic like data management, technological tools emerged as a major component influencing data quality and integrity. Most strikingly, participants in both countries raised the need for robust institutional servers to better support clinical research data management activities. The case studies highlighted that servers are particularly important for synchronising data across computers (hence across team members), thus influencing version control and long-term preservation of data (beyond the primary researcher's computer).

My case study data suggests that technology partly determines data management processes. For example, data validation procedures differ when using spreadsheet or database management software. In addition, available servers influence (or impose) the choice of data collection and management tools: the limited reliability and capacity of existing servers slows down the adoption of EDC in Senegal, and their near absence hampers the use of client-server database management software in the Congo (among other factors). The choice of technological tools is further driven by soft tangible and intangible components, including the available knowledge and skills, as well as individuals' values and norms underlying their personal preferences. These multiple elements at stake in the selection of data management tools and processes show that interconnections between systems' components are complex and extend beyond the apparent hard/soft divide of the Aragón framework (Figure 11.1).

While a participant in Senegal argued that making servers available goes beyond the responsibility of an institution (but rests on the national level), others demonstrated a proactive behaviour in building technological capacity from the bottom up. The values, strong commitment and beliefs of 'change agents' [263, 264] that they can improve a situation enable them to progressively upgrade existing tools and implement new systems, eventually demonstrating their benefits to others and encouraging further investment and improvement. In practice, the responsibility to provide robust servers is certainly shared between institutions and nations. The establishment of a robust ICT network providing internet access and enabling the use of online systems indeed requires national action, while the establishment of local networks does not.

Re-examining the pros and cons of EDC in light of existing literature

Participants' perspectives on the opportunities and challenges of using new technology (and especially, EDC) in LMICs largely aligned with those reported in the literature. Major opportunities mentioned by Senegalese participants and documented by others who used EDC in similar research contexts include gains in data collection and processing time [93, 94, 262, 271], as well as enhanced data quality [262, 271]. As a reminder, these were also the major advantages reported in articles on the use of EDC for the collection of routine data [85, 87–90].

In parallel, key technological challenges experienced in Senegal (and elsewhere [93, 94]) relate to the dependency on external providers for data transmission or storage, and to limited reliability of the national ICT infrastructure and associated connectivity for data transmission from rural areas. Some authors have also mentioned challenges related to the poor-quality of mobile devices [93, 262]. Finally, an important consideration when using EDC is to ensure encryption of the data from collection up to reception into the database, so that participant confidentiality is maintained [94]. Senegalese data managers were aware of that necessity to encrypt data, but they did not perceive it as a challenge.

On top of the technical challenges summarised above, key barriers to the adoption and implementation of EDC in LMICs (including Senegal) concern the need for training in programming and in the use of tablets or smartphones (especially for fieldworkers with lower education levels) [89, 92–94]. Despite the significant learning curve for fieldworkers, several authors also report "enthusiasm" and high usability

of EDC tools among this staff group [94, 271], suggesting the latter barrier can be overcome thanks to people's values and norms, to which 'enthusiasm' would pertain. Similarly, Knipe *et. al.* found a positive impact of EDC on the overall perception of their research work by community fieldworkers, as the use of hand-held devices by data collection staff elevated their status [272]. Of note, the need for advanced programming skills is in fact not restricted to DMS using EDC: as an illustration, Park *et al.* explain that developing their study-specific data system (in which source data was collected on paper) necessitated "three-month time of a skilled system designer and programmer" [111]. Similarly, Watson *et al.* built a DMS relying on remote data entry and electronic transmission to a data management centre; and they found that this necessitated considerable planning and piloting efforts [44].

As a concluding remark on EDC and technologies, Senegalese participants regretted the high cost associated with the use of EDC, which deterred some from using this tool unless it is imposed and paid for by the study sponsor. Some authors reported that re-using mobile phones or tablets from one study to the next has potential to make this approach to data collection cheaper in the long run [262], which supports DM11's view that the real barrier to EDC adoption is its perceived upfront cost rather than its running cost.

11.1.2 Finance: The impact of participants' values and norms on funding accessibility and use

More generally, financial resources are another hard component whose influence I expected to observe in my case studies of DMS in LMICs. Past qualitative work has documented the many barriers to clinical research in low-resource settings [30, 31], which include the limited availability of financial resources locally, as well as challenges with attracting funding through competitive mechanisms for academic research funding. My qualitative findings mirror those of Franzen *et al.* in that respect, considering that participants in both countries and across institutions I visited expressed very similar difficulties with securing funding for their clinical research and capacity strengthening activities. Participants in the Congo also emphasised the lack of national commitment towards funding these areas, a problem also raised in a recent report by The World Bank [64]. I further found an influence of participants' values and norms on funding availability, which was evidenced by researchers' willingness to invest out-of-pocket to establish their own research laboratory (mentioned in the Congo), or by investigators' willingness to devote variable amounts of the typical study budget to data management (mentioned in Senegal).

Some of my study participants (especially in the Congo) also emphasised access to funding as a barrier to clinical research in general, and further suggested that the difficulty in attracting funding relates to a lack of a "*culture of searching for funding*", itself linked to the absence of a research culture altogether within the country. Establishing such a culture again requires, among others, to address indi-

viduals' resistance towards something new, that is to say, it requires acting at the level of their values and norms. Congolese participants also regretted the absence of a data-driven culture of policymaking, another soft intangible component that constitutes a barrier to raising awareness on the necessity for clinical research and capacity development in that important area, which implementation science often seeks to address [95, 96].

11.1.3 The unexpected influence of organisational hierarchy on human resources establishment

While I was not surprised to uncover limitations related to available technology and funding, and to some extent, to a skilled workforce, I did not foresee the influence on HR of the organisational setup and associated structural constraints. In my initial reading of the literature on data management practices and systems, I indeed did not identify this potentially impactful component. Yet, the lack of recognition of the data manager role, hence impossibility to recruit this staff group on permanent contracts at UCAD, appeared as a key barrier to the establishment of a long-term data management workforce within the institution, and imposed to draw on per-project funding and contracts. Difficulties in retaining skilled clinical research staff also related to their need to earn a decent living, emphasising the absence of career development pathways and opportunities for continuous employment within clinical research systems in both countries.

In addition, I did not anticipate the varied backgrounds and competencies of the available data management workforce in the two countries (a key finding further discussed in section 11.2). The available skilled workforce within a country is partly influenced by the priority subjects taught in scientific curricula at universities, which may limit the visibility of clinical research as an avenue for data management skills (as opposed, for example, to business and economics).

11.1.4 Relationships, communication and informal rules complement formal management skills and processes

Data management processes in the two countries seemed on their way to becoming formalised, with a stated intention to establish written standard procedures or policies at both sites. Until these processes are established at the institutional level and/or further harmonised across teams, informal rules and personal habits however prevailed for data quality control and for the distribution of data-related tasks (especially, in the Congo). For example, there often was a tacit expectation that statisticians should handle data management, or that epidemiologists are well equipped to do the job. This perception seems to neglect the computing and software programming aspects of data management, which are well covered in a recent competency framework for clinical research data managers [101], and highlighted as critical in different methodological articles describing study data management procedures [44, 92, 111]. This also reflects the lack of a 'formal' definition for the clinical research data manager's role (the focus of section 11.2).

In addition, the two qualitative case studies highlighted the line-management duties of data managers, who notably have to manage data entry staff and field-workers. In the absence of formal training to take on this people's management role, participants relied on their own interpersonal skills; a 'soft' skill reflecting participants' values and relationships. In parallel, investigators acted on their friendly relationships with officials to help resolve difficulties with communication and formalised processes for the rollout of SMC evaluation studies. Good communication and professional relationships between operational staff (including data managers and data entry staff) and their line-managers also influenced workers' satisfaction, perceived job recognition and credit, with consequences on the DMS through workers' involvement and motivation for the job. Although soft components are rarely at the forefront of discussions related to the implementation of clinical research DMS, Style *et al.* mention a similar influence of data collection staff's "motivation and organisational skills" on their ability to timely collect data and follow up participants [93]. These authors also see a role for new technologies in supporting these planning tasks, further illustrating the interplay between soft and hard components in DMS functioning.

11.2 Raising the profile of clinical research data management

The role of biostatisticians and other data staff is evolving in the current context of open science [3, 273], and the role of the clinical research data manager has only recently gained attention as a professional role. Table 11.1 below lists key tasks of the clinical research data manager, as synthesised from the TDR Global Competency Framework for Clinical Research [274, 275] (whose development I significantly contributed to, in conjunction with colleagues from The Global Health Network).

More recently, Zozus *et al.* [101] published an excellent list of competencies for the clinical research data manager. The activities listed in their framework [101] exclude data analysis and planning of the trial methodology, which are traditionally the duty of statisticians. However, in an attempt to propose bridges between roles and promote career development for clinical research support staff [276], the TDR Global Competency Framework proposes that specialist data managers may have knowledge of statistics. A clear distinction between activities that are the remit of the clinical research data manager or of the biostatistician (considered as professional roles) is indeed 'theoretical'. In fact, I demonstrated how participants' job titles do not always reflect their expected, theoretical role (section 10.2); and how Senegalese data managers perceive statistics as the next competency to master (section 10.5.2).

CLINICAL RESEARCH DATA MANAGER ROLE
Creating data
• Plan the data entry and management processes (format, storage, etc.)
• Assist IT teams in the creation of the infrastructure for data management
• Lead the process of database and CRF specification design and creation
• Develop test scripts and execution logs for User Acceptance Testing (UAT)
• Provide input in any study document relating to data collection (including CRF), management (and analysis)
• Write SOPs for database construction and requirements (security, backup, storage…)
• Write or contribute to SOPs for communication with other parties (for example, laboratories and sponsor), in particular for aspects relating to data transfer and export/import
• Ensure that the data and data systems meet the sponsor, funder and regulatory requirements
• Develop data agreements between contractors (for example, external laboratories, data vendors) and the research team
• Develop training on the Central Data Management System that is suitable for users including PI, data entry staff, study coordinators and monitors
• Train team members and staff on data entry, editing, cleaning and validation procedures
• Recruit an appropriate team of data collectors to gather study data
Processing data
• Manage and supervise a team of data entry staff
• Propose suitable means to improve the data quality management systems within the organisation
• Audit databases to validate specifications, programming and quality checks
• Monitor and evaluate the activities of data staff in own team and/or at different sites
• Verify data from source documentation
• Create reports of errors and missing data
• Design and use algorithms to query the database and resolve discrepancies
• Resolve queries, liaising with appropriate team members who generated the incomplete data if necessary
• Review data for accuracy/consistency; validate query resolution, discrepancy checks and data cleaning
• Prepare summary of data activities from databases, compile reports on data queries and their resolution, in particular for audits
• Ensure consistent execution of the DMP throughout the study lifecycle
• Perform database-lock as required by PI
Preserving data
• Ensure that source documents are kept, that a clear log is maintained and that data quality is traceable
• Archive CRFs and study database as laid-down in SOPs
• Perform close-out audits; ensure future availability of data

Table 11.1: **A 'theoretical' set of tasks and definition for the clinical research data manager role**. The proposed tasks summarise those listed in the draft TDR Competency Framework for Data Management Staff [275], and align with recent literature on research data management competency frameworks [101, 103]. Tasks were mapped to the research data lifecycle's steps, using the terminology from the UK Data Service [52] used throughout this thesis.

The 'mismatch' between theoretical roles and the job titles of those executing data management tasks emphasises that in some settings, these tasks may have to be distributed across different skilled individuals hired in diverse positions. As my qualitative case studies evidenced, reasons for not having a (permanent) 'data manager' may range from the non-existence of such a professional occupation in institutional organograms, to the unavailability of a single worker with the desired (theoretical) mix of competencies. Arguably, spreading data management amongst clinical research staff is not ideal, considering that it diverts attention from this key area of activities in clinical research. However, in practice, this task distribution approach may be the only currently feasible strategy to access the necessary data management skills. This idea of moving beyond the concept of 'rigid' roles and associated competencies, and of accounting for the capacity of the clinical research team as a whole, was also a strong motivation for the development of the TDR Global Competency Framework [276].

As a final consideration on competency development, it is interesting to note that a strong demand for statistical training emerged across participants, institutions and countries. In many cases (with the exception of individuals with clear expertise in data management), it was unclear whether researchers were fully aware of the separate requirements of data management (centred on the programming and data 'processing' aspects described in Table 11.1), as compared to those of statistics (including methodological design and analyses). Consequently, it is hard to estimate in absolute terms whether the evident request for statistics capacity strengthening voiced by participants reflects a greater need than for data management.

In fact, and echoing the challenges around delegating and sharing credit within clinical research teams, investigators may be more acutely aware and willing to address their limitations in statistics. Competency in that area is rewarded in the academic world of quantitative medical sciences, as a sound statistical approach is necessary for publication in high-impact journals. To raise the profile of data management, the academic system may therefore need to include similar mechanisms to 'reward' good data handling, for which incentives are not as clear [3]. This was also the message of data managers, who suggested funders should enforce stronger requirements for transparent, robust DMS; and this echoes the complaint that their significant contribution to clinical studies was not academically recognised (for example, through authorship of publications).

11.3 Trustworthiness of the qualitative work

In this section, I discuss different aspects that challenge the trustworthiness of findings derived from a qualitative two-case study [122]. First, I critically examine the possible transferability of results to other settings (section 11.3.1). I then turn to the issue of voices heard or missed, and to influences on the accounts given by participants, especially in light of my positionality (section 11.3.2).

11.3.1 Transferability of findings to other settings

The qualitative case studies were critical in enabling me to explore the perceptions, practices and unanticipated factors that underpin the strengths and weaknesses of DMS; something that quantitative methods did not allow me to do. However, this in-depth, focussed approach inevitably comes with the typical limitations of qualitative research [253]. I investigated two contrasting cases, and sought to gain a thorough understanding of contexts, which suggests that some key mechanisms may be transferable to other settings sharing 'determining' characteristics [254].

As emphasised above, the Senegalese case study stood out for the on-going transition from paper to EDC (and other upgrades in pre-existing technological systems). This situation exacerbated individuals' actions and reactions to change, and made the role of 'soft' components in a DMS functioning all the more prominent. In similar contexts of constrained resources for data management but a push for technological change, it is likely that the importance of workers' relationships, values and norms will drive DMS functioning and improvement. The challenges faced by clinical research teams at UCAD Parasitology are thus of relevance to many institutions planning a switch to EDC. For example, the perceptions of a threat to some jobs is likely transferable, as well as the challenges in acquiring new, necessary skills.

In contrast, institutions in the Congo were at earlier stages of DMS setup. This provides an opportunity for building robust DMS from scratch, but certainly makes access to 'basic' (yet adequate) technology and skilled HR one of the most pressing issue. Use of 'sub-standard' data management software like MS Excel® is likely

to be observed in similar settings where hard components are not more conducive to good practice. Furthermore, a particular strength of the FCRM concerning data management capacity strengthening is its independent status, as well as its very mission to induce change through the 'creation' of a clinical research culture within the country. Institutions with a similarly positive and pro-active attitude to change are likely to provide a fertile ground for advancement of DMS (within the limits of what other systemic, hard component-related constraints enable).

In addition, it is noticeable that despite the contrast between cases, the need for robust servers at the institutional and national levels emerged in both countries; an issue related to ICT infrastructure, which is likely common to multiple LMICs [85, 86]. Participants in Senegal and the Congo had different baseline servers, networks and internet connectivities, yet they all considered that improvements were necessary to enable adequate data management. Bezuidenhout *et al.* conducted a qualitative study [121] followed by a quantitative survey [37], to explore the impact of research environments on the data sharing practices of LMIC-based researchers. Through this work, the authors emphasised that research on the challenges and opportunities of existing ICT infrastructure in LMICs needs to move beyond the question of its presence or absence in a given contest, and to focus rather on what it enables local scientists to do in terms of data management and sharing.

Further qualitative case studies would reinforce the transferability of findings, and provide insights into other, different contexts (including, institutions at different stages of establishing DMS, and in different geographic settings in and out of francophone Africa), but carrying these out was beyond the scope of a mixed-

methods doctoral project. As a final note on transferability, many of my qualitative results on the specific issue of data management capacity echo the work by Franzen *et al.* [30,31] and The World Bank [64] on clinical research capacity as a whole. Key common findings thus include the necessity for sustainable financing mechanisms, the importance of creating a national culture of clinical research, as well as the key role of change agents in promoting improvements.

11.3.2 Sampling effects and participants' openness

While I performed background (desk-based) research on the institutions I visited, the fieldwork and immersion at sites were critical in understanding local clinical research organisations. Field facilitators were immensely helpful in that respect, and they ensured that I met all the key personnel (especially but not only, those with a data management role) from day one of my visits. Their own perception of who is involved in data management in and around their institution means that I have possibly missed some data management workers; and indeed the field facilitators remembered about more individuals I should meet as my fieldwork progressed, and as participants suggested others.

This field facilitator 'bias' certainly impacts the picture I gained of data management practices outside of UCAD Parasitology and FCRM themselves. However, from my repeated (and often daily) interactions with researchers and data managers within the two case study organisations themselves, I am confident that I did

not miss participants primarily employed in a clinical research data management role within these. Furthermore, in light of the small number of individuals actually performing data management, I could interview all of them, rather than having to (purposively) sample, thus increasing the likelihood that I captured all relevant perspectives within this group. Perspectives from investigators may be more contingent on the sampling, as these participants tended to be less accessible than data managers, which made the scheduling of interviews difficult. However, I had other forms of interactions with senior investigators and head of institutions (for example, exchange and review of study reports for Senegal, and presentation of early findings at seminars while on-site for the Congo), thus ensuring I also heard their perspective.

As an important note, I delivered research seminars to present my work at both institutions. The seminar at UCAD Parasitology also included a lecture in data management for postgraduate students, and was in fact attended by postdoctoral fellows and data managers as well. Coupled to the 'University of Oxford' affiliation, this inevitably placed me in a position of 'expert', with likely implications for the information participants were willing to share during interviews centred on the strengths and challenges of their data management practices. Several interviewed investigators indeed insisted on their good publication record, as if they perceived my questions (or project overall) as a criticism, or felt the need to justify their own status. This positionality challenge adds on to classic issues of participants' recall and of unintendedly directive probing by the interviewer (myself) [253].

Another potential influence on the information shared was that some interviews took place in shared offices, in the presence of colleagues. Apart from a few excep-

tions, interviewees seemed largely unconcerned by this relative lack of privacy, as suggested by the fact they smiled at each other, openly laughed when commenting on given issues, or even asked for each other's viewpoint in one occurrence (where both participants took part in independent in-depth interviews, hence no additional issue related to consent). As illustrated by this quote from the FCRM's project manager, participants may have withhold some sensitive information, but in a way that would not compromise the overall 'truthfulness' of their account: "*The problem is that it will be written ... I cannot lie and I must be frank, but without either pouring oil [on the fire]*".

Finally, while findings and quotes are reported in an 'anonymised' fashion, I cannot guarantee the full confidentiality of participants with regard to their line-managers where reports are returned to the unit and otherwise put into the public domain (published). This is due to the low number of participants involved, and to the specificities of their projects and/or contrasting preferences in terms of data management. Sharing preliminary reports orally or in writing with study participants again suggests they do not see this limited anonymity as an issue.

11.4 Conclusions on the qualitative piece of work

The qualitative piece has been highly informative in shifting the perspective from the challenges and opportunities in data management 'strategies' (theoretical data management plan) to 'practices' (what happens on the ground, and why).

In particular, this work highlighted that in terms of technology, a key element shaping DMS is the availability of robust servers, which are notably a pre-requisite to the use of client-server databases, and enable remote data collection as well as synchronisation of data files across a network of computers. This observation sheds an interesting light on the question of which software programme may be used for which clinical study. In the absence of such ICT infrastructure, clinical research data management practices are unlikely to improve drastically, even if open-source software and freely available GCDMPs© become widely accessible.

The case study sites revealed varying levels of awareness and competency in data management among investigators, some of whom were also more reluctant than others to delegate this important set of tasks to skilled data managers. In parallel, individuals with desirable data management skills came from varied backgrounds, emphasising a lack of established educational and professional pathway for the corresponding role. This confusion around the tasks and responsibilities of the clinical research data manager occasionally resulted in a lack of recognition for the work performed by the concerned individuals within their clinical research teams and organisations, further constraining their career progression and sustainable employment.

Importantly, the qualitative work demonstrated that strengthening the capacity for data management will require much more than an upgrade in technological tools and accessible skills. Indeed, I found a critical role for change agents in improving both clinical research and data management practices within their institution, through their proactive attitude. Good communication and relationships within research teams also contribute to functional DMS, but they can be challenged by the introduction of new technologies.

Overall and thanks to the perspective brought by the Aragón framework, this qualitative piece played an essential part in uncovering the influence of human factors and 'soft components' on DMS, thereby significantly complementing the partial understanding of capacity strengthening needs provided by the previous quantitative pieces. In the next and final chapter, I take further the findings from my three empirical pieces of research, by re-examining them in light of each other, and through the lens of the Aragón framework.

Part IV

General discussion and conclusion

Chapter 12

Strengthening data management systems: global lessons from the mixed-methods study

The present thesis opened with proposed 'working definitions' for clinical research data management and related concepts (Chapter 1), which provided a framework to examine 'complex' data management systems (DMS) and to formulate hypotheses regarding their strengths and weaknesses, and regarding the external factors influencing their implementation (Chapter 2). As my doctoral work unfolded, I explored different cases within the boundaries set by my research question – as a reminder: **What are the opportunities and challenges in clinical research data management for investigators in Low and Middle Income Countries, and how could capacity-strengthening initiatives address them?**

Lessons learnt along the way led me to progressively focus on two infectious diseases of poverty affecting tropical countries (schistosomiasis and malaria); as well as on a limited geographical area, from Africa as whole down to francophone African countries (specifically, Senegal and the Republic of the Congo).

In this final chapter, I jointly interpret findings from the quantitative and qualitative parts that I have independently presented above, in light of my research aims. I discuss what the three empirical pieces taught me about the variability in clinical research data management practices in LMICs (section 12.1); and about the challenges and opportunities in establishing different DMS (section 12.2). Drawing on the specific or transferable lessons from my in-depth exploration of the aforementioned cases, I then propose recommendations for contextualised data management capacity strengthening (section 12.3). The final conclusion summarises key messages from this thesis, and restates its original contribution to the field (section 12.4).

12.1 Characterising the variability of clinical research data management practices in LMICs

The first aim of my thesis was to characterise the variability in data management practices across clinical research on specific diseases relevant to LMICs. In the two quantitative studies, I contrasted the case of schistosomiasis and malaria research, with a focus on data management strategies and outcomes in clinical trials. The qualitative piece extended the exploration of data management practices in malaria

research, and participants shared perspectives on further study types (especially, case-control and cross-sectional designs).

In this section, I describe the sources of variability in data management practices that my mixed-methods work highlighted, as well as their implications for data use and re-use, where applicable. In reference to the complex systems viewpoint [128] first introduced in Chapter 2 (Figure 2.2), I successively examine the variability in processes (section 12.1.1), tools (section 12.1.2) and personnel (section 12.1.3) that typically make up a DMS. This 'artificial' separation reflects a need to structure this discussion; however, systems' components are tightly interconnected, and it is irrelevant to consider them in isolation when seeking to strengthen a health or research system [31, 65, 277, 278].

12.1.1 Data management processes vary among studies and institutes, and are documented with variable depth

Data management processes encompass the procedures and plans put in place to handle data throughout and beyond the study lifecycle, and all three research pieces showed these processes vary across studies and institutes.

The schistosomiasis piece first demonstrated that the variability in study methods for a given disease can induce great heterogeneity in data aggregation and reporting choices. My work furthered the understanding of issues previously noted by other authors [137, 167]. The observed variability is even more interesting in

that WHO published SOPs to harmonise processes in anthelmintic efficacy studies commissioned by ministries of health and national control programmes [141], with apparently mitigated impact on practices in the global schistosomiasis research community. This demonstrates a limitation of international recommendations, which can change over time and must be properly disseminated and communicated, or otherwise risk remaining little implemented and open to interpretation.

While most evident in late data analysis steps, this variability in procedures underlying antischistosomal efficacy studies is likely to translate into variable early data management processes too. As explained in the schistosomiasis background section (Chapter 3), diagnosing infections with *Schistosoma* worms necessitates the collection of often multiple urine or stool samples, each resulting in a certain number of microscopy slides [134]. Some research teams may capture and preserve data down to the egg count per slide and sample, but others may only enter an average for the concerned individual in their study database. This variable granularity in collected data has important implications for sharing and re-use: for optimised comparability, diverse antischistosomal efficacy study datasets should be pooled at the same and most precise level, which here goes further than the individual-level suggested in the term 'individual participant data' (IPD).

When progressing to the meta-analysis of praziquantel efficacy among schoolchildren in Africa (Chapter 5, section 5.3), I found that data manipulations performed to derive population outcomes are often ambiguously reported in publications. If this ambiguity reflects the documentation of original datasets (as findings from the sub-

sequent quantitative piece would support), there is an important gap in researchers' capacity to create transparent audit-trails of their data management activities.

The quantitative analysis on malaria data (WWARN piece, Chapter 7) best evidenced the global variability in data management strategies. Despite the focus on a single disease and study type (antimalarial clinical trials), variations in shapes and contents of DMPs, meta-data and databases were remarkable. I initially considered exploring correlations between end-data validity and initial data management choices (as explained in Appendix B). As I could not conduct this analysis, associations between diverging data management strategies and data quality remain unclear. Most DMPs and dictionaries shared by malaria investigators were nonetheless incomplete when assessed against desirable features of those documents, with 57.3% to 66.7% of studies scoring low or medium on these indicators. Some WWARN data contributors also seemed puzzled by my request for a 'dictionary', and many did not share one, suggesting limited awareness regarding the critical role of this document or similar meta-data in the (re-)usability of research data (also noted by others in different explorations of data sharing practices [233, 234, 236]).

Finally, the qualitative case studies (Part III) evidenced variable levels of awareness regarding data management processes across institutions and individuals. Most institutions I visited were establishing data management policies or standard procedures; but in their current absence or incompleteness, the responsibility for good data management rested on data managers (if accessible), or on investigators themselves (whom past work suggests they are poorly equipped for that job [3, 4]). Pro-

cesses thus varied along with individuals' preferences for different software or strategies. The re-use of pre-existing procedures was frequent, with potential conflicts in versions held by team members. The latter observation shows that challenges with version control and audit-trailing noted in the quantitative pieces do not only concern the contents of databases, but also the accompanying study documentation.

12.1.2 Data management tools used by local investigators in LMICs are not so varied

Data management tools include database management software, hardware and the wider ICT infrastructure, as well as the medium supporting data creation (paper or electronic device) [130]. Although the three research pieces evidenced variability in tools, the latter was not as great as I anticipated. For quick reference, Table 12.1 jointly summarises the quantitative data on reported software that I have already presented independently for each disease in Chapters 5 and 7, respectively for schistosomiasis and malaria. The two samples of antischistosomal and antimalarial efficacy studies suggested that both research communities use surprisingly similar programmes for data management, and that these are mostly Epi Info^{TM} and MS Excel®. Findings from the qualitative study in the Congo reinforce this quantitative observation, considering that all investigators and statisticians I interviewed in this country mentioned a preference for one of these two programmes.

In contrast, findings from the qualitative study in Senegal suggested a more frequent use of MS Access® and SQL databases, and more recently of ODK to support

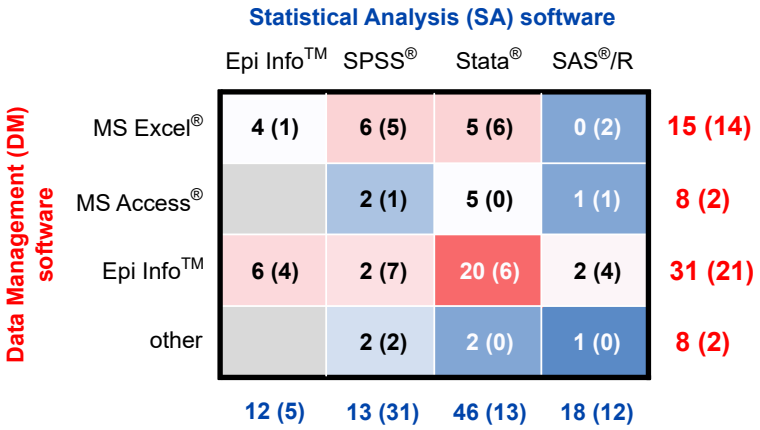


Table 12.1: **Data management (DM) and statistics software used in trials of antimalarials (and antischistosomals)**. Numbers are counts of trials reporting the use of the corresponding combination of DM and statistics software, for the malaria quantitative sample, and for the schistosomiasis one (into brackets). Grand totals include studies which specified one type of software but not the other. As a reminder, the samples are little representative of studies sponsored by the private sector.

EDC. No study in the quantitative sample of antimalarial trials mentioned the use of EDC, and UCAD appears as an outlier in its use of this technology. However, the introduction of EDC at UCAD was relatively recent, and participants mostly reported its use for case-control or cross-sectional studies; two elements of context that are poorly captured by the quantitative sample drawn from the WWARN repository. Altogether, findings from my mixed-methods work thus support others' claim that researchers in LMICs still rarely use EDC for clinical trials [93, 94]. Investigators in Switzerland and Tanzania recently developed an additional module to enhance data management functionalities of the ODK suite, which may promote its greater use for interventional trials [114].

Finally, the two qualitative case studies highlighted variability in exact institutional and national ICT infrastructure, but rather similar limitations in terms of

what DMS it enables setting up and operating. Congolese research institutes rarely had a local network and a desktop computer to use as server; and if they did, participants expressed doubt in the sufficiency of such systems to sustainably host and preserve the volume of data involved. Senegalese participants benefited from better infrastructure within their institutions and country, enabling them to take greater advantage of online systems and remote servers; however, the data managers and statisticians at UCAD Parasitology also considered that existing servers in their department were sub-standard compared to what would be needed.

12.1.3 Data management personnel and actors vary greatly

The two quantitative datasets had limited value to explore variations in data management personnel. Nonetheless, the few malaria study protocols that listed staff emphasised statistical needs and stated that an epidemiologist was involved. This assignment of data management tasks to individuals with statistics expertise strongly aligns with qualitative findings from Senegal and (even more) from the Congo, where statisticians or researchers with a public health background were frequently perceived as equipped with data management skills. More generally, variability in background and job titles concerned both roles of data manager and data entry clerk. Investigators at UCAD Parasitology seemed more aware than several investigators in the Congo (especially, outside of FCRM) regarding the requirements of clinical research data management; however, according to data managers, this raised awareness did not translate into investment of (limited) resources in that activity.

12.2 Clarifying the reasons behind variability in data management practices

Having comprehensively reviewed the sources of variability in clinical research DMS evidenced by my three complementary studies, I now critically examine the factors underlying such variability. The second aim of this thesis was indeed to clarify the reasons behind variability in data management practices (as well as their implications for data use and re-use, which I already alluded to in the section above).

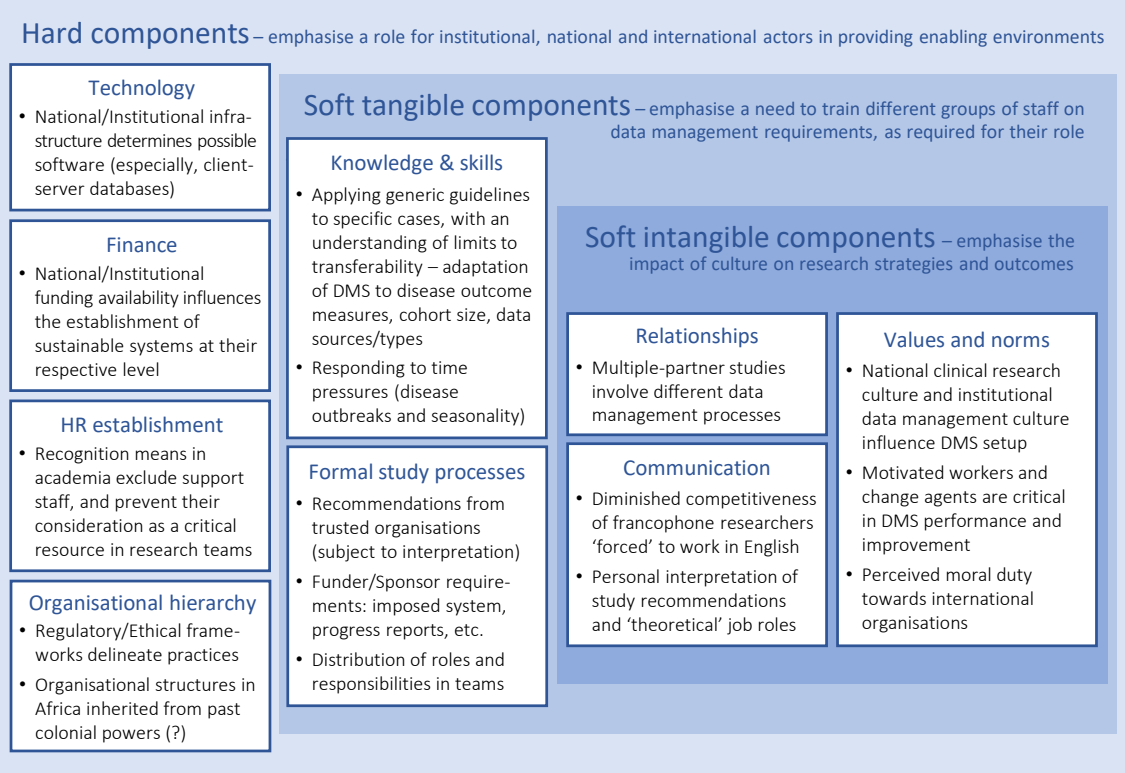


Figure 12.1: **The Aragón framework applied to clinical research DMS**. This mixed-methods study highlighted a variety of influences in the setup and functioning of a DMS, which strongly speak to the CHEPSAA interpretation of the Aragón framework adapted here [125, 269], to become inclusive of research-specific elements and of wider influences from international actors.

To discuss the key factors influencing the establishment and operation of a DMS, I draw once again on the Aragón framework (Figure 12.1). In this section, I thus successively focus on hard (section 12.2.1), soft tangible (section 12.2.2) and soft intangible (section 12.2.3) components which influence DMS; although again these systemic components are an interconnected whole, and themes cutting across several components are only mentioned once.

12.2.1 Hard components

Using the Aragón framework [125] and a 'complex systems' lens [128] draws attention on the impact of institutional, national and international resources on local DMS, and highlights the influence of global frameworks on individual practices in clinical research and data management.

The availability and sustainability of technological, financial and human resources constrain possible data management strategies

As one of the diseases mentioned in the sixth Millennium Development Goal ("Combat HIV/AIDS, malaria and other diseases" [279]), malaria attracted much attention from different global health stakeholders long before NTDs like schistosomiasis [222]. Greater investment and learning opportunities stemming from overall more research efforts could translate into improved data management practices among malaria investigators. Yet, the picture provided by the two quantitative datasets suggested

that surprisingly similar software programmes are used for clinical research data management in both research communities (Table 12.1).

In fact, this apparent similarity of tools in both communities likely reflects the limited representation of studies sponsored by the private sector within the subsample of antimalarial trials. Considering that hardly any of the schistosomiasis treatment efficacy studies considered here was sponsored by a pharmaceutical company, the software information for the two disease communities comes from similar populations of academic researchers (worldwide) and local investigators (based in LMICs). Beyond comparing studies funded by the private sector with others (funded by the public or non-profit sectors), I could not gain a precise understanding or assess the impact of resources assembled to conduct the clinical studies constituting the quantitative samples.

In contrast, the qualitative case studies enabled an in-depth exploration of resources accessible to local teams and for different studies. The availability of financial resources through sponsors came up as an influential factor in the DMS setup, especially when deciding to collect data with paper or EDC in Senegal. Participants emphasised the need for sustainable funding mechanisms, which are critical in maintaining an institutional workforce, including competent data managers and data entry clerks. In a context of limited recognition for the latter two groups of staff by academic institutions, continuous funding from other sources becomes even more critical to their sustainable employment. Finally, the qualitative work also highlighted that the ICT infrastructure is instrumental in enabling the use of different technological tools, such as client-server databases or EDC.

The situation in the Congo globally reflected the importance of a favourable 'culture' to enable health and research systems strengthening, which is certainly more determining of a country's clinical research context than its 'overall resources'. The World Bank categorises the Congo as a lower-middle income country [243], and Senegal as a supposedly 'less-resourced' (low income) one [241]. Yet, Congolese participants deplored the lack of a clinical research culture in the region, which made their 'wealthier' country overall less conducive to clinical research than Senegal, and emphasises the need to gain an understanding of how resources are made available and used in national research systems and for specific studies. Of note, Franzen *et al.* also identified what they call a "pro-research culture" as an essential determinant of a country's clinical research capability, with Sri Lanka faring better than Ethiopia or Cameroon in that respect [31].

Higher education, academic and organisational hierarchy do not facilitate careers in clinical research data management

Most participants in both countries mentioned statistics in the context of career development, and several in the Congo associated complex statistical analyses with the possibility to publish their results in higher-impact, international journals. No mechanism currently rewards good research data management in such an obvious way for academic researchers [3], which likely acts as a disincentive to invest time and resources in this essential activity, and may explain the limited awareness of some interviewees regarding the data management skills they would need beyond

the statistical analysis stage. Overall, the current academic landscape thus provides fewer incentives for clinical investigators to improve their own knowledge of data management, or for institutions to consider data staff in their typical HR.

In parallel, national research systems in Africa (and likely, elsewhere) are probably influenced by the organisational structure formerly imposed by colonial powers, whether France or other European countries. Testing this hypothesis was not a goal of my work, hence I collected little evidence to support or refute that claim. However, I noticed parallels between the higher education system in France and in the two countries I visited, which somewhat differ from the Anglo-Saxon model [280]. These systemic differences may result in different curricula and accepted academic career paths, and by extension, in different workforces for clinical research data management within francophone and anglophone countries.

Beyond the specificities of each country's higher education and research system, the case studies highlighted some barriers that French-speaking countries specifically face, most of which are likely transferable to other, non-anglophone nations. Participants in both Senegal and the Congo mentioned the lack of French material to support their clinical research efforts, as the academic world is dominated by English publications and conferences, and much of the internet content for international audiences is in that language [281]. As a reminder, one Congolese investigator described undergoing a double-shock when reading content on an unfamiliar topic and in a foreign language. He further called for greater political engagement from France and for coordinated efforts among French-speaking countries, to create a culture of francophone research, which he envisioned as competitive against

the anglophone one. Interestingly, the Organisation for Economic Cooperation and Development (OECD) proposes that the 'domination' of English in academia may become challenged by rising powers with large populations like China and India, or jointly by hispanic and lusophone countries [281]; but developing an 'independent' research community to compete with the existing one still seems hard to realise, and counter-productive in terms of global knowledge exchange.

Regulatory and ethical frameworks impose higher standards for clinical trials, but show value for other study designs

In health-related investigations, whether interested in the efficacy and safety of drugs or in the performance of services and public health programmes, the boundary between research and evaluation or surveillance is sometimes tenuous [75]. The difference between the two resides in their pre-specified aim. Research generates data to derive new and generalisable knowledge, or in other words, to understand what would be best to do; whereas evaluations monitor what is typically done, to understand if it unfolds as planned [75]. This distinction is important, for it bears implications in terms of ethics and regulatory frameworks, but in practice, both approaches may require similar methods and data management procedures. Observational studies were the most frequently discussed by Senegalese and Congolese interviewees. Some participants also had experience ranging from interventional trials down to HDSS household surveys feeding into HIS (Figure 12.2), and they holistically discussed the associated challenges and opportunities in data handling.

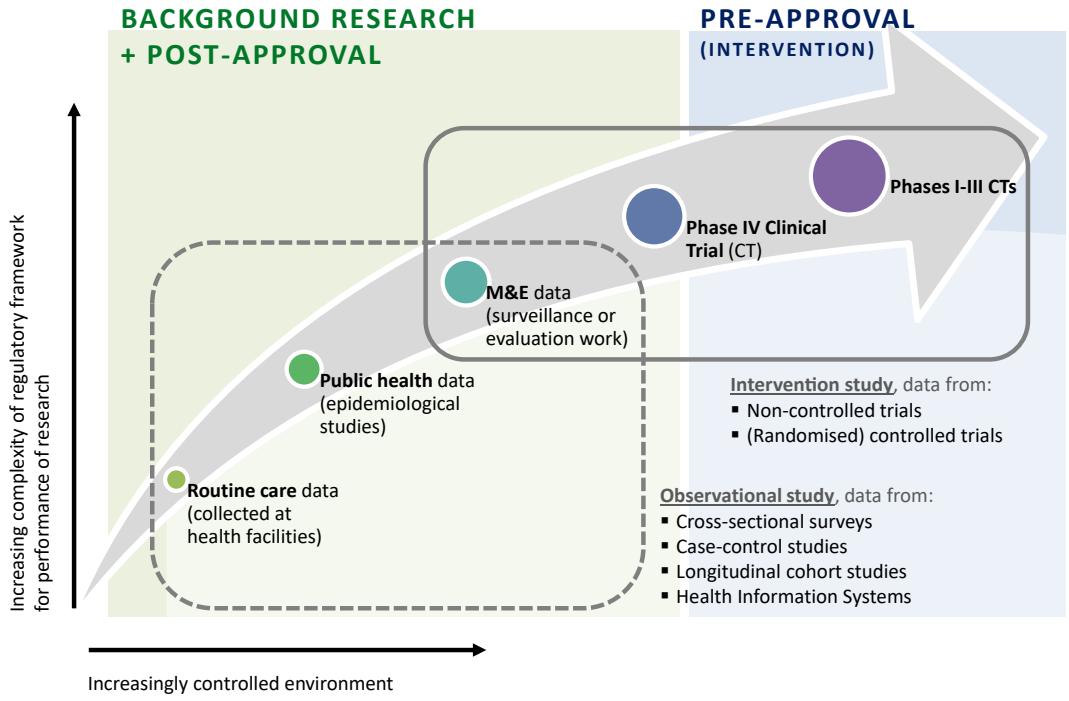


Figure 12.2: **Sources of clinical data for research and types of studies it may support**, in relation to the stringency of regulatory framework and controlled environment when planning such research. CT: Clinical Trial; M&E: Monitoring and Evaluation.

In the schistosomiasis scoping review, I considered varied study designs; not restricting myself to trials, but considering studies with a similar design, yet not considered to be trials because the investigator is not the one assigning drugs. Provided that their data is valid, such studies would interestingly contribute to a pooled database of IPD, hence their consideration in this work [166]. At the schistosomiasis data platform stakeholders' meeting [133], some however voiced concerns over the lack of quality control measures for data generated outside of 'rigid' research frameworks. This supports the view from Senegalese data managers that best practices should be applied beyond cases where they are imposed by regulations, so that all

clinical data can be meaningfully and maximally (re-)used.

Equally monitoring all data however goes against a pragmatic, 'risk-proportionate' approach to data quality control in clinical research [227]. Applying all possible validation measures may seem ideal for data reliability, but some studies suggest marginal benefits of whole-inclusive data verification [282]. Promoting a stringent approach should also balance out the risk of rendering processes too burdensome, which could deter local investigators from performing most needed and sometimes 'simple' studies [43, 48]. Without altering the overall DMS, flexible quality control procedures may justifiably apply to different data, as done for example in the PERCH study, for which "data checks were prioritised based on the potential impact on primary study objectives" [44].

12.2.2 Soft tangible components

Although GCDMP© [116] cover elements of project management, such soft components and their significance for DMS are rarely the focus of data management guidelines. The perspective I borrowed from the health system and international development fields [125, 267, 269, 270] was highly valuable in going beyond an exploration of processes and tools, traditionally at the forefront of discussions around data management or similarly 'technical' topics [114, 283].

In applying the CHEPSAA interpretation of the Aragón framework [269] to research systems, I re-framed the 'Management knowledge & skills' component as 'Knowledge & skills', to allow consideration of both research and management com-

petencies; and I rephrased 'Formal management processes' as 'Formal study processes' (Figure 12.1). Thus framed, those two tangible components appeared to exert a strong influence on the way clinical research DMS operate.

Specific elements of study design and size may call for different DMS

While 'best' data management practices should apply across study types, specific designs may call for variations in the DMS setup, and explain some of the variability mentioned in section 12.1.

In the quantitative piece on malaria research, data management practices appeared 'better' for trials that had a safety objective on top of an efficacy one: safety and efficacy studies indeed scored higher on the constructed data management indicators (Chapter 7). The possibility to test the impact of other design elements (notably, randomisation and blinding) was compounded by the overall homogeneity in the sample (Appendix B), but the number of treatment arms did not correlate with any data management score.

When asked about study-related elements affecting DMS, few participants in the two qualitative case studies raised any. The detailed analysis and coding of interviews however pointed towards a few exceptions, where participants' views suggest that different study designs lead to different data management practices.

I extensively discussed in Part III that Senegalese participants held different views about EDC, and that several of them questioned the associated possibility

to perform source data verification in the way they were accustomed to [282]. For this reason (among others), one data manager at UCAD suggested that paper-based data collection is preferable for clinical trials. His colleague did not support that view, arguing that monitoring can still occur with EDC, by checking and digitally annotating electronic forms before their transfer to the central database. While these conflicting opinions first related to the functionalities of EDC, they also suggest a belief from some participants that trials require distinct DMS.

In providing a detailed account of the data management procedures for a RCT in Nepal, Style *et al.* also largely discuss how their study procedures directed the choice of different (EDC) software, alongside costs and training considerations [93]. They concluded that EDC was valuable for some but not all data collection events occurring in their RCT, thus suggesting that DMS require flexibility to accommodate intrinsic specificities of the study design and procedures.

Finally, the analysis of data shared with WWARN also highlighted the cohort size as an impactful factor on the chosen software. This aspect did not come up as such in the qualitative data, but one Senegalese data manager discussed the issues and coping strategies he had to find to maintain a very large database (and over a long period of time), thus suggesting larger cohorts may indeed require more elaborate DMS. This was in fact a HDSS database, whose specific challenges were induced by population dynamics, and are also reported by INDEPTH [84].

Disease-specific requirements may further shape the DMS

In line with the same diagnostic methods for both disease areas, the STHs literature suffers from similar limitations as the schistosomiasis one [157], which impede the comparability of aggregated, population-level measures. The similarities in procedures and overlap of the two research communities actually prompted the development of a single data platform integrating IPD on both diseases, currently ongoing under the umbrella of IDDO. While this occurs at the end-point of the data lifecycle (sharing, preserving and re-using data), the technical feasibility of a pooled database suggests that the same initial DMS could be implemented across studies on either disease (all other things being equal).

Throughout Part III, I presented the qualitative data in a holistic manner, without commenting on potential differences between research on malaria or other diseases. In practice, interviewees primarily discussed malaria studies, in line with the great emphasis on that disease within their research institutions and with my own stated focus. The detailed coding and analysis of interview data again pointed towards a few situations where the disease influences the DMS; however, the differences noted remain largely transferable to other diseases as well.

The seasonality of infections is a frequent characteristic of these diseases that may have implications for DMS. Several participants at UCAD mentioned the challenges imposed by the malaria season, from pressure to have all elements in place promptly to run the study within the appropriate seasonal window, up to the cycles of hiring and releasing data entry clerks to perform this activity at regular but 'short'

timeframes. Many infectious diseases have peak seasons, hence challenges described by Senegalese participants for seasonal malaria could apply to multiple other communicable diseases. Readiness and time pressures may be even more intense when conducting research on epidemic diseases occurring unexpectedly [64].

In contrast, non-communicable diseases would impose different constraints on DMS. Many of the latter are also chronic, with timescales requiring tools and databases that, once set up, demonstrate reliability over several years to follow long-term outcomes [111, 113]. Of note, infections can also be chronic or predispose to non-communicable diseases (for example, HIV/AIDS and Chagas disease), hence again a rather 'irrelevant' distinction between diseases as such.

Finally, when asked about study- or disease-specific elements influencing data management practices, the data manager at FCRM rather contrasted studies conducted internally by FCRM biologists with studies involving external partners too. As a reminder, he described the latter as more burdensome, owing to the 'soft' yet real challenges of managing data from different sources and among multiple actors.

Sponsors exert a 'soft' tangible influence on DMS performance, beyond the resources they provide

In reference to national control programmes (whose data management practices are synthesised in Appendix C.2.3), a few Congolese participants suggested that some are better than others at storing and retrieving their data and reports. One statistician boiled these differences down to "*a people's issue*", suggesting that "*the*

one who is there [in charge of the 'model' programme] ... loves his work", thus emphasising the importance of 'soft components' (especially, individual motivations) in the functioning of DMS. The same statistician acknowledged a role for financial resources and funders, specifying that this 'model programme' used to be supported by French funding, and collapsed when support ceased. Interestingly, on top of the financial benefit, this participant suggested that funders' requirements for progress reports and results were critical in creating incentives for good practices. Similarly, a junior investigator in Senegal mentioned an ongoing switch from MS Excel® to a more adequate data management software for a study he was involved in, triggered by a request of the sponsor.

The qualitative case study at UCAD Parasitology further suggested a strong influence of managerial processes in overall DMS performance, especially in the context of a perceived competition and of evolving demands for skills. The introduction of a new technology for data management can be disruptive, and lead to a redefinition of job roles. In that context, data entry clerks in Senegal would have appreciated acquiring experience as fieldworkers, but these were perceived by their line-managers as separate functions. In the literature, recommendations for the management of data personnel include sending data entry clerks to the data collection point in the field, and conversely, having fieldworkers perform some data entry [130]. This strategy enables both groups of staff to become aware of the challenges of each other's work, hence encouraging more considerate data collection and entry, and eventually enhancing the quality of study data [130].

12.2.3 Soft intangible components

The 'boundaries' of many clinical research roles are loosely defined [276], especially in small-scale teams and LMICs where one individual may take on multiple tasks and responsibilities. Importantly for DMS, a lack of clarity in the definition of a professional role particularly concerns the research data manager (in and out of clinical sciences [3, 101]). In the absence of 'formal' job descriptions in institutions or teams, the remit of the data manager *versus* statistician role remains particularly open to interpretation for different investigators and research stakeholders. What people perceive as data management thus strongly depends on their own personal understanding and what they have been exposed to in that respect [51].

Here, I discuss these soft intangible issues around communicating and interpreting the data manager's role. I then turn to the influence of participants' norms and values on their behaviour and work as a clinical researcher or data manager, some of which provide a strong leverage to improve data management practices.

Variations in the understanding of statistics (analysing data) and data management (processing data)

As specified in introduction (Chapter 1), the UK Data Service, whose framework I drew on throughout this mixed-methods study, has since updated its breakdown and representation of the research data lifecycle [53]. Importantly, the new version includes a "planning research" step before the "creating/collecting data" one (Figure 12.3), thus highlighting the significance of this formerly overlooked step, which

data managers at my case study sites emphasised. To balance this change, the UK Data Service merged the processing and analysing data steps into one. Despite the remaining 'uncertainty' around the exact tasks of the clinical research data manager, the statistical analysis competency is usually not at the forefront of their desirable skill set [101]; and data processing tasks are clearly at the core of the data management definition proposed by CDISC [1]. I therefore see this merging of data analysis and processing into one step as an unfortunate limitation of the updated framework: it risks increasing the confusion between data management and statistics, which my quantitative and qualitative results both highlighted as a current issue.

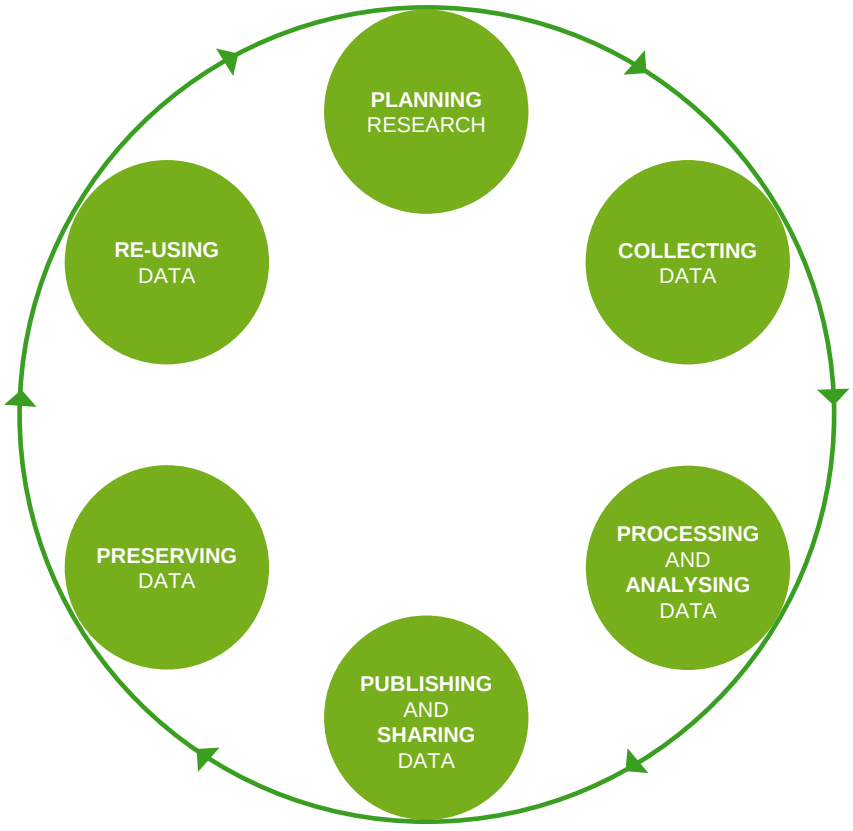


Figure 12.3: **Research data lifecycle, reproduced from the UK Data Service [53].** Current as of June 2018 (see Chapter 1, Figure 1.3 for the previous version used throughout this thesis).

In the quantitative work, some software programmes that investigators reported using for data management suggest a potential conflation with statistics. Ten antimalarial trials used software designed for data entry and database management, such as MS Access® (8 trials), REDCap or OpenClinica (1 each). A few more trials reported the use of Stata® or SPSS® for data management (Table 12.1). These programmes offer useful functionalities for data annotation, which some authors indeed report using as part of their trial data management strategy [93]; however, Stata® and SPSS® are primarily made for statistical analyses. Some studies mentioning their use for data management may thus refer to late data preparation ahead of analyses, rather than to early validation. Importantly, a 'database lock' (or freeze) should occur between quality control and analyses [116, 130], to guarantee the reproducibility and validity of findings.

The possible confusion between statistics, epidemiology and data management was most apparent in the qualitative data. Many participants (especially those with a medical background) associated public health training with data management expertise, although such curricula often dwell on study methodology and statistics [284, 285]. Among interviewees in Senegal and the Congo, demands for training in statistics were also more frequent than demands for training in data management *per se*, even though my interview questions directed towards the latter.

As a final note, the neglect of data management and its frequent conflation with statistics among researchers also reflect a greater emphasis on 'data skills' as opposed to 'software skills' when thinking about the role of the research data manager; an issue raised in the very informative Wellcome Trust report on this

subject [3]. Interestingly, this report also notes that "at present, the data scientists see their ideal career paths as being outside the research domain", with associated challenges in HICs for the sustainable hiring and career development of such research support staff, which may be all the more acute in LMIC settings.

Values and norms can drive top-down and bottom-up change

In the opening literature review (Chapter 1, section 1.4), I outlined challenges in navigating the literature on clinical research data management. Key messages from this early review included the lack of past empirical research on the barriers and enablers of data management capacity development in LMICs [97]; in stark contrast with multiplying policies and guidelines for research data sharing [18], some of which partly cover data management and/or capacity strengthening aspects [3, 118–120]. Potentially conflicting guidance and tools raise the question of who should prescribe which are 'good', and how to disseminate best practices in the clinical research community. The experience of discussing GCDMP© with trial experts throughout my doctoral studies suggests that many are unaware of the exhaustive document produced by the SCDM [116], which questions their current benefit.

Moreover, there is a 'tacit' agreement that MS Excel® is not appropriate for clinical research data management [230]. Yet, this claim can be hard to justify to investigators in resource-limited settings, who frequently use (and appreciate) spreadsheet-based tools developed by high-profile international organisations like WHO [49]. In the Congo, one investigator felt a 'moral duty' to report and share

all antimalarial trial data with WHO, regardless of whether this organisation sponsored the trial or not. Likewise, several of the schistosomiasis stakeholders invited to discuss directions for a global data platform insisted on the need for WHO or TDR to steer the project [133]. These opinions suggest high levels of trust in the WHO. Developing data management guidelines may not be WHO's role, but this organisation could certainly help to disseminate good practices among researchers in LMICs. In its recent report on how to develop the capacity for clinical research in situations of disease outbreaks, the World Bank also found limitations in the legal frameworks of many African countries in that respect, and directly called on the WHO to "develop and disseminate examples of broadly applicable legislation and policies to support and enable efficient conduct of clinical research" [64].

In parallel to these 'top-down' approaches, my qualitative work emphasised the importance of change agents as a 'bottom-up' driving force for improvement in DMS. Individuals in both Senegal and the Congo were obviously critical in "*imposing*" change or "*shaking up*" others (Chapter 10, section 10.4), with success associated to their perseverance. In their similar case studies of the barriers and enablers of clinical research in general, Franzen *et al.* also found a role for such motivated leaders in making clinical research happen locally [31]. In fact, many of their generic recommendations for clinical research capacity development quoted hereafter (and especially the three first) would greatly transfer to the case of data management, as I further explore in the next section: "(1) fostering pro-research cultures, (2) developing trial leaders and staff, (3) providing a facilitative operational environment and (4) ensuring trial research has an impact" [31].

12.3 Understanding what is needed to develop the capacity in LMICs

The third aim of my thesis was to gain an understanding of what is needed to develop the data management capacity of clinical research staff and institutions based in LMICs. I achieved this final aim by drawing on my systematic characterisation of the variability in clinical research data management practices, and on the subsequent clarification of deep underlying reasons for such variations. Importantly, I found that influential factors include (as illustrated in Figure 12.1):

- **Hard components**, such as availability of adequate technology (especially, servers enabling the use of dynamic database software), funding sustainability (hampered by limited domestic input in research financing, and by tough competition at the international level for newcomers in clinical research), or the possibility to hire data managers in research organisations (compounded by a lack of recognised educational and professional paths for that role);

- **Soft tangible components**, including competencies in project, staff and data management, which are essential to design a flexible DMS meeting study-specific requirements, and to assemble and optimally use locally-accessible resources (especially where institutional support is lacking in that respect);

- **Soft intangible components**, with a particular influence of values and norms, which underlie individuals' motivations and work performance, as well

as their sense of duty and will to promote progress in clinical research and data management practices.

In light of this greater understanding of key driving forces in the establishment and functioning of DMS, I outline below some priority areas for clinical research data management capacity strengthening, seeking to reconcile the challenges of proposing holistic strategies while also accounting for the real constraints of varied settings.

12.3.1 Developing a culture of data management crossing disease and discipline boundaries

Capacity strengthening initiatives for health research have often taken a vertical approach, whereby resources and efforts are invested on disease-specific programmes [71]. I found overall limited 'disease-specificity' of DMS, as data management tools and processes varied as much if not more within communities working on the same disease than between distinct communities. This emphasises the greater value of 'horizontal approaches', seeking to strengthen the clinical research capacity by targeting cross-cutting areas and by addressing systemic weaknesses [31, 65].

Career pathways remain fragmented for individuals in data management roles (and other clinical research roles as well [30, 43]). Building bridges between disciplines for 'generic' activities and transferable skills could create more job opportunities, thus helping to make these critical research support roles viable careers in LMIC settings (and elsewhere [3]). In parallel to multiplying job options across

medical specialities, some mechanisms are needed to enable greater (academic and other) recognition of those involved in cross-cutting areas of clinical research [3]. As a start, data managers and others whose contribution to clinical studies is significant should be more systematically included as authors on publications; and in parallel, journal publishers should provide solutions for greater transparency in the specification of 'who did what' in these often lengthy authors' lists [286].

Transforming career prospects for data managers however requires a real shift in 'culture' among the international research community, so that good data management is rewarded and contributes to professional advancement in academia, thus making clinical research data management a recognised and desirable role [3]. Through their soft tangible and intangible influence (especially in LMICs), research funders and international organisations have potential to encourage this shift in mindset among academics, by making it clearer what good data management is, and by 'imposing' best practices where their role allows.

In a similar fashion as a few advocates are building a culture of clinical research in the Congo, some key individuals are also needed to advocate for data management and ensure buy-in at different levels – institutional, but also within the clinical research community as a whole and in LMICs, where large-scale efforts to build infrastructure are needed. My qualitative data could not support more the World Bank's claim that "a supportive research culture is a sine qua non for strong clinical research capability" [64], and I would add this goes for data management too. Indeed, despite a key role for resources (skilled workforce and technology alike), my qualitative work evidenced that one-off investment in setting these up is unlikely to

bridge the capacity gap alone. Sustainable financing mechanisms for research and training are missing in many LMIC settings; and greater national engagement would be much needed to decrease the dependency on discontinued, project-specific and often disjointed international funding efforts [64, 287].

12.3.2 Targeting data management training to audiences

Anyone involved in research will handle data, and as such they should receive some (adapted) training on the critically important topic of data management. In line with the necessary clarification of roles and responsibilities mentioned above, data management tasks (from database creation to data entry and validation, etc. [101, 116]) should be fully covered in clinical research teams, and distributed in accordance with existing skills (which may require to carefully manage perceived competition and team dynamics). Training that is commensurate to each individual's contribution to the clinical study should be made available. Past work has however evidenced that research data management is rarely covered in researchers' curricula (including in HICs), and that existing training courses do not meet the varied needs of the many workers involved in that enterprise [3, 4].

In particular, advanced clinical research data management courses are needed for computer scientists and data managers to fill their potential knowledge gaps, in light of their initial and varied backgrounds. Competency-driven curricula and flexible modules, aligned with recent and evolving competency frameworks for different data staff [101, 103], should help the latter to identify appropriate training and to

develop the right skills. In addition to core and advanced data management concepts, training for data staff could include aspects of pedagogy, so that competent individuals become empowered to teach others, thus extending the reach of initial training efforts following a 'train-the-trainer' model [288].

In parallel, all clinical researchers must receive at minimum basic training on data management. Training targeted to investigators should emphasise what they have to do themselves in terms of data management, but also what they would better delegate, in light of the specific and complex skills that data managers bring to the clinical research team (and despite the occasional challenges of delegating and sharing credit). Of note, I emphasised this point in the introductory data management course I delivered in Dakar, and received positive feedback from one local data manager, with whom a young researcher got in touch for advice as a result of attending my course. Clinical investigators should also be equipped with the skills to communicate with data managers on specific constraints related to their area of expertise. While capacity strengthening needs cut across disease areas, some characteristics that are intrinsic to each disease and study design may call for adaptations in data management tools and processes from one to the next. Clinical investigators should be able to guide data managers in that respect, so that everyone's expertise is used at best to design robust DMS.

I recognise that this ideal, collaborative approach to data management would not work in all LMIC settings, especially where specialist data managers are a rare resource and investigators have to plan and oversee the data flow themselves. In such settings, it is arguably even more important that data management becomes an

integral part of research methodology courses. Major training needs highlighted in this mixed-methods study include version control and audit-trailing of study datasets and documentation. I also found in both the literature and among participants I interviewed some confusion between EDC (where the data is directly created and captured in an electronic format [116,130]) and decentralised data entry (where data is first collected on paper and entered remotely at the collection site, the entered data thus becoming the source data for verification and monitoring, as done at some of the PERCH study sites [44]). This mixed understanding of EDC or other data management models suggests a need to raise awareness of what each enables.

Ideally, training should take account of existing organisational hierarchies, ICT infrastructure and technological resources. It is unrealistic to expect a clinical research organisation to hire and train a specialist for each data management task in a context of limited HR and finances. In such settings, all-encompassing competency frameworks for the clinical research team like TDR's can help ensuring all skills and tasks are still covered, but flexibly distributed across available workers [276]. As another example, training individuals on top-notch data management software will have limited impact if they cannot implement this tool in their institution or country. In such cases, workshops with a 'participatory action research' component [289, 290] (as further described in the next section) may be more beneficial as a capacity strengthening strategy. Clinical research teams may in fact need support to accommodate new technologies, as suggested by the mixed reactions and relative tension following the adoption of EDC in Senegal.

12.3.3 Empowering change agents

As emphasised by Potter and Brough [70], capacity development requires joint action at the individual and institutional levels. Trained individuals need an enabling institutional environment to apply their skills, and conversely institutions can hardly progress without impulse from skilled workers. My qualitative case studies and those of Franzen *et al.* [31] demonstrated the critical role that individual change agents play in attracting resources and inducing improvement from the bottom-up. They can also serve as role models for others, and lead to further improvements through the snowball effect of motivating others to follow suit.

Identifying such individuals and empowering them to further this change agent role could have great impact in strengthening DMS. For example, such 'champions' could work with support from 'experts', to identify key areas in need of improvement in their own setting, and to develop solutions addressing both the requirements of 'best practices' and the constraints of their environment (following a participatory action research approach [289, 290], already tried for clinical research monitoring [291]). In other words, individuals with prior awareness of data management, and with the potential to play a change agent role in their home institution, could be target recipients of advanced training in data management.

Training would further include brainstorming sessions facilitated by a researcher with expertise in capacity development, where 'students' are invited to reflect on the potential barriers (and enablers) towards implementing best practices and systems within their setting. Working together with other students and course facilitators,

participants could develop concrete action plans for their own institution. Maintaining a connection with the 'action research lead' after the workshop, students could start implementing the conceived plan and exchange with their fellows about lessons learnt, and ask advice to overcome unanticipated barriers. This model amounts to establishing a 'community of practice', whereby likely-minded professionals come together and share their experiences for the benefit of all [292–294]. The Global Health Network and its online science park demonstrate the demand for such networking and knowledge exchange opportunities among clinical researchers [295].

The success of this approach remains strongly reliant on funding availability, to run workshops bringing together change agents in the first place, and to invest in institutional DMS under the lead of these empowered individuals once they go back to their home institution. Furthermore, mechanisms would be necessary to avoid exclusion of institutions where no change agent can be identified, so that capacity strengthening initiatives are equitable and fair, and reach those in greatest need.

12.4 Conclusion

Taking a step back from the active debate around current issues with research data sharing [8, 12–14], this thesis set out to explore the challenges and opportunities of the prior data management step, in the context of clinical research. Investigators' practices being largely undocumented in the literature, this work started with a systematic depiction of the topic, drawing on different disease communities and countries. Although the impact on data quality of different approaches to data

management could not be explored, the exploration of published articles, study datasets and accompanying documentation pointed towards sub-standard practices, which call into question the DMS ability to guarantee the integrity of end-data. In particular, my mixed-methods study emphasised that researchers in LMICs (and elsewhere) are either unaware of the importance of annotating their clinical study data (through the use of meta-data or variable dictionaries), or they do not know how to achieve this task with the tools currently available to them [4, 233, 234].

Developing skills and ICT infrastructure for the appropriate use of tools appeared critical in LMIC settings. Personal preferences and competencies accessible at the institutional or national level strongly guide the choice of data management software; but ultimately, the possibility to deploy adequate technologies for even most experienced data managers rests on the potentialities of underlying infrastructure (local or remote servers and networks). My mixed-methods study (and especially, its qualitative part) demonstrated that strengthening skills in African countries will jointly require a better understanding of the clinical research data manager role (and other data staff roles), to create curricula meeting competency needs, and to streamline career paths for this under-recognised academic role [3]. This requires further advocating for good data management in clinical research using bottom-up and top-down approaches – both of which showing potential for impact according to my results. Through advocacy, academic institutions, nations and the international health research community must realise the necessity to commit resources and develop sustainable HR and ICT infrastructure for clinical research data management. Despite remaining practical issues, data sharing is becoming an accepted

and integrated part of research work [9, 19, 21, 34, 35], and a similar culture for the advancement of good practices in research data management is long overdue.

The challenges and opportunities of data management in LMICs had previously been investigated in the context of 'routine' data collection [73, 79–82, 85–87]. In parallel, health research capacity development is building up attention [30,31,48,64, 65,68,69]. The novelty of this thesis resides in the exploration of a focussed area at the intersection of these important fields, thus bringing the neglected topic of clinical research data management capacity development in the spotlight. The lack of specific literature with which to confront my results and the focus on just a few cases may limit the strength and transferability of my findings; nonetheless, I could identify general underlying trends in clinical research data management, thus providing a starting point for further empirical work. Furthermore, the conclusions I reached largely echo those from the two above-mentioned fields, with a particular need for systemic approaches to capacity development, starting from the establishment of enabling environments where the significance of health research data management is finally acknowledged.

Part V

Appendices

Appendix A

Description of datasets compiled for the schistosomiasis scoping review and meta-analysis

Eleven data tables were created as part of the work on schistosomiasis (Table A.1); first for the quantitative synthesis (datasets A and 1 to 5), then for the meta-analysis (datasets B and 6 to 8). This appendix describes the contents of each data table, and includes a dictionary of variables and some clarifications on the different row-units. The data quality assurance procedure followed to limit mistakes in the final datasets is also described (section A.4). Datasets underlying the quantitative synthesis are provided as supporting documentation to Julé *et al.* [166].

Table	Description	File name	Appl.
A	Search results and eligibility screening (1)	sDB-0search.csv	QUAN
B	Eligibility screening (2)	sDB-article.csv	META
1	Study design	sDB-1design.csv	QUAN
2	Setting (site and year)	sDB-2setting.csv	QUAN
3	Methods underlying outcome assessment	sDB-3outcomes.csv	QUAN
4	Participants' characteristics	sDB-4participants.csv	QUAN
5	Assigned treatment(s)	sDB-5treatment.csv	QUAN
6	Study-level information (design and methods)	sDB-study.csv	META
7	Reported outcomes by group	sDB-grp-eff.csv	META
8	Detailed spatial data by site	sDB-grp-loc.csv	META

Table A.1: **List of data tables created as part of the schistosomiasis scoping review and meta-analysis**. Data files are (or will be) available online upon publication of results. Appl.: applies to; QUAN: quantitative synthesis; META: meta-analysis.

A.1 Datasets A and B: Search results and two-stage screening process

The dictionary below describes the variables (columns) found in data tables A and/or B, as specified. Both datasets contain one row per reference (article or conference abstract) screened for eligibility in the corresponding part: A for the quantitative synthesis, B for the meta-analysis. The main purpose of those two data tables was to provide a track record of the eligibility screening process. Therefore, datasets A and B document the reasons for exclusion or inclusion of each reference examined, which then enabled creating the PRISMA diagram (Figure 5.1).

Variable dictionary for schistosomiasis work – Datasets A and B
16th Oct. 2017 [version 2.0]

- Publication reference and source

VARIABLE	DESCRIPTION	DATASET
X	Unique row identifier (X00010, X00020, etc.) given to the reference at entry into the database. This identifier X enables linkage between datasets. NB: A reference may require several rows to capture information on all cohorts in other datasets (see esp. 'Design' dataset for details), in which case children rows will be numbered X0001**1**, X0001**2**, etc.	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]
X0_ref	Free text: Reference of the journal article or conference abstract, including all authors and full title. Formatting style: 'Elsevier Harvard (with titles)'.	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]
X1_source	Factor variable, reflecting the source(s) of the reference (several may apply): • "**published secondary analyses**" – reference found in the bibliography of at least one of the 5 previously published meta-/pooled analyses used as sources and manually screened for references; • "**haematobium**, **mansoni** or **japonicum**" – reference found at search for studies on *S. haematobium* (respectively, *S. mansoni* and *S. japonicum*), performed on 27 May 2015; • "**MFQ/OXA**" – reference found in the additional, separate search for studies on mefloquine and oxamniquine, performed on 12 June 2015.	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]
X2_level	Factor variable, specifying if eligibility of the reference was assessed at the level of the "**full text**" (journal article or conference record) or using only a "**summary**" (abstract).	**(A)** sDB0_search.csv [quantitative synthesis]
X0_PMID	PubMed ID, with hyperlink to the corresponding PubMed webpage. Leave blank if the study was screened at the summary level only; enter "not found" if the full text was accessed but no corresponding PMID exists.	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]

- Landscaping exercise eligibility screening

VARIABLE	DESCRIPTION	DATASET
X2_screen	Factor variable, describing reasons for excluding the reference before (pre-screen) or after (post-screen) detailed assessment of its content (see below for details).	**(A)** sDB0_search.csv [quantitative synthesis]
X3_treat	Dummy variable (**1** = Yes/**0** = No): Did the study involve the delivery of an antischistosomal drug (experimental or registered) to human participants? Of note, the study may have taken advantage of episode(s) of PC/MDA for drug delivery to participants. Interventions, other than drugs aiming at clearing the infection with *Schistosoma* (e.g. preventive *Schistosoma*-repellent cream, clinical management/surgery procedures, especially to mitigate symptoms), are out of scope (excluded at pre-screening).	**(A)** sDB0_search.csv [quantitative synthesis]
X4_pre (respectively, **X4_post**)	Dummy variable (**1** = Yes/**0** = No): Did the study involve screening of participants for infection with a species of *Schistosoma* shortly before (resp., after) delivery of the intervention? The screening procedure should involve a recognised diagnostic test (egg count in urine/stool, Circulating Cathodic Antigen – CCA). Isolated surveys for prevalence monitoring, as well as studies screening different sets of participants before and after treatment, are out of scope ("**0**").	**(A)** sDB0_search.csv [quantitative synthesis]
X5_inTime	Dummy variable (**1** = Yes/**0** = No). At least one outcome assessment (post-treatment diagnostic test) occurred within 60 days (including, exactly 60 days or 2 months) after treatment, where administration of the first study drug counts as Day 0.	**(A)** sDB0_search.csv [quantitative synthesis]
X6_eligible	Auto-filled variable, establishing whether the study is overall eligible. Equals **1** (= Yes) if and only if **X3_treat** + **X4_preTest** + **X4_postTest** + **X5_inTime** = 4 and **X2_screen** = "none"; otherwise, equals **0**.	**(A)** sDB0_search.csv [quantitative synthesis]
X7_dup	Factor variable, listing references that seem to be duplicated reports of the same study. Reasons for excluding/including potential duplicates are discussed in **#notes** (see below).	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]

- Meta-analysis eligibility screening

VARIABLE	DESCRIPTION	DATASET
X6_PZQ40	Dummy variable (**1** = Yes/**0** = No): Was praziquantel delivered at the PC/MDA dose and regimen (i.e. once 40 mg/kg)?	**(B)** sDB_article.csv [meta-analysis]
X6_5to19yrs	Dummy variable (**1** = Yes/**0** = No): Did the study participants include some from the 5-19 years age group (or schoolchildren, even if ages unknown)? Leave blank if non-eligibility was already ascertained from previous question(s).	**(B)** sDB_article.csv [meta-analysis]
X6_CRreport	Dummy variable (**1** = Yes/**0** = No): Was at least a Cure Rate (CR) reported as a measure of drug efficacy (within the relevant age group)? Leave blank if non-eligibility was already ascertained from previous question(s).	**(B)** sDB_article.csv [meta-analysis]
X6_exclu Reason	Semi-free text: Other reason for excluding the study from the meta-analysis (as opposed to previous inclusion in the scoping review). Leave blank if non-eligibility was already ascertained from previous question(s).	**(B)** sDB_article.csv [meta-analysis]
X6_retain	Dummy variable (**1** = Yes/**0** = No): Overall, is the study eligible for the meta-analysis?	**(B)** sDB_article.csv [meta-analysis]

- Has the reference been analysed and coded, and were issues encountered?

VARIABLE	DESCRIPTION	DATASET
X8_status, **X9_status**	Factor variable, capturing the progress of the review: "To be Analysed"; "**Analysed**"; "**Analysed (conference)**"; "**Excluded**"; "**Excluded (unclear)**"; "**Exclude (meta-analysis)**"; "**Inaccessible**"; "**Inaccessible (unclear)**".	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]
#date	Date (**DD/MM/YYYY**) of last update of the complete reading, analysis and coding of that reference.	**(A)** sDB0_search.csv [quantitative synthesis]
#notes, **#notes-MA**	Free text: Comments on problems encountered during analysis and coding of the article (missing or unclear data, impossibility to record the data in the desired format, inconsistency between different sources, etc.). Comments may also include details on reasons for inclusion/exclusion of studies whose eligibility was difficult to assess. Leave blank if no comment is required.	**(A)** sDB0_search.csv [quantitative synthesis] **(B)** sDB_article.csv [meta-analysis]

Detailed explanation of "**X2_screen**" variable and decision-tree for study eligibility assessment

Some references can be excluded quickly, before even reading their full summary. The "pre-screens" are successive 'checkpoints', which a reference must all go through before progressing to further assessment.

- "**Pre-screen 0: conference prior 2014**" – excluded because 'too old', as judged from the dates of the conference
- "**Pre-screen 1: case report**" – excluded because it reports on a single or series of case(s), but does not involve execution of a pre-planned clinical study
- "**Pre-screen 2: review**" – excluded because it is a review of the literature (textbook chapters, scoping or systematic review, meta-analyses, etc.), or a comment/opinion paper, but does not involve primary data collection
- "**Pre-screen 3: in vitro/ex vivo, laboratory work, no human subject**" – excluded because the study is performed in the laboratory and/or does not involve human subjects
- "**Pre-screen 4: no *Schistosoma spp.* infection, no antischistosomal drug delivery, non-endemic setting**" – excluded because the study is not on *Schistosoma spp.* (no participant assessed for this infection), or the intervention delivered is not aiming to clear infection with this parasite, or the study is not performed in the endemic setting where infection occurred

When progressing to more detailed assessment, excluded studies are categorised in one of 6 broad "topics" to clarify the reason for their exclusion.

- "**Topic 0: individual patient data for drug efficacy, with late outcome assessment**" – fulfils all inclusion criteria, except that the outcome measure was taken too late (>60 days post-treatment)
- "**Topic 1: associated morbidity and immunological responses, co-infections, risk factors**" – IPD would not enable antischistosomal efficacy assessment; the study focusses on symptoms, co-infections and morbidity associated to schistosomiasis, or on immunological responses to infection and/or treatment
- "**Topic 2: diagnostic approach, prevalence estimation method**" – IPD would not enable antischistosomal efficacy assessment; the study focusses on development of new diagnostic assays, or on optimised use of existing tests into the field (e.g. for epidemiological mapping)
- "**Topic 3: behaviour, education, qualitative research, implementation piloting and safety**" – did not involve diagnostic testing of participants, as the focus is on the safety/accuracy of programme delivery (e.g. multiple drugs within national preventive chemotherapy, dose assessed using height as a proxy for weight), or on behavioural/educational factors associated to (re-)infection
- "**Topic 4: cost-effectiveness, modelling, projections and simulations *in silico***" – did not focus on treatment efficacy data collection into the field, but on mathematical modelling (using past, secondary data or not)
- "**Topic 5: endemicity, disease mapping, surveillance, epidemiological survey**" – did not involve a pre-/post-treatment diagnostic assessment (or at least not at the individual level), as the focus is on prevalence estimation and monitoring

Finally, some studies are excluded after detailed assessment, and reason for their late exclusion is explained by the "post-screens".

- "**Post-screen 1: study end prior 2000**" – although published within the chosen timeframe, the study is too 'old', i.e. completed before 1st January 2000
- "**Post-screen 2: redundant, several references reporting on same study**" – several articles reporting on the same study were identified, and only the most exhaustive one was included

If no reason was found to exclude the reference, enter "none". If the situation is still unclear (e.g. due to inaccessible full text), enter "unknown".

A.2 Datasets 1 to 5: Quantitative synthesis data

Data synthesis and descriptive analyses performed for the scoping review rely on five data tables, all of which have one row per cohort of participants, as defined in Halder *et al.* [157] (Table A.2).

Term	Definition
Study	A study or trial, where a pre-defined protocol was followed for all participants, with the exception of differences between cohorts and arms.
Cohort	Groups within studies are considered as separate cohorts if at least one of these reasons applies: there were differences in the protocol followed for different groups of participants (other than those covered by different intervention arms); study recruitment took place in multiple distinct time periods, or participants were in different countries; participants were in *a priori* defined categories which depended on participant characteristics (e.g. diagnosis of different soil-transmitted helminth infections or other co-infections; ethnicity).
Arm	Different arms of a study often correspond to different drug regimens under comparison. Study- and cohort-level data are the same for different arms; only treatments/interventions vary. The arms are such that participants (or clusters, as per design) could be randomized to them (though randomization was not implemented in all studies).

https://doi.org/10.1371/journal.pntd.0006053.t001

Table A.2: **Glossary of key terms from Halder *et al.*** [157], reproduced from their article in PLoS Neglected Tropical Diseases.

Concretely, the definitions translated into clear conditions outlined in the box below, which facilitated consistent data extraction.

A **study** is counted as *N* different **cohorts** (i.e. corresponds to *N* rows in the database) if and only if at least one of the 4 following conditions apply:

(1.) Multi-centric study, with centres in distinct countries
→ 1 cohort per country

(2.) Multi-centric study, with centres possibly in the same country but infested with different *Schistosoma* species (including mono- *vs.* mixed endemicity sites)
→ 1 cohort per endemic setting

(3.) The study clearly differentiates *a priori* between different categories of participants to be studied, in particular in terms of age group (e.g. mothers/children, infants/schoolchildren)
→ 1 cohort per participants' group

(4.) The protocol is otherwise amended depending on site/group (e.g. different diagnostic approach, different follow-up time-points)
→ 1 cohort per set of study procedures

Variable dictionary for schistosomiasis work – Dataset 1
17th Oct. 2017 [version 2.0]

- Study identification

A1_trialID	Trial identifier, if the trial was registered and registration could be found (i.e. mentioned in the publication, or identified through search for 'schistosomiasis' + 'country of study' in ClinicalTrials.gov, Controlled-Trials.org and WHO ICTRP registries). Enter "not found" otherwise. In alphabetical order of registry, and separated by semi-colon if several registries apply.
A1_trial Registry	Factor variable, corresponding to the name of the trial registry: "**ClinicalTrials.gov**"; "**Controlled-Trials.com (BioMed Central)**"; "**Pan African Clinical Trials Registry (PACTR)**". Enter "not found" otherwise. In alphabetical order and separated by semi-colon if several registries apply.
#studyTitle	Scientific title of the study/trial, as stated at registration. Enter "not found" otherwise.
#cohortsJustif	Free text: justification for counting a single 'study' reported in this article/abstract as several 'cohorts' = set of participants sharing the same study meta-data characteristics (see box above for details).
A0_S	Unique study identifier (S1, S2, etc.), derived from the row identifier (X00010, X00020, etc.). If two or more rows are required to capture information on a same study (i.e. multiple cohorts), then they must have the same identifier A0_S. NB: In the rare case where several studies are reported in the same publication, record them as S1a, S1b, S1c, etc.
A0_C	Unique cohort identifier (S1C1, S2C1, etc.) given as soon as a new cohort (set of participants) is added to the database.

- What were the study objectives?

A2_efficacy	Dummy variable (**1** = Yes/**0** = No). Was the study designed primarily to assess efficacy or effectiveness of the treatment? NB: Some studies were conducted for another purpose (e.g. diagnostic sensitivity assessment, immunology studies), in which case they score "**0**"; but they still collected data that would enable drug efficacy assessment in secondary analyses using IPD, hence their relevance to the landscaping exercise.
A2_safety	Dummy variable (**1** = Yes/**0** = No). Did the study assess the safety or tolerability of the treatment (i.e. collected data on side effects)?
A2_pharma	Dummy variable (**1** = Yes/**0** = No). Did the study involve some pharmacokinetics/pharmacodynamics (PK/PD) measurements?
A2_other	Factor variable, describing other objectives of the study, namely: • "**diagnostic approach**": diagnostic test study (e.g. sensitivity assessment), including search for biomarkers of infection which could potentially lead to new diagnostic approaches; • "**immunological responses**": assessment of specific antibody responses, etc.

	• "**nutrition**": impact of nutrient supplementation on treatment efficacy or other outcomes (e.g. performance at school). If several levels apply, enter them in alphabetical order, separated by "+". Enter "none" otherwise.
#objComment	Free text: Additional comment/clarification on study type or aims, e.g. nested study, trial phase, significance for national programme. Enter "none" otherwise.

- How many arms did the study involve, including control-arm?

A3_control	Dummy variable (**1** = Yes/**0** = No). Was the intervention assessed against a comparator (or, did the study involve >1 arm)? NB: Some studies are formally considered 'non-controlled' but they include a 'comparison' (e.g. of two age groups/settings), in which case they score "**1**".
A3_ctrlCat1, 2	Factor variable, describing the category of comparator used in controlled studies. • "**drug comparison**"; "**dose comparison**"; "**regimen comparison**" – comparison of several antischistosomals against each other • "**placebo**"; "**delayed treatment**"; "**untreated controls**"; "**healthy controls**" • "**other** (**endemic context**, **season**, **treatment history**, or **education**)" – e.g. "other (endemic context)", for a comparison between low/high or new/old focus of transmission If several levels apply, record them in the order suggested above. Enter "unclear" otherwise, or "N/A" for non-controlled studies and in superfluous columns. NB: studies falling only into the category "other" are rather 'comparative' than 'controlled' studies (see also **A3_control**).
A3_arms	Total number of arms in the study, including comparative groups if applicable.
#armComment	Free text: Description of the different arms, especially for unusual/complex designs. Enter "none" otherwise.

- Was the assignment to different arms randomised and blinded?

A4_random	Dummy variable (**1** = Yes/**0** = No). For multi-arms studies, was the assignment to one arm or the other randomised? If not explicitly specified, the assumption is "**No**".
A4_rand Method	Semi-free text: method of randomisation into the study if applicable. (e.g. "**computer-generated sequence**"). Enter "unclear" otherwise, or "N/A" for non-randomised studies.
A5_blind	Dummy variable (**1** = Yes/**0** = No). For randomised studies, was the person delivering the drug and/or assessing the participant blinded to the intervention administered? If not explicitly specified or if drugs were distinguishable, the assumption is "**No**".
A5_blind Level	Factor variable, describing the level of blinding used in the study if applicable: "**assessor-blinded**"; "**carer-blinded**"; "**double**"; "**triple**". Enter "unclear" otherwise, or "N/A" for non-blinded studies.

Variable dictionary for schistosomiasis work – Dataset 2
17th Oct. 2017 [version 2.0]

- Where was the study conducted/cohort recruited?

B0_country	Full name of the country of study – UN classification: https://unstats.un.org/unsd/methodology/m49/
#siteLocation	Free text: most precise name and address of each and every site of recruitment, including reported longitude and latitude GPS coordinates, if available.
B1_siteCat	Factor variable, describing the category of the site chosen as a recruitment centre of target participants: "**school**"; "**hospital/health-centre**"; "**other**" (village/city, households, etc.). Enter "unclear" otherwise.
B1_setting	Factor variable, describing the setting where recruited participants live: "**rural**"; "**semi-urban**"; "**urban**". Use Google Earth if unclear from publication. If several apply, enter them in increasing level of urbanisation, separated by "+". Enter "unclear" otherwise.
B0_region	World region of the country of study – UN classification (same as above)
B0_income	Income group of the country – World Bank: http://data.worldbank.org/country
B2_endem	Factor variable, reflecting the prevalence of infection with *Schistosoma* in the study area – from "**0**" (prevalence of 0 to 9.99%) to "**9**" (90 to 99.99%), as found at screening (alternately, past estimates) based on egg count in excreta. Consider only the highest prevalence known for mixed endemicity contexts and/or if several sites are involved. If the prevalence is only qualitatively reported, enter "**0***" for 'low', "**4***" for 'medium/intermediate' and "**8***" for 'high' prevalence. Enter "-9" if unknown.
B2_species1	Factor variable, describing the *Schistosoma* species of interest at the study site.
B2_species2	Factor variable, describing the second *Schistosoma* species present at the study site, if applicable. If the prevalence of this second species among participants was checked only at baseline but not at follow-up, append "[partial data]". If the prevalence of this second species was checked and found to be null, append "[absent]", or possibly "[rare]". Enter "N/A" otherwise.

- When was the study conducted and reported?

B3_studyStart	Start date of study (**DD/MM/YYYY**). If the exact date is unknown, then, arbitrarily, DD=15 and MM=03. Enter "unknown" otherwise.
B4_studyEnd	End date of study (**DD/MM/YYYY**). If the exact date is unknown, estimate it from start date and duration of follow-up, or, arbitrarily DD=15 and MM=09. Enter "unknown" otherwise.
B5_reportDate	Date of publication (earliest known, i.e. chose 'ePub ahead of print' date if available), or of presentation at a conference. If exact day unknown, DD=15. If exact month unknown, chose 'manuscript accepted' date + 1 month.
B5_report Means	Factor variable, describing the medium with which research results were reported: "**journal article**"; "**conference**".

- What (standard) analyses were performed to diagnose urinary schistosomiasis?

#urinary Standard	Semi-free text: description of the sample collection strategy and of the laboratory test performed to diagnose infection with *S. haematobium* – e.g. "2 to 3 samples (only 1 at baseline) collected between 10.00 am and 2.00 pm; 1 repeat (?); 10 mL filtration". Enter "none" if no diagnosis of urinary schistosomiasis by egg count was performed.
C0_uSamp	Minimum number of independent urine samples collected (and required for consideration of the participant in the analysis) to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.
C1_uRep	Minimum number of laboratory test repeats made on one same sample to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.

- What (standard) analyses were performed to diagnose intestinal schistosomiasis?

#intestinal Standard	Semi-free text: description of the sample collection strategy and of the laboratory test performed to diagnose infection with *S. mansoni/japonicum* – e.g. "aim for 3 samples but 2 tolerated; 2 repeats; Kato-Katz thick smears (cellophane, 50 mg)". Enter "none" if no diagnosis of intestinal schistosomiasis by egg count was performed.
C0_iSamp	Minimum number of independent stool samples collected (and required for consideration of the participant in the analysis) to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.
C1_iRep	Minimum number of laboratory test repeats made on one same sample to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.
C3_iTech1, 2	Factor variable, describing the technique employed to analyse stools. "**Kato-Katz thick smears (XX mg)**"; "**Kato-Katz method** (= technique, test)"; "**Kato-Katz smears (modified)**"; "**Katz thick smears**"; "**Kato thick smears (XX mg)**"; "**Kato technique (qualitative)**"; "**concentration (formaldehyde)**"; "**hatching test**"; "other". Information in brackets may or not be recorded depending on details in the publication or lack thereof. If several apply, use additional columns and record in the order suggested above. Enter "N/A" if no test was performed and in superfluous columns.

- What other tests were performed to diagnose and characterise schistosomiasis, as well as other biomedical assays (for additional, secondary outcome measures)?

C4_testClinic	Semi-free text: Additional point-of-care (POC)/rapid test or clinical assessment to diagnose and characterise schistosomiasis (including symptoms); mainly biopsy, dipsticks/reagent strips for haematuria, proteinuria or leukocyturia, visual aspect of excreta, POC circulating antigen, ultrasonography. Enter "none" otherwise.

C4_testLab	Semi-free text: Additional molecular biology assay performed to diagnose and characterise schistosomiasis, mainly ELISA and PCR. Precise the source of biological material (blood, stool, urine…), and the type of molecules assessed in molecular assays (e.g. *Schistosoma* DNA, soluble egg antigen…). Enter "none" otherwise.
C5_other Assays	Factor variable, describing other biomedical assays: "**Hb concentration**"; "**other serology/immunology assays**" (serology for HIV/AIDS diagnosis, cytokines' profile after PBMC culture, etc.); "**other blood analyses**" (e.g. haematocrit, iron absorption).

- When were participants followed-up for outcome assessment(s)?

D1_T1, 2, 3, last	Time of first (respectively, second, third or latest) follow-up, in days, counted from the day of first treatment dose (Day 0), up to the day of first re-sampling. If the time is specified in weeks, months, or years in the publication, use 1 week = 7 days, 1 month = 30 days and 1 year = 365 days to convert. Enter "none" if only one (resp. 2 or 3) time point(s) applies.

- How was the primary, drug efficacy endpoint expressed?

D3_CR	Dummy variable (**1** = Yes/**0** = No). Was the cure rate (CR) calculated at any time of follow-up? CR: percentage of participants cured (i.e. excreting no egg) at follow-up, among participants with confirmed infection (i.e. excreting eggs) at baseline.
D4_ERR	Dummy variable (**1** = Yes/**0** = No). Was the egg reduction rate (ERR) calculated at any time of follow-up? ERR: ratio between mean egg count at follow-up and at baseline, among participants with confirmed infection.
D4_intens	Dummy variable (**1** = Yes/**0** = No). Was there a quantitative comparison of pre- and post-treatment intensity of infection? This variable is particularly relevant for studies which do not report the ERR, but still report and compare the intensity of infection before and after treatment as a mean, median and/or other quantitative measure (i.e. not only qualifying the intensity as 'low' or 'heavy').
D5_eggsMean	Factor variable, describing the population mean utilised to calculate the ERR: "**arithmetic**"; "**geometric**"; "**both**". Enter "unclear" otherwise, or "N/A" if no ERR was calculated.
#endpoint Comment	Free text: Specifications on other times of follow-up that could not be recorded in dedicated columns ("**Additional time-points**"), on the way the outcome was assessed ("**Limits**" – e.g. time-point applicable to a subset of participants only, efficacy endpoint derived from other measure than egg count), or additional notes ("**NB**").

- Number of participants and type of schistosomiasis infection that they were required to have for inclusion in the study

D0_included	Total number of participants in the initial cohort; i.e. who were *included into the study and received the intended intervention* (drug(s), control, placebo, no-treatment, etc.), with the intention to follow them up for outcome assessment. Enter "-9" if unknown.
E0_infection	Factor variable, describing the infection that was necessary *to be included in the study and to receive antischistosomal treatment* (i.e. do not consider participants infected with other parasites who were specifically treated against those): "**Sh+**"; "**Sj+**"; "**Sm+**". Enter "not required" otherwise. If several infective species apply, specify into brackets which infection was "**required**" for inclusion; and which co-infection was either "**tolerated**" or "**excluded**" among study participants. If applicable, record several species in the order suggested above ("required", "tolerated", "excluded"), and then in alphabetical order of species. Separate different levels by "**and**" (if both infections were concomitantly required)/"**or**" (if one or the other infection was required), accordingly – e.g. "Sm+ (required) and Sh+ (tolerated)"; "Sh+ and/or Sm+".
D1_infected	Number of participants whose individual data would enable assessment of drug efficacy; i.e. participants who *were infected with Schistosoma spp. at baseline and received an antischistosomal drug or comparator, with the intention to follow them up* for outcome assessment within 60 days post-treatment. In some cases (especially, studies in which individuals non-infected with *Schistosoma spp.* at baseline were still included in the study), this exact number of participants is unknown, and a conservative estimate should be calculated – i.e. total included in analyses (**D0_included**) x Prevalence in area. If no correction is needed, re-enter **D0_included**. Enter "-9" if unknown.
D2_followed	Number of relevant participants (i.e. among **D1_infected**) who were followed up. If follow-up is detailed for each of the different time-points, enter the maximum number of participants followed-up. If the exact number is unknown (i.e. for studies in which individuals non-infected with *Schistosoma spp.* at baseline were still included in analyses), use the same correction factor as was used to obtain **D1_infected** from **D0_included**. Enter "-1" if not applicable (e.g. when 'provide samples at follow-up' is an inclusion criterion, meaning that the study cohort was defined *a posteriori*, based on compliance); and "-9" if unknown.
#corrected Calculation	Free text: way in which the infected number of participants (**D0_infected**) was obtained from the total of included participants (**D0_included**), if applicable. Enter "N/A" otherwise.
E1_schisto Include	Free text: other criteria related to schistosomiasis and that were applied to include participants – e.g. minimum number of eggs required. Enter "none" otherwise.
E1_schisto Exclude	Free text: other criteria related to schistosomiasis and that were applied to exclude participants – e.g. symptomatic form of the disease.

- What major (non-*Schistosoma*) co-infections were controlled for, and possibly resulted in exclusion from the study?

E2_malaria	Dummy variable (**1** = Yes/**0** = No). Was co-infection with malaria assessed (at least at baseline) using a diagnostic test?
E2_excluMal	Dummy variable (**1** = Yes/**0** = No). Did infection with malaria (as confirmed by diagnostic test or clinical assessment only) result in exclusion from the study?
E3_helminth	Dummy variable (**1** = Yes/**0** = No). Was co-infection with at least 1 other helminth assessed (at least at baseline) using a diagnostic test?
F3_excluHelm	Dummy variable (**1** = Yes/**0** = No). Did infection with at least 1 other helminth result in exclusion from the study?

- What is the age range of enrolled participants and could they be pregnant?

F0_ageMin (resp. Max)	Minimum (resp. Maximum) age, in years, among enrolled participants. Strictly speaking, this does not necessarily correspond to the age range specified in inclusion/exclusion criteria for the study. Enter "-9" otherwise.
F5_pregnancy	Semi-free text, stating first if pregnant women were "exclude[d]" or "include[d]", and how pregnancy was defined (e.g. "exclude: pregnant or lactating women, pregnancy test"). Enter "N/A" if age range (<12 years old) suggests no pregnant female could have been included; enter "unclear" otherwise.

- Specific inclusion criteria[1]

G0_inclu Healthy	Free text: general health status required for inclusion, along with criteria for judgement (e.g. "appear healthy at enrolment, as assessed by the study physician", "no diarrhoea").
G1_incluMorph	Semi-free text: weight and/or height limits for inclusion (e.g. "weight > 25 kg").
G3_inclu Compliant	Free text: ability and willingness to follow/minimal compliance to the protocol required for inclusion, especially in analyses (e.g. "able and willing to be examined by the study physician at baseline and follow-up").
G4_inclu Consent	Free text: method for securing informed consent, including who gave consent for children (parent, head of the school, etc.) and how (orally, in writing, etc.).
G5_inclu School	Free text: school grades among which children were recruited, and/or other criteria for inclusion related to the school (e.g. "be present at school on screening day" or "school accessible by road during rainy season").
G6_inclu Residency	Free text: duration of residency in the study area, and/or other criteria of inclusion related to the area of residence (e.g. "living in the study area and intending to deliver at the Entebbe Hospital" or "residing in the study area since birth").
G7_incluOther	Free text: other inclusion criteria not falling into any of the previous categories.

- Specific exclusion criteria[1]

H0_exclu Malnutrition	Free text: state of malnutrition and/or anaemia leading to exclusion, along with decision rule if available (e.g. "signs of micronutrient deficiencies" or "signs of severe malnutrition (defined as children with weight/height ratio <3 SD or <70% of the median of WHO standardised reference values, or still with symmetrical oedema affecting both feet)").
H3_excluIllness	Free text: list of severe illnesses or systematic diseases leading to exclusion.
H4_exclu Allergy	Free text: precisions on allergies leading to exclusion, essentially hypersensitivity to study drugs.
H5_exclu Medication	Free text: medication(s) taken during a specific period prior to (or during) the study and leading to exclusion, essentially any antischistosomal within the past 28 days.
H7_exclu Participation	Free text: multiple participation in the same or similar studies, or concomitant participation in another trial leading to exclusion.
H8_excluOther	Free text: other exclusion criteria not falling into any of the previous categories.

[1] When inconsistency between sources (e.g. trial registry and journal article), report most stringent criteria; unless additional information permits to say which source is accurate.

Variable dictionary for schistosomiasis work – Dataset 5
17th Oct. 2017 [version 2.0]

- Detailed treatment and corresponding number of participants in group 1, 2, etc.

#detail Regimen1, 2, etc.	Free text: exact names and doses of drugs assigned to group 1, 2, etc. Provide details on how the number of participants *treated* with that regimen was obtained, if unclear (e.g. for studies in which individuals non-infected with *Schistosoma spp.* at baseline were still given treatment). Enter "none" in superfluous columns. NB: For the purpose of this descriptive review, participants who received an antischistosomal (standalone) or the same antischistosomal together with a placebo are considered within one same group.
I0_regCat1, 2, etc.	Factor variable, describing the category of the treatment regimen. If several factors apply, enter them in the order suggested below, separated by "+". • "**PZQ (XX mg/kg)**": praziquantel and corresponding dose in mg/kg body weight – where the dose corresponds to the *overall dose received within 1 week* • "**antischistosomal (other)**": any non-PZQ, antischistosomal drug(s) • "**anthelmintic (other)**": any drug used against other helminths • "**ACT**": Artemisinin-based Combination Therapy • "**antimalarial (other)**": any non-ACT, antimalarial drug(s) • "**control**" or "**control (healthy)**": placebo or no treatment, and possibly uninfected at baseline • "**other**": any drug(s) not falling in one of the previous categories Enter "N/A" otherwise, i.e. in superfluous columns.
I1_reg1, 2, etc. **I2_reg1, 2, etc.** **I3_reg1, 2, etc.**	Abbreviated name of the first (**I1_regX**), second (**I2_regX**) and third (**I3_regX**) drug(s) given as part of the assigned regimen. Record drugs in the order suggested below. Drugs in brackets may have been given to some participants in that group, but not all (e.g. albendazole, if infected with other helminths; placebo or not in addition to an interventional drug). Enter "N/A" in superfluous columns. • Dose of praziquantel, recorded in the format "**(NxQQ):Tu**", where: ▪ N = total number of split doses received within the study. Add a star (*) if this dose is a repeat of a past treatment, or will be possibly repeated for a subset of participants, e.g. at third follow-up if treatment failure. ▪ QQ = quantity (in mg/kg) per split dose administered. ▪ Tu = time T (and its unit u) that separated 2 split doses. If several times were possible (e.g. every 4 to 6 hours), enter the shortest interval. • "**OXA**" = Oxamniquine; "**myrrh**" = Mirazid • "**As**" = Artesunate; "**AM**" = Artemether • "**A**" = Amodiaquine; "**MFQ**" = Mefloquine • "**SP**" = Sulfadoxine-Pyrimethamine; "**SMP**" = Sulfamethoxypyrazine-Pyrimethamine; "**LU**" = Lumefantrine • "**TCBZ**" = Triclabendazole; "**ALB**" = Albendazole; "**placebo**"; "**none**".
I4_treated1, 2, etc.	Number of participants infected at baseline and who fully received the treatment they were assigned. If the number of participants successfully treated is not specified, enter the number of *participants initially assigned to that arm* – i.e. assuming they all completed the treatment. When several doses of the same drug were possibly repeated, the number recorded is the number of participants who received at least 1 dose of the treatment, if known – i.e. maximum number treated with any number of doses. Enter "-9" if unknown, and "-1" if not applicable (superfluous columns).

Data management or statistics software used in schistosomiasis studies

Of note, variables regarding the software used for data management and statistical analysis are not described in the dictionaries presented above, because software-related data for the schistosomiasis work was collected and synthesised later on, using the same approach as in the malaria work (Chapter 7).

A.3 Datasets 6 and 7: Meta-analysis data

The schistosomiasis meta-analysis work relies on three data tables. Dataset 6 covers information already captured during the quantitative synthesis (except for one newly defined 'reportBias' variable): it is a copy of the subset of articles and variables relevant to both parts of the landscaping exercise. In contrast with datasets from the quantitative synthesis, Dataset 6 (sDB-study.csv) has one row per study, because for the concerned subset of studies, study-level information (related to study design and methods) was consistent across cohorts and groups. Dataset 7 captures information on exact study site location, and has one row per group; the relevant unit of analysis in line with the purpose of the meta-analysis and constraints of aggregated data. The definition of a group is provided in results (Chapter 5) rather than in methods or Appendix, as this unit dynamically evolved with increasing understanding of the complexity of comparing non-homogeneous aggregated data. Dataset 8 captures information on the reported drug efficacy, and has one row per group considered in the meta-analysis.

Variable dictionary for schistosomiasis work – Dataset 6
17th Oct. 2017 [version 2.0]

- General information and study descriptors

SID	Unique study identifier (X0001, X0002, etc.) – this corresponds to the first 5 digits of the X identifier (see dictionary for Datasets A and B above).
AuthorDate	Factor variable, corresponding to the name of the first author and publication year of the corresponding article.
A1_trialID	Trial identifier, if the trial was registered and registration could be found (i.e. mentioned in the publication, or identified through search for 'schistosomiasis' + 'country of study' in ClinicalTrials.gov, Controlled-Trials.org and WHO ICTRP registries). Enter "not found" otherwise. In alphabetical order of registry, and separated by semi-colon if several registries apply.

- What were the study objectives?

A2_efficacy	Dummy variable (**1** = Yes/**0** = No). Was the study designed primarily to assess efficacy or effectiveness of the treatment?
A2_safety	Dummy variable (**1** = Yes/**0** = No). Did the study assess the safety or tolerability of the treatment (i.e. collected data on side effects)?
A2_other	Factor variable, describing other objectives of the study, namely: • "**diagnostic approach**": diagnostic test study (e.g. sensitivity assessment), including search for biomarkers of infection which could potentially lead to new diagnostic approaches; • "**immunological responses**": assessment of specific antibody responses, etc. • "**nutrition**": impact of nutrient supplementation on treatment efficacy or other outcomes (e.g. performance at school). If several levels apply, enter them in alphabetical order, separated by "+". Enter "none" otherwise.
#objComment	Free text: Additional comment/clarification on study type or aims, e.g. nested study, trial phase, significance for national programme. Enter "none" otherwise.

- What arms did the study involve?

A3_control	Dummy variable (**1** = Yes/**0** = No). Was the intervention assessed against a comparator (or, did the study involve >1 arm)? NB: Some studies are formally considered 'non-controlled' but they include a 'comparison' (e.g. of two age groups/settings), in which case they score "**1**".
A3_ctrlCat1, 2	Factor variable, describing the category of comparator used in controlled studies. • "**drug comparison**"; "**dose comparison**"; "**regimen comparison**" – comparison of several antischistosomals against each other • "**placebo**"; "**delayed treatment**"; "**untreated controls**"; "**healthy controls**"

	• "**other** (**endemic context**, **season**, **treatment history**, or **education**)" – e.g. "other (endemic context)", for a comparison between low/high or new/old focus of transmission If several levels apply, record them in the order suggested above. Enter "unclear" otherwise, or "N/A" for non-controlled studies and in superfluous columns. NB: studies falling only into the category "other" are rather 'comparative' than 'controlled' studies (see **A3_control**).
A3_arms	Total number of arms in the study, including comparative groups if applicable.
#armComment	Free text: Description of the different arms, especially for unusual/complex designs. Enter "none" otherwise.

- Was assignment to different arms randomised and blinded?

A4_random	Dummy variable (**1** = Yes/**0** = No). For multi-arms studies, was the assignment to one arm or the other random? If not explicitly specified, the assumption is that the study was not randomised.
A4_rand Method	Semi-free text: method of randomisation into the study if applicable. (e.g. "**computer-generated sequence**"). Enter "unclear" otherwise, or "N/A" for non-randomized studies.
A5_blind	Dummy variable (**1** = Yes/**0** = No). For randomized studies, was the person delivering the drug and/or assessing the participant blinded to the intervention administered? If not explicitly specified or if drugs were distinguishable, the assumption is that the study was not blinded.
A5_blind Level	Factor variable, describing the level of blinding used in the study if applicable: "**assessor-blinded**"; "**carer-blinded**"; "**double**"; "**triple**". Enter "unclear" otherwise, or "N/A" for non-blinded studies.
reportBias	Dummy variable (**1** = Yes/**0** = No). Does this study potentially suffer from a bias of "over-reporting" of positive results (i.e. reporting 100% CR, where the study did not originally aim at assessing drug efficacy but the article reported that all treated participants were cured)?

- Standard analyses: urinary schistosomiasis

#urinary Standard	Semi-free text: description of the sample collection strategy and of the laboratory test performed to diagnose infection with *S. haematobium* – e.g. "2 to 3 samples (only 1 at baseline) collected between 10.00 am and 2.00 pm; 1 repeat (?); 10 mL filtration". Enter "none" if no diagnosis of urinary schistosomiasis by egg count was performed.
C0_uSamp	Minimum number of independent urine samples collected (and required for inclusion of in the study) to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.

C1_uRep	Minimum number of laboratory test repeats made on one same sample to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.

- Standard analyses: intestinal schistosomiasis

#intestinal Standard	Semi-free text: description of the sample collection strategy and of the laboratory test performed to diagnose infection with *S. mansoni/japonicum* – e.g. "aim for 3 samples but 2 tolerated; 2 repeats; Kato-Katz thick smears (cellophane, 50 mg)". Enter "none" if no diagnosis of intestinal schistosomiasis by egg count was performed.
C0_iSamp	Minimum number of independent stool samples collected (and required for inclusion in the study) to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.
C1_iRep	Minimum number of laboratory test repeats made on one same sample to perform diagnosis at follow-up(s). Assume 1 if unclear; enter "N/A" if no test was performed.
C3_iTech1, 2	Factor variable, describing the technique employed to analyse stools. "**Kato-Katz thick smears (XX mg)**"; "**Kato-Katz method** (= technique, test)"; "**Kato-Katz smears (modified)**"; "**Katz thick smears**"; "**Kato thick smears (XX mg)**"; "**Kato technique (qualitative)**"; "**concentration (formaldehyde)**"; "**hatching test**"; "other". Information in brackets may or not be recorded depending on it being detailed in the full text or not. If several apply, use additional columns and record in the order suggested above. Enter "N/A" if no test was performed and in superfluous columns.

- Type of parasitic infection

E1_inclu Parasites	Semi-free text: type of parasitic infection required (or tolerated, regarding other helminths) for participants to be included in the corresponding study – e.g. minimum number of egg output at baseline.
E1_exclu Parasites	Semi-free text: type of parasitic infection leading to exclusion of a participant from the corresponding study – e.g. symptomatic schistosomiasis, co-infection with malaria/helminths.

Variable dictionary for schistosomiasis work – Dataset 7
17th Oct. 2017 [version 2.0]

- General information and study/group descriptors

MID	Unique row identifier within this meta-analysis dataset (M001A, M001B, M002A, etc.): one per distinct study group for which an antischistosomal drug efficacy measure was reported – i.e. mainly one new row per arm, per species, and per separately studied geographical location. NB: The numerical part of the identifier (001, 002, 003…) is incremented by one for each new study; and the alphabetical part of the identifier (A, B, C…) is incremented by one for each new study group.
SID	Unique study identifier (X0001, X0002, etc.) – this corresponds to the first 5 digits of the X identifier (see above), and enables linkage with the study-level dataset.
AuthorDate	Factor variable, corresponding to the name of the first author and publication year of the corresponding article.

- Site area, study year and information related to PC (MDA) start in country

country	Categorical variable (reminder): country of study arm.
studyYear	Year of efficacy study: corresponds to the study *start* year (i.e. year of drug delivery). Record "-1" if unknown (not reported in publication), and "-9" if not applicable (study not eligible for meta-analysis).
MDAyr	Presumed year of first Massive Drug Administration (MDA) campaign in the country of study; using WHO PCT databank (and additional information from WHO Neglected Tropical Diseases department). Remark: When using WHO PCT databank, record the first year at which the national coverage reaches more than 1%.
siteName	Free text: Most specific information describing the study site, as could be found in published article.
siteClass	Factor variable, corresponding to the name of the area covering the newly defined class – takes the form *Name* [ADM_i], where ADM_i indicates the administrative level at which the class is defined and that provided the class name. If the class name makes use of a specific study site appellation rather than an administratively defined one, then ADM_i is categorised as "site".
classPrecision	Factor variable, corresponding to the estimated maximum distance between GPS coordinates of 2 sites aggregated together in a given class. Distances between sites are straight-line distances, and can take the value <30 km (strictly less than 30 km), <50 km (strictly between 30 and 50 km), 100+ km (about 100 km, likely less and possibly up to 150), 300+ km (about 300 km, likely less and possibly up to 400). If a class corresponds to one unique study site, then the distance is set to 0.
classComment	Free text: rationale for creating the corresponding site class.

- Site geospatial data – information relevant for mapping and graphical representation of site nearness. For statistical analyses, use site class information.

ADM1, 2, 3, 4	Factor variable – location as matching Global Administrative Unit Layers (GAUL), as described by the FAO GeoNetwork: http://www.fao.org/geonetwork/
latitude, longitude	Numerical variables: geographical positioning system coordinates of site, most often derived from Google Map search of study site name (occasionally provided in articles).
GPSaggreg	Dummy variable (Yes/No): Do assigned GPS coordinates correspond to a central point between several aggregated sites? If Yes, the estimated maximum distance between 2 aggregated sites is specified into brackets (same levels as classPrecision). Remark: To assign GPS coordinates to an ADM, the gCentroid() function from R 'rgeos' package was used.
siteFrom	Free text: Explanations for how the assigned GPS coordinates were obtained, and especially for whether they were derived or reported from the article.
GPSprecis	Factor variable, corresponding to the diameter (in km) of the circle (centred on the assigned latitude/longitude) within which the real site is expected to be found.
bestMathc	Factor variable – when mapping the site, which is the best (i.e. most precise) geospatial information available for that study site? Levels: ADM1, ADM2, ADM3, ADM4, LatLong
siteComment JBH/AMJ	Free text: further comments or assumptions made by the 2 researchers involved in geospatial data extraction from articles for this meta-analysis.

Variable dictionary for schistosomiasis work – Dataset 8
17th Oct. 2017 [version 2.0]

- General information and study/group descriptors

MID	Unique row identifier within this meta-analysis dataset (M001A, M001B, M002A, etc.): one per distinct study group for which an antischistosomal drug efficacy measure was reported – i.e. mainly one new row per arm, per species, and per separately studied geographical location. NB: The numerical part of the identifier (001, 002, 003…) is incremented by one for each new study; and the alphabetical part of the identifier (A, B, C…) is incremented by one for each new study group.
SID	Unique study identifier (X0001, X0002, etc.) – this corresponds to the first 5 digits of the X identifier (see above), and enables linkage with the study-level dataset.
AuthorDate	Factor variable, corresponding to the name of the first author and publication year of the corresponding article.

- Are extracted efficacy estimates from independent study groups?

group BreakDown	Semi-free text, to explain the breaking down of one study cohort into several groups (one row and one value of MEC/CR/ERR per group, i.e. per recruitment site and treatment time). Of note, some specific "sites" (for example and most often, several schools) with one MEC/CR/ERR reported for each were actually all in the same village or nearby area. In such a case, the latter geographical unit is considered as the "site"; and a weighted average (weighted by number of participants recruited at each school) of MEC/CR/ERR is computed. Weighted averages for the group were also computed when several study arms recruited in the same place were assigned PZQ40 and PZQ40 + placebo (here considered as the same treatment as PZQ40 alone), or when studies provided the MEC/CR/ERR for each baseline infection intensity subset rather than within the whole study population

- Spatial, temporal and infected participants context for the corresponding efficacy measure(s) recorded in that row

species	Factor variable: "haematobium" or "mansoni". *Schistosoma* species for which the drug efficacy has been assessed.
timeDays	Time (in days) of follow-up for outcome assessment at which the reported CR and/or ERR was calculated. Reminder: day 0 = administration of first treatment dose. Only data for measures within 14-60 days have been extracted. When several time-points measures were available, the one closest to the 14-21 days window recommended by WHO was retained.
regCat	Drug(s) administered to treatment group for which the CR and/or ERR is reported. Refer to dictionary from scoping review for details (Dataset 5).

treatedPpts	Number of participants enrolled in the corresponding study arm and having received the assigned treatment. Record "-9" if unknown/unclear.
analysedPpts	Number of participants included in the calculation of the reported CR and/or ERR (mainly, followed-up at the relevant time-point). Record "-9" if unknown/unclear.
FUposPpts	Derived variable – Number of participants egg-positive (uncured) at follow-up, obtained from: analysedPpts * (100 – CR) / 100

- Efficacy measures and way they were computed

CR	Reported cure rate within corresponding study group.
reportERR	Reported egg reduction rate within corresponding study group. Record "-9" if unknown (not reported in publication). If several ERR types are reported, record *geometric* one (most frequently reported measure across all studies).
repERRtype	Factor variable: "geo", "ari", "unclear" – corresponding to the type of ERR retained in the analysis (with occasionally "geo (both)" to remind the analyst of the availability of an arithmetic ERR as well). Record "N/A" if not applicable (ERR not reported).
intensityCat Baseline	Semi-free text, to record whether the group is comprised of participants with light, moderate or heavy infections at baseline; using WHO definitions.
subgrp Outcomes	Dummy variable (Yes/No): Are there efficacy outcome measures available for each category of infection intensity at baseline? If so, fill in the "Sub-group_intensity" spreadsheet; otherwise record "none".
meanPre	Mean egg count (MEC) in the population before treatment. Record "-9" if unknown.
meanPost	Mean egg count (MEC) in the population after treatment. Record "-9" if unknown.
corrMEClog1	Corrected MEC, assuming the "correct" (standard) GM calculation should be $GM = \text{antilog}(\log(X_i + 1))$, i.e. considering all participants in final ERR calculation.
meanType	Semi-free text, to record the type of population mean egg count reported (AM for arithmetic, GM for geometric, or a mix of both). Record "N/A" if not applicable (ERR/MEC not reported).
derivERR	Derived variable: $(1 - \text{meanPre}/\text{meanPost}) * 100$. Record "-9" if reported data does not allow the derivation of any ERR for that study group.
corrERR1	Corrected ERR, derived from corresponding corrected MEC: $(GM_{pre} - GM_{post}) / GM_{pre} * 100$, with GM_{pre} = corrMEClog1
Correction1	Correction applied to derive a corrected MEC from the post-treatment MEC reported in the publication.
ERRtype	Factor variable: "geo", "ari", "unclear" – corresponding to the type of ERR retained in the analysis (with occasionally "geo (both)" to remind the analyst of the availability of an arithmetic ERR as well). Record "N/A" if not applicable (ERR not reported, nor possible to derive).

MECformula1, 2	Semi-free text, describing the formula used by authors to derive the population Mean Egg Count (MEC) from the individual ones (x). For easiness of read, the calculation is expressed in 2 subsequent stages. Record "(no MEC)" if no MEC was calculated.
consideredPpts	Semi-free text, describing whether the whole study population or only participants uncured at follow-up were considered in the calculation performed to obtain the reported ERR. Record "N/A" if not applicable (ERR not reported, nor possible to derive).

- Characteristics of participants considered to obtain efficacy estimate

ageMin, ageMax	Numerical variable: minimum/maximum age (in years) of analysed (or enrolled, if detailed age range among analysed participants is unknown) participants. Record "-9" if unknown.
ageMean	Numerical variable: mean age (in years) of analysed participants. Record "-9" if unknown.
treatNaive	Dummy variable 1 (Yes) / 0 (No): Were the participants in the study arm naïve to the study treatment (praziquantel)? Record "-9" if unknown.
daysNaive	Numerical variable: number of days prior delivery of the study drug during which we know that participants did not receive an antischistosomal treatment. Record "-9" if unknown/not applicable.
sitePrevStudy	Numerical variable: percentage of infected individuals among those screened for inclusion at the study site. Record "-9" if unknown.
#issues	Free text, to record any additional issues regarding data extraction for this study.

A.4 Quality assurance

After data extraction and coding as per the dictionaries above, I verified all datasets collated for the quantitative synthesis. To that end, I re-examined each article to ensure I accurately recorded data for the most 'impactful' variables. These corresponded to variables capturing numbers of participants (*Schistosoma*-positive at baseline, and who successfully received the assigned treatment) at both enrolment and follow-up, and in each arm; because these impacted on the estimated total of IPD. I also double-checked the recorded *Schistosoma* species and study drug(s), as this information served to filter articles by category (*S. haematobium* or *S. mansoni*, praziquantel *versus* other drugs, etc.) and to create datasets for sub-analyses. Finally, impactful variables also included the participants' follow-up timepoints, as this data influenced inclusion/exclusion of articles from analyses. This double-checking led to minor amendments, with uncovered errors mainly due to ambiguity in the original articles (for example, unclear flowchart of included, treated and followed-up participants). I also randomly selected a subset of articles for full, independent re-examination. This step unravelled 3 typos, which, spread across a total of 45 variables (excluding the impactful variables previously double-checked) and 12 articles, suggest an error rate of 0.56% in the data as a whole, and of $< 1\%$ per variable.

For complementary information extracted in the second phase (meta-analysis), no systematic quality control procedure was performed; however, two researchers[1] were involved and discussed unclear data whenever necessary.

[1] Amélie M. Julé and Julia B. Halder (Imperial College London)

Appendix B

Supplementary methods and analyses of malaria data shared with WWARN

This Appendix complements Chapter 7, where I presented key findings from the quantitative analysis of data management practices in the malaria community, based on a systematic examination of datasets and documentation shared with WWARN. At the onset of this piece of quantitative work, I anticipated a larger sample of eligible antimalarial trials, coming form a wider set of 'independent' data contributors and organisations. In addition, I expected to access more study and data documentation after requesting it from the concerned investigators, and to find the information I needed in more of the available documents than happened in practice. Difficulties in accessing documentation and information led to a substantial decrease in sample

size, leaving 48 to 84 studies with scores on the data management indicators I had defined. Consequently, I abandoned a whole set of statistical analyses I had initially planned, to focus on testing correlations between data management indicators and a few study attributes. The abandoned analyses aimed to test correlations between data management indicators and end-data quality or validity, as shown on Figure B.1 (a former version of Figure 2.2 first presented in Part I).

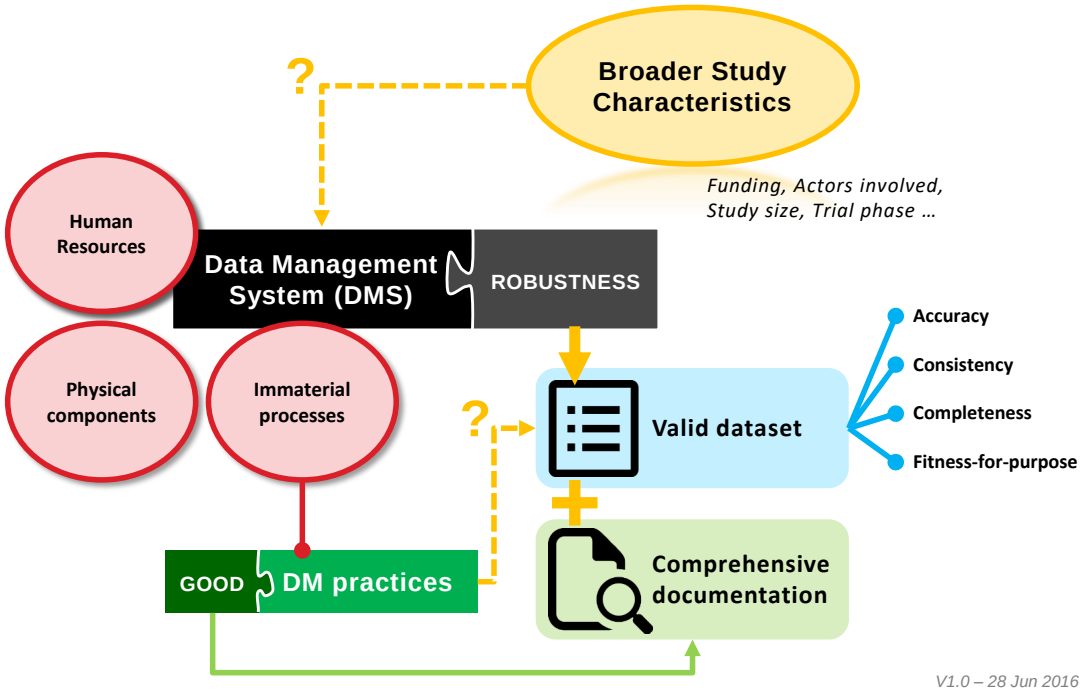


Figure B.1: **Key components and outcomes of a Data Management System (DMS)**, and hypothetical correlations meant to be tested in the quantitative piece (yellow dashed arrows). A DMS encompasses the human resources, physical components (offices and workspace, stationery, hardware and software, etc.), and immaterial processes (governance structure, standard procedures for data management, activities of data handling and manipulation) within an organisation, which altogether contribute to enabling or facilitating data management. The robustness of a DMS is reflected by its ability to deliver a valid dataset for analysis (that is to say, a dataset most accurately reflecting the original, true set of observations intended to be recorded in the field), together with comprehensive meta-data facilitating its understanding and use.

The first part of this Appendix describes how I developed a 'theoretical' analytical framework for the initially forecast work (Appendix B.1), including considerations on data quality that ultimately did not contribute to the analysis. Subsequent parts of this Appendix clarify the communication addressed to WWARN data contributors as well as its outcomes (Appendix B.2), and describe the data I collected to search for correlations between study attributes and data management indicators (Appendix B.3).

B.1 Development of the analytical framework to characterise data management practices

When developing a protocol for the systematic exploration of data from the WWARN repository (second semester of 2016), I could not identify an instrument enabling to quantitatively characterise data management practices based on shared datasets and accompanying study documentation. Van Tuyl *et al.* designed an interesting framework to evaluate sharing practices against desirable outcomes reflecting the purpose of research data sharing, which they define as making data "discoverable", "accessible", "transparent" and "actionable" [296]. Its scope restricted to data sharing makes Van Tuyl and colleagues' framework unlikely to cover wider dimensions of data management, but I adopted a similar methodology to create a suitable data collection instrument for my analysis. I first clarified the desirable outcomes of data management and dimensions of data quality, by drawing on published literature

and experts' knowledge. I then listed observable attributes of clinical trial data that reflect initial data management strategies and final outcomes. The list of attributes was progressively refined through pilot-testing on a sub-sample of antimalarial trials, until it provided a list of relevant variables for subsequent statistical analyses. Several methodologists have recommended such a multi-stage (broad scoping from the existing body of knowledge, then refinement through piloting) and mixed-methods (literature review and discussions with experts) approach for the development of data collection instruments in quantitative research [297, 298].

B.1.1 Elements drawn from the literature

Test searches on PubMed and Clarivate Analytics (Web of Science) highlighted the difficulty of constructing meaningful (specific and sensitive) queries of journals' databases around concepts as generic as 'data quality' and 'data management' in relation to clinical research. Consequently, and consistently with the approach taken for the literature review reported in Chapter 1, I performed a traditional review, based on evolving search criteria and non-exhaustive examination of the references retrieved. As an example, I used the following keywords: ("data management" OR "data quality") AND (guideline* OR practice* OR strateg*), with or without "clinical". I sorted search results by relevance whenever the browsed database offered this functionality, and I examined the first pages of reference titles to assess their interest for my work. I also considered perspectives from the grey literature, including reports from stakeholders actively promoting data management and sharing.

This literature scoping highlighted four essential dimensions of end-data that reflect the robustness of initial data management tools and processes:

- **Availability** – accessibility of the study dataset, together with comprehensive supporting documentation enabling its use at any time (extends the concept of data timeliness [299–301] and transparency [296]);
- **Accuracy** – correctness of the study data. Accurate data is representative of the reality it intends to reflect, or in other words, the error between a recorded value and a true value is null or negligible (depending on the potentiality of the data capture instrument used) [299–301];
- **Completeness** – exhaustiveness of a dataset. A complete trial dataset covers all enrolled participants and all collected variables (as described in the trial protocol). A complete dataset also clearly states the reasons for missing (blank) values, which should be minimised [300, 301];
- **Reliability** – validity of the study data and findings, as ascertained by their reproducibility. A reliable dataset can be reproduced through repetition of the experimental trial, and yield similar conclusions upon analysis [302, 303].

The literature mentions other dimensions reflecting the robustness of a DMS and quality of delivered data, but these were either irrelevant to the case of clinical research, or variations around the broader themes covered by some of the dimensions I retained. For example, the "fitness-for-purpose" [299] of a dataset is an important attribute, but not relevant to assess in the WWARN case examined here. Indeed, my

analysis of data management practices in the malaria community draws on clinical trial data initially collated for the purpose of IPD meta-analyses; therefore, all datasets will be fit for this purpose. Another example of synonymous or embedded dimension is "data consistency" (defined as the coherence between data recorded across variables, for example between gender/age and pregnancy status) [304], as this is a specific demonstration of data accuracy.

B.1.2 Input from experts

To balance the difficulty in systematically capturing information from the literature, a colleague[1] and I organised a brainstorming session. We invited colleagues from the Centre for Tropical Medicine and Global Health and other clinical research units within the University of Oxford, some of whom had experience working in LMICs. This informal, group conversation aimed to draw on participants' expertise of data management, in order to identify elements potentially missed during the literature review, as well as to provide a more practical standpoint on the matter. The group of data managers, computer scientists and statisticians was encouraged to think about the data lifecycle, using as a prompt the representation proposed by ADMIT (reproduced in Figure 2.1) [129]. We particularly encouraged participants to describe how the trial end-data may reflect each planning and execution step of the data management process. Their 'hands-on' perspective was critical to identify concrete data attributes that I could objectively assess from one clinical trial to the next, and using the often-limited information accessible regarding shared study

[1] Laura Merson, Infectious Diseases Data Observatory (IDDO)

data. Tangible evidence (indicators) of robust DMS suggested by experts included the use of data management software enabling audit-trail, as well as the existence of detailed DMPs and reports specifying the data verifications performed.

B.1.3 Pilot application and instrument refinement

The preliminary list (Tables B.1 and B.2) linking theoretical dimensions to practical indicators of 'good' data management practices formed the basis for further specification of the quantitative (and semi-quantitative) variables underlying statistical analyses. The resulting, draft collection tool was confronted to the data sources I could use to assess data management strategies and outcomes. These sources included published articles, study protocols or reports, CRFs, variables' dictionaries, dataset annotations, as well as notes and reports from WWARN data curators. I performed iterative phases of piloting on a random sample of 21 studies (about 5% of the over 500 WWARN entries initially screened for eligibility), to confirm the practicality and usefulness of the variables when seeking to characterise and highlight variability in data management practices across antimalarial trials.

The pilot phases demonstrated the impossibility to collect some desirable information for the vast majority of studies, whether due to its sheer absence or to the impossibility to measure it in an objective manner. For example, accessible study documentation rarely covered the data verifications performed or the data standards applied; and the "*neatness*" of documentation or "*smoothness*" of data distribution were too subjective for consistent assessment. Finally, some of the suggested indi-

Output	Desirable attributes of shared data
Overall study documentation and meta-data	Availability → "*the more, the better*" Documents should be "*neat*", i.e. well organised and easily readable, self-explanatory and understandable, complete/exhaustive, detailed. They should also demonstrate "*continuity*" between protocol, CRF contents and database structure.
Study protocol	From the data manager's perspective, should at least cover: • inclusion/exclusion criteria – guidance on the definition of protocol violations and on expected variables' range • diagnostic method – with details on validation and standardization of equipment/procedures across labs
Annotated CRF or data collection form + Data dictionary*	Variables' names and labels should demonstrate some "*logic*": • e.g. 'Temperature at day N' consistently labelled TempD01, TempD28 (vs. T-day1, Temp-28 mixed in the same dataset) In addition, variables' dictionaries should cover: • the data type – numerical, textual, categorical… • possible values (esp. for categorical variables) and corresponding codes, incl. distinct codes for 'missing' and 'not applicable' • value range [seen as an optional feature]
Data Management Plans (DMPs) and reports	DMP should specify: • data standards used (e.g. ISO, CDISC) • data validation strategy – total/partial verification of collected data, coherence and consistency checks, etc. Existence of data management reports, study queries report and other data manipulation/audit-trail/QC report suggests well thought through and anticipated data modifications

Table B.1: **Desirable attributes of outputs from a clinical research Data Management System (DMS) – part 1**. Elements italicised and inter-quotes are paraphrased from experts. *Data dictionaries may also be referred to as codebooks or variables' dictionaries. CRF: Case Record Form; ISO: International Organisation for Standardisation; CDISC: Clinical Data Interchange Standards Consortium; QC: Quality Control

cators were remarkably homogeneous across the studies selected for piloting. This included the specification of inclusion/exclusion criteria and diagnostic methods, which are virtually always described in protocols. I dropped all those variables

Output	Desirable attributes of shared data
Dataset (data table(s))	Dataset should demonstrate accuracy → have been checked for variables' range and "*smoothness*" of distribution (no spikes or separation of distribution suggestive of inappropriate rounding/digit preferences) + plausible data combination per participant (e.g. pregnancy and female) Dataset should demonstrate completeness → document reasons for "*missingness*" – different scenarios possibly explaining it (e.g. variable added as the study occurred, participant missing a visit, test failure at laboratory…) In (quantitative) clinical research studies, dataset should contain limited amount of overall text → excessive text can suggest investigators gave little attention to breaking down the information into simple variables, i.e. to what should be collected and why.
Database system	The chosen system should allow: automatic audit-trail, built-in edit-checks, queries' generation/resolution, thus facilitating achieving the desirable attributes of a dataset Database robustness evidenced in the end-data by: • consistent formatting, i.e. constant format per variable – especially in multiple-site studies → date and measurement units conversion • constant data type per variable (not mixing numbers/text) • absence of typos (no "s*loppy*" entry), blanks and duplicates • structure of disseminated data – file format + flat/relational tables (indicative of source software and database structure)

Table B.2: **Desirable attributes of outputs from a clinical research Data Management System (DMS) – part 2**. Elements italicised and inter-quotes are paraphrased from experts.

with an excess of missing or homogeneous data from the analytical framework, as they would not contribute to a quantitative characterisation and contrasting of data management strategies and outcomes across the sample of clinical studies.

Eventually, the retained variables were operationalised into an electronic survey using LimeSurvey [305], to facilitate systematic data extraction.

B.2 Communications with data contributors

To limit bias in documentation availability, which the initial theoretical framework considered as a key dimension of DMS robustness that I sought to 'measure', I contacted at least one data contributor per eligible trial and requested missing documents (if applicable). Requests were addressed by e-mail to the assigned WWARN contact point for the trial, and another collaborator was put in copy whenever possible, to increase chances of documents' retrieval. An example of the initial communication is reproduced (in an anonymised version) on the next page. This letter was followed by a reminder after about a week.

Overall, I invited over 100 distinct contributors to provide additional study documentation regarding 251 studies. The response rate was similar among investigators having conducted studies in Africa and Asia, with 60.3% and 64.1% of e-mails answered to, respectively for each continent. Most responses were to confirm agreement for inclusion in the analysis, but without providing the requested documents, nor clarifying reasons for their missingness. Before the round of communications, the ratio of available over inaccessible protocol or report was 0.8 (47 already provided *versus* 61 missing documents) and 0.7 (22 *vs.* 33) for studies performed in Africa and Asia, respectively. The documents newly shared by contributors on this occasion inverted the picture and increased those ratios to 1.4 in both cases (63 provided *vs.* 45 still missing, and 32 *vs.* 23, respectively in Africa and Asia).

16th January 2017

Dear Dr. *< Name >*,

We are writing to you because some study data you have shared with WWARN are eligible for inclusion into our study of **Data Management Systems (DMS) across Malaria Clinical Trials**. However, additional information (as listed in the table below) would greatly help us to perform this work.

To enable this **DMS study**, as well as the improved curation of your dataset going forward, we would like to review the protocol and data dictionary, and any other documentation related to the data management procedures for your studies.

Please click here (or go to https://survey.msd.ox.ac.uk/index.php/491165?lang=en) **to upload the missing data if available.** Feel free to also include any additional information regarding your study processes, such as data management plans.

SID	Title	PubMed ID	Protocol	Dictionary
ID-1	*< Study Title 1, as recorded in WWAN repository >*	PMID-1	**Missing**	Partial – please provide complete if available
ID-2	*< Study Title 2, as recorded in WWAN repository >*	PMID-1	**Missing**	**Missing**
ID-3	*< Study Title 3, as recorded in WWAN repository >*	PMID-1	Already provided – please share additional information (e.g. Data Management Plans) if possible	Already provided

Through this work, we want to better support you and other malaria researchers in future data management and sharing activities. We plan to provide all data contributors to WWARN with feedback on our findings by September 2017, and use these findings to work towards the development of specific data management resources. We later plan to publish a report in an open-access journal, where your contribution will be acknowledged.

We may have to contact you for additional details, but we will keep this contact minimal. We thank you in advance for your help with this important project.

To opt out of the DMS study, please reply to this e-mail by Friday 3rd of February 2017.

Find out more about the study: http://www.wwarn.org/data-management-systems-study

Amélie

WWARN Data Management Systems (DMS) study | Contact: **data@wwarn.org**
Infectious Diseases Data Observatory (IDDO) | WorldWide Antimalarial Resistance Network (WWARN)
www.wwarn.org | www.iddo.org | www.facebook.com/AntimalarialResistance | twitter.com/WWARN

B.3 Data preparation for statistical testing (correlation analyses)

Information extracted from datasets and documentation for the 108 antimalarial efficacy trials relate to three topics: (i) study, (ii) contributors & actors, (iii) data management. In this section of the appendix, I present all variables I collected in relation to the first two topics. I jointly describe the manipulations performed to prepare these variables before testing my hypotheses; that is to say, before searching for potential associations with data management strategies and outcomes. The Appendix does not mention data management variables, as these were extensively described in the main part (Chapter 7).

B.3.1 Study 'complexity' – trial design

Arms, randomisation, blinding

	Open-label	Blinded
single-arm	19*	-
2 arms	43 (3*)	10
3 arms or more	17 (3*)	13

Table B.3: **Number of arms and blinding level**. Values are numbers of studies with the corresponding design. An asterisk (*) indicate non-randomised studies, counted as open-label although the blinding-level attribute does not apply *sensu stricto*.

To measure the 'complexity' of study design, I collected data on the number of treatment arms, the use of randomisation when assigning participants to treatments,

and the use of blinding when administering or assessing treatment outcomes. In fact, the design of the 108 trials turned out quite homogeneous (Table B.3). Notably, 89 were multi-arm; 83 were also randomised (93.3% of multi-arm studies), of which 60 were open-label (72.3% of randomised studies). In view of this relative homogeneity, I only retained the number of treatment arms to account for the study design and associated 'complexity'.

Safety, pharmacology (pharmacokinetics or pharmacodynamics, PK/PD) and molecular data

In addition to clinical efficacy data, seventy-one studies (65.7%) were designed to collect clinical safety data; and forty-six studies (42.6%) collected pharmacokinetics/pharmacodynamics (PK/PD) and/or molecular data about parasites' resistance markers (Table B.4). The existence of such 'laboratory' data correlated with the number of study arms ($\chi^2 = 6.45$, p-value = 0.040), hence only the binary variable reflecting the existence of safety data was kept for analyses.

	No safety data	Safety data
No PK/PD and/or molecular data	24	38
PK/PD and/or molecular data	13	33

Table B.4: **Variety of data collected in addition to clinical efficacy data**. Values are numbers of studies with the corresponding data type(s). A data type was considered as existing if it was observable in the datasets shared with WWARN for the study; but also if it was mentioned in articles or supporting study documentation, despite its absence from the WWARN database.

B.3.2 Study 'complexity' – trial size

Number of enrolled participants

The number of enrolled participants per study varied from 13 to 6,026 (with a mean value of 538.1 participants). The cohort size distribution was extremely skewed, as 98 studies (90.7%) enrolled less than 1,000 participants, thus the log-transformed representation on Figure B.2.

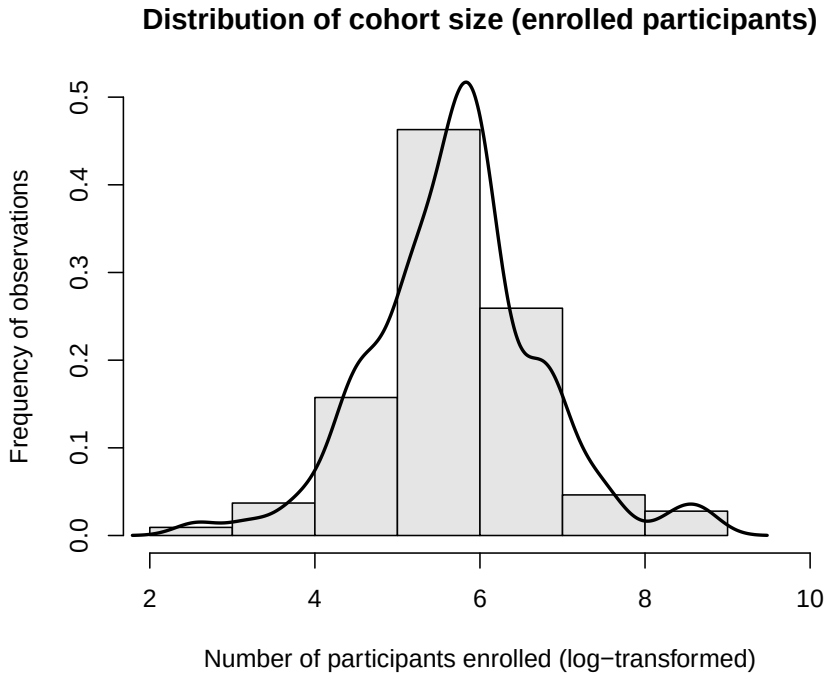


Figure B.2: **Distribution of cohort size, using the log-transformed count of participants**. The number of enrolled participants reported in published articles or reports sometimes differed from the number of participants found in the dataset shared with WWARN. If so, the latter number was retained, to allow meaningful comparisons with other metrics extracted, such as the database size or data 'quality' indicators within WWARN data management reports (metrics eventually not considered).

Total of distinct recruitment countries and sites

	single site	2 to 4 sites	5 sites or more
single country	57	32	4
2 to 4 countries	-	6	5
5 countries or more	-	-	4

Table B.5: **Amount of recruitment countries and sites**. Values are numbers of studies with the corresponding amount of sites and countries involved in recruiting participants.

	Single-site	Multiple-site
DMP score	(N = 48), p-value = 0.695	
Low	12.5% (6)	8.3% (4)
Medium	18.8% (9)	27.1% (13)
High	14.6% (7)	18.8% (9)
Dictionary score	(N = 52), p-value = 0.937	
Low	9.6% (5)	9.6% (5)
Medium	17.3% (9)	21.2% (11)
High	21.2% (11)	21.2% (11)
Software (DM)	(N = 84), p-value = 0.702	
MS Excel®	9.5% (8)	8.3% (7)
MS Access® or Epi Info™	28.6% (24)	23.8% (20)
other	13.1% (11)	16.7% (14)

Table B.6: **Observations of DMP, dictionary and software scores by recruitment site context**. P-values are from Fisher's exact tests for the corresponding sub-table.

Few (15, or 13.9%) studies were carried out across multiple countries, and even less (13, or 12%) recruited participants at five or more sites (Table B.5). I thus aggregated the number of recruitment countries and sites into one binary variable

distinguishing between single- and multiple-site studies, but data management indicators did not depend on the number of study sites defined in that way (Table B.6).

B.3.3 Study setting and sources of support

Recruitment country (macro-region)

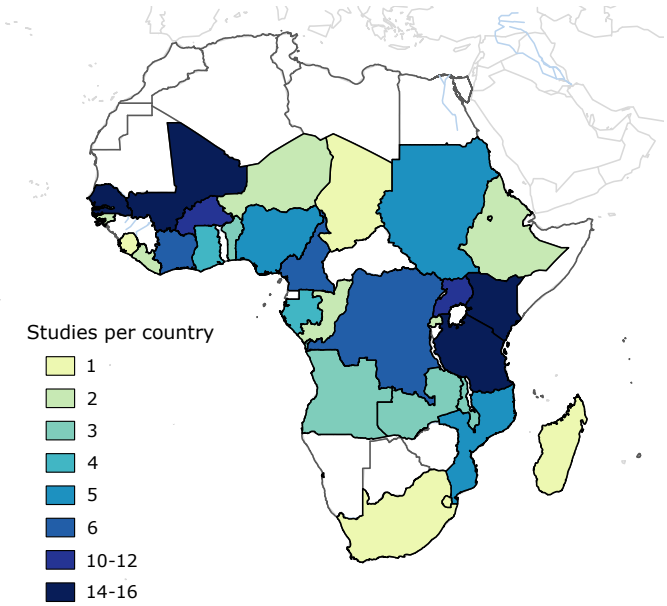


Figure B.3: **Included trials per African country**. This map is not fully representative of the overall WWARN data available for Africa, owing to the exclusion/inclusion criteria applied in this quantitative analysis (see methods in Chapter 7).

The included 108 studies covered 30 distinct countries (Figure B.3), of which six countries have contributed 10 or more datasets to WWARN: Senegal and Mali (10 trials each), Tanzania (12), Kenya (14), Uganda (15) and Burkina Faso (16). To compare data management strategies between countries while limiting the number of different classes for the 'study place' variable, I grouped countries by regional area: Western (49 studies), Eastern (46) and Middle (22) Africa (as categorised by

the United Nations [232]), and others (as only six studies altogether were performed in Southern or Northern Africa). These regional categories were non-exclusive, as 12 multi-country studies involved countries across different macro-regions.

Urbanisation setting at recruitment site(s)

	Western Africa	Eastern Africa	Middle Africa	other/multiple macro-regions
site(s) in rural area only	17	23	4	6
site(s) in urban area only	13	10	10	2
sites in mixed rural and urban areas	8	2	2	8

Table B.7: **Urbanisation setting by macro-region**. Fisher's exact test suggests a statistically significant correlation between the two categories (p-value = 1.704×10^{-3}).

Data on the level of urbanisation was captured, to test the hypothesis that rural settings make it more difficult to implement complex DMS, for example relying on modern ICT infrastructure. However, the urbanisation level variable was not exploited, because it is somewhat subjective (depending on the overall development of the country), and it correlated with the macro-regions defined above (Table B.7).

Source of support (funding)

Sources of funding varied and were categorised as: (i) public, for studies funded by a governmental body, such as a Health Ministry or National Research Council; (ii) non-profit, for studies funded by a charity or private foundation not seeking to generate profit; or (iii) private, for studies funded by the pharmaceutical industry.

	Western Africa	Middle Africa	Eastern Africa	other/ multiple	Total
Public source of funding	**26**	**9**	**39**	**5**	**81**
WHO, WHO-TDR or WHO-AFRO	3	2	4	3	12
CDC	1	1	5	0	7
EDCTP	3	1	2	1	7
SAREC/SIDA	2	0	5	0	7
DFID	0	0	6	0	6
European Commission	1	1	3	0	5
National Institutes of Health (USA)	2	0	3	0	5
President's Malaria Initiative/ American Relief & Recovery Act	3	1	1	0	5
The Global Fund	3	0	0	0	3
International Atomic Energy Agency	2	0	0	1	3
National Research Council*	2	2	3	0	7
Ministry of Foreign Affairs/ International Development**	2	0	3	0	5
Ministry of Health/MNCP/Hospital*	4	1	0	0	5
Other (3 academic institutes)	0	0	4	0	4
Private source of funding	**9**	**2**	**1**	**8**	**20**
Sanofi	4	0	1	2	7
Mepha	3	1	0	0	4
Other (6 pharmaceutical companies)	2	1	0	6	9
Non-profit source of funding	**17**	**7**	**11**	**6**	**41**
Médecins Sans Frontières (MSF)	4	7	2	2	15
Bill and Melinda Gates Foundation	4	0	1	1	6
DNDi	2	0	1	1	4
MMV	0	0	1	2	3
Wellcome Trust	1	0	2	0	3
Other (10 charities or foundations)	6	0	4	0	10

Table B.8: **Sources of funding, by regional area where studies took place**. Observations of the corresponding funder in the sample of 108 studies. Totals add up to more than 108, as studies may receive funding from multiple sources. CDC: Centers for Disease Control and Prevention (USA); DFID: Department for International Development (UK); DNDi: Drugs for Neglected Diseases initiative; EDCTP: European and Developing Countries Clinical Trials Partnership; MMV: Medicines for Malaria Venture; MNCP: Malaria National Control Programme; PMI: President's Malaria Initiative (USA); SAREC/SIDA: Swedish International Development Agency. (*) any country; (**) any other than UK (DFID) and Sweden (SAREC/SIDA).

As a reminder, the initial intention was to also capture information about the sponsor (considered as the legally responsible party, whether it is also the funder or not), but piloting the data collection instrument demonstrated that sponsorship was too rarely known to be of interest for quantitative analyses.

Most studies performed in Eastern Africa had at least one public source of funding: 76.5% of the stated funders in this macro-region were indeed public funders, compared to 50% in both the Western and Middle Africa regions (Table B.8). In light of this association, I did not further explore variations in data management indicators by macro-region (country) of study conduct.

B.3.4 Interconnectedness of investigators in the malaria community (network analysis)

When coming up with a communication plan, which required identifying one primary contact point and one collaborator to put in copy of e-mails, while ensuring no data contributor would receive more than one request, I noticed the interconnectedness (joint work on trials) of malaria investigators involved with WWARN. This interconnectedness was further evidenced when contacted investigators forwarded my e-mails to each other. These observations prompted me to perform a network analysis of the malaria community represented in my study sample, considering that 'connected' researchers may also exchange their data management tools and procedures.

Definition of nodes, edges and attributes

For each trial identified and included in the exploration of data management practices, I extracted information regarding: (i) the authors (first, last and corresponding authors) on the published article, if applicable; (ii) the WWARN contact points (data contributors); (iii) the principal investigator(s) (for academic studies) or study coordinator (for pharmaceutical studies) as specified on the protocol, if available. Naturally, many contributors simultaneously belonged to multiple categories for one same trial.

I then compiled a list of unique contributor identifiers (network's nodes), with associated attributes – namely, institute of affiliation and country of institute. For contributors with multiple affiliations occurring across the sample, a PubMed search and examination of several articles served to clarify the most recent and/or most frequently used affiliation. Two different contributors appearing as authors, contact points, and/or investigators on the same trial are connected by this trial, which therefore constitutes an edge of the network. The lists of nodes and edges thus compiled were imported into Gephi [306] for the graphical representation and calculation of metrics for the network (see [307] for an accessible introduction to metrics used in network analysis and discussed below).

Network representation and characteristics

As specified above, I contacted about 100 contributors. When also considering first, last and corresponding authors of publications, as well as stated investigators on

protocols, I identified 180 unique contributors for the sample of 108 African studies. They formed a network connected by 312 edges (distinct pairs of contributors to at least one same study – Figure B.4). The average degree of this network is 3.467, suggesting that the considered contributors worked with three to four others on average. The average weighted degree is 4.911, meaning that connected contributors worked together on an average of 2.456 studies within this sample.

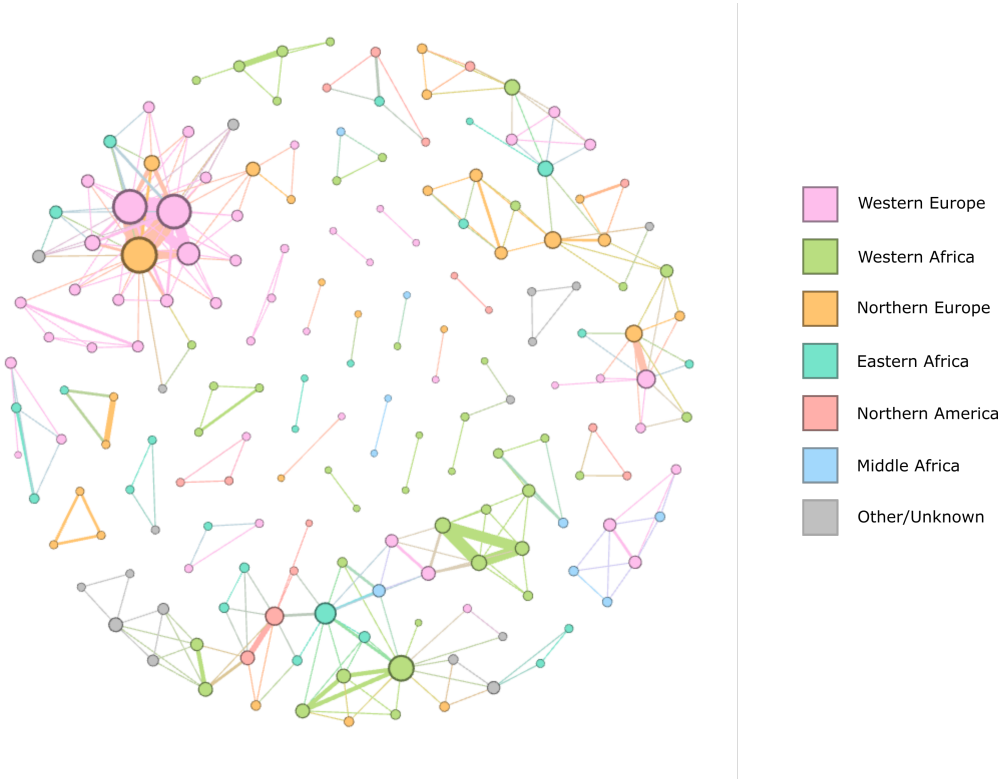


Figure B.4: **Network of authors, investigators and WWARN contact points among African studies of antimalarial efficacy**. Regional groups correspond to the location of the institute where the corresponding contributor (node) is affiliated. Nodes' sizes are proportional to their degree (number of connections). Edges' thickness is proportional to the number of studies that the connected contributors worked on.

The network has a diameter (longest path between two contributors) of 9, and an average path length of 3.044, which suggest a rather high connectivity. Yet, this

network is not very dense (with a density of 0.019), which suggests a few contributors are very well connected while most others have only a couple of connections. Indeed, 31 contributors (17.2% of the network's nodes) have only one connection, and 56 contributors (31.1% of nodes) have exactly two connections, making this value the median degree. By contrast, six contributors have strictly more than 10 connections, of which three have 20 or 21 connections. Four of those six 'influential' contributors are affiliated with MSF/Epicentre, which further highlights the excessive representation of that group within the sample, as already observed among funders (Table B.8). Interestingly, the two next highly connected nodes are also connected with each other, thus introducing an important connection between contributors affiliated with institutes based in Western and Eastern Africa.

Appendix C

Topic guides and intermediate qualitative data analysis

This Appendix covers additional methodological and analytical material produced to support the two qualitative case studies discussed in Part III.

C.1 Methodological tools and procedures

C.1.1 Participants' information brochure

The following information brochure (translated in French by myself) was distributed to study participants, before gathering their consent for an in-depth interview (consent given in writing) and/or for me to take field notes on some personal information they had shared informally (consent given orally).

14th October 2016 – Version 1.1
OxTREC reference: **5120-16**

This Information Sheet and Consent Form has been created based on a template provided by the KEMRI / Wellcome Trust, via The Global Health Network and freely accessible here: https://globalhealthtrials.tghn.org/resources/templates/

Research Participant Information Sheet

Project Title

Data Management Practices across Malaria Clinical Research: exploring variability to understand the needs

Responsible researcher	Amélie M. JULÉ [1,2], amelie.jule@gtc.ox.ac.uk *Candidate DPhil in Clinical Medicine*	
Supervisors	Prof. Trudie A. LANG [1,2] *(primary supervisor)* Prof. Philippe J. GUÉRIN [1,3] Dr. Vicki MARSH [1,4] Prof. Piero L. OLLIARO [1,5]	trudie.lang@ndm.ox.ac.uk philippe.guering@wwarn.org vicki.marsh@ndm.ox.ac.uk olliarop@who.int
Co-investigators	*Senegal* – Prof. Oumar GAYE [6] Dr. Magatte NDIAYE [6] *Republic of the Congo* – Prof. Francine NTOUMI [7]	oumar.gaye@ucad.edu.sn magatte1.ndiaye@ucad.edu.sn francine.ntoumi@uni-tuebingen.de
Affiliations	1. Centre for Global Health and Tropical Medicine, Nuffield Department of Medicine, Oxford University (United Kindgom) 2. The Global Health Network, Oxford University (United Kindgom) 3. WorldWide Antimalarial Resistance Network (WWARN), Oxford University (United Kindgom) 4. KEMRI/Wellcome Trust, Kilifi (Kenya) 5. UNICEF/UNPD/World Bank/WHO Special Programme on Research and Training in Tropical Diseases (TDR), Geneva (Switzerland) 6. Faculté de Médicine, Université Cheik Anta Diop, Dakar (Senegal) 7. Fondation Congolaise pour la Recherche Médicale, Brazzaville (Republic of the Congo)	

What are the institutions involved and what is this research about?

The Global Health Network[1] and *WWARN*[2] are two academic research groups working collaboratively within the *Centre for Tropical Medicine and Global Health*, an institute based at Oxford University (UK). Joint interests of *The Global Health Network* and *WWARN* include research capacity-strengthening and training: both groups are highly committed to the development of open-access resources to facilitate the conduct of clinical research, especially in low-resource settings within Low and Middle Income Countries (LMICs). In order to design the most appropriate tools for researchers based in LMICs, *The Global Health Network* and *WWARN* carry out needs assessment and

[1] https://tghn.org/
[2] http://www.wwarn.org/

exploratory research with local partners. Those mainly involve asking questions to research staff, and in some cases observing activities, so that opportunities and concerns are better understood.

All research at the *Centre for Tropical Medicine and Global Health* involving human subjects has to be approved by the *Oxford Tropical Research Ethics Committee* (OxTREC) before it begins. This university-based committee makes sure that every research project is important, and that participants' safety and rights are respected.

In this specific project, we want to learn more about clinical research data management practices at your institution. Amélie Julé is leading the research, and would like to hold discussions with participants representing a variety of research roles, including: data managers, statisticians, computer scientists, data entry personnel and investigators. She would like to talk to people individually/in groups in a convenient meeting room of your research institute, and spend some time observing usual activities of a few participants in their day-to-day job.

Why do you want to talk to me and what does it involve?

Your research institution has been selected because it is based in a LMIC; and an investigator *[insert name]* based here is currently collaborating with WWARN, and they expressed an interest in being included in the present data management needs assessment study. Your colleague suggested that we talk to you, because you play an important role in relation to the generation, management or quality-control of clinical research data at your institute. We feel that your experience of working with malaria study data can contribute much to our understanding and knowledge of data management practices in *[specify country]*.

[Select paragraph below that applies]

*For **in-depth interviews*** – I would like to ask you a number of questions about your experience of working with malaria study data. If you do not want to answer any of the questions you may say so and I will move on to the next question. The discussion will take place in a meeting room of your research institute. No-one else but the interviewer needs to be present, unless you would like someone else there with you. The interview will be tape-recorded, and a written transcript will be created to assist in later analysis. No-one will be identified by name in the transcript, and the tape will be destroyed at the end of the project.

*For **focus group discussions*** – We would like you to take part in a discussion with 5-9 other persons with similar experiences. The discussion will be guided by a trained facilitator. We will ask questions about *[summarise topics detailed in protocol, based on corresponding staff]*. You do not need to discuss any information you are not comfortable in sharing. The discussion will take place in *[specify location/setting]*. Only the people involved in the discussion, the person asking the questions, and a note-taker will be present. The discussion will be tape-recorded, and a written transcript will be created to assist in later analysis. No-one will be identified by name in the transcript, and the tape will be destroyed at the end of the project.

Are there any risks or disadvantages to me of taking part?

[Select sentence below that applies]
*For **in-depth interviews*** – The interview should take about 1 hour or less.
*For **focus group discussions*** – The group discussion should not exceed 2 hours.

We do not anticipate major inconvenience for you other than the time you will devote to participating.

Some individuals may feel uncomfortable to describe their work activities if they are aware of not following the recommended best practices. We will not make any judgement, and on the contrary, we would very much like to identify what prevents you from following the best practices that you know.

Are there any benefits to me of taking part?

There are no individual benefits to taking part, but in answering our questions you will help us improve our understanding of clinical research data management practices, which could help address the needs and strengthen research capacity of your colleagues in the future. You may also get new insights on your own work and acquire greater confidence to perform it in the future.

Who will have access to the information I give?

During the project duration, we will not share individual information about you or other participants with anyone beyond a few people who are closely concerned with the research. It is very unlikely that such people will not have affiliation with one of the institution listed above.

All of our documents and tapes are stored securely in locked cabinets and on password-protected computers. The tapes will be destroyed after completion of the research.

The knowledge gained from this research will be shared in summary form, without revealing individuals' identities. Your research institute will receive a report of the findings; which we then aim to publish in an open-access scientific journal.

The anonymised data from this research will be archived for at least five years and may be used to support new research on similar questions by other researchers. Any future research using information from this study must first be approved by a local or national ethics committee to make sure that the interests of participants and their communities are protected.

*For **focus group discussions** only* – We will ask everybody in the discussion to keep what is said in the group confidential, but it is important to recognize that we cannot stop participants sharing what they have heard.

What will happen if I refuse to participate?

All participation in research is voluntary. You are free to decide if you want to take part or not. If you do agree you can change your mind at any time without any consequences.

[Select paragraph below that applies]

*For **In-depth interviews*** – If you decide to withdraw over the course of the interview, we will stop it there and all the recordings and notes from our encounter will be destroyed.

*For **Focus group discussions*** – If you decide to withdraw over the course of the discussion, you can leave the room at any time (please note that the group will carry on). We will seek to delete your previous contributions from the transcript of the conversation, although we may be in touch to ask for your agreement to keep some fragments that are necessary to the understanding of the general discussion.

What if I have any questions?

You are free to ask me any question about this research. It is important that your interview should be about your experience, and I would not want to influence your responses. In case you think about a question during the interview, I may have to ask you to wait until the end of the interview before I provide you with a detailed answer.

You can ask me any question now before we start, and if you have any further questions about the study after our discussion, you are free to contact the research team using the contacts below:

Amélie JULÉ

Centre for Tropical Medicine and Global health, Nuffield Department of Medicine Research Building, University of Oxford, Old Road Campus, Roosevelt Drive, Oxford, OX3 7FZ, UK

Telephone: *[insert mobile]* or *[insert landline]*. E-mail: amelie@theglobalhealthnetwork.org

C.1.2 Interview topic guides

During qualitative interviews, I exchanged with participants on the topics listed below. Questions correspond to those I jotted down in my initial guide, but they partly evolved as interviews and case studies unfolded. In addition, I soon became very familiar with this tool, which enabled more interactive conversations and adaptive phrasing depending on the participant's experience. Bullet points below each question indicate possible prompts or clarifications, which I used to invite participants to share further details, if needed and as relevant.

Topic 0 (warm-up): Work experience and motivations

How did you come to work as a < job role > in this institute?

- Background and initial training
- Past work experience – at other institute(s), in another research/health role

What prompted you to pursue this career?

- Motivations: interest for the work itself, appealing work environment (flexible working hours), interesting opportunities for career progression/training
- Necessarily (or randomly) – job availability, salary, stable employment

What do you like most about this job?

- Could you tell me more? [Why?]

Topic 1: Clinical research environment

What research projects typically happen at your institute?

- Diseases under study?
- Type of study (clinical, pre-clinical, epidemiological), trial phase...

Who initiates these projects?

- Pharmaceutical industry, governmental bodies, university/academic institutes
- Local investigators – external/internal to the visited institute

Topic 2: Breadth of data management – define and describe

'Data management' is a generic term. Different persons think about different things when they hear this term. Could you tell me what it means to you?

- Think about the data you create/manipulate/analyse yourself. What does it involve 'to manage' this data?
- Type of data to handle: structured/tabulated or more qualitative; meta-data and associated documentation...
- Note: This question first lets participants express their ideas in an unguided manner.

Topic 3: Concrete activities and processes (+ Data quality)

How is data managed at your institute?

- Centralised management; or *ad hoc*, on a case-by-case basis for each study
- Tools and software used + Involved human resources
- Type of information collected (where? at the clinic, in the lab, in the field) and transferred (when? to whom?)

As a < job role >, how do you contribute to data management?

- Refer to steps previously cited by the participant (and to steps of the research data lifecycle, if not cited): You mentioned X or Y as an important step of data management. Can you tell me more about your involvement in that step?
- Can you tell me how you facilitate these activities [For participants in a study oversight role only]? For example, training opportunities provided to team members, HR management, introduction of new technology...

How do you prepare research data management plans? [OR: How are data management (and statistical analysis) plans usually developed?]

- Re-using templates from past studies

- Using guidelines developed at the institutional level
- Other templates/guidelines: SCDM, The Global Health Network...
- What should they cover (according to you)? What do they cover (in practice)? For example, roles and responsibilities, archiving and sharing...

What do you think about the preparation/planning of data management?

- Is it important and useful, or is it pointless paperwork?

What do you do to ensure the quality of study data?

- Processes: checklists, routine controls (graphical representation, sum-checks...)
- Personnel (who does what in terms of QC): audits, independent monitor/dedicated team member, investigators/authors on publications
- Data that are checked – socio-demographic data, clinical data (primary outcome, safety data)

Topic 4: Underlying reasons for data management choices

Why is data managed in this way?

- Institutional policy
- Practical reasons *vs.* constraints (for example, available resources)

What makes data management procedures differ from one study to the next?

- Disease under study?
- Type of study (clinical, pre-clinical, epidemiological), trial phase...
- Funding or sponsor – especially if pharmaceutical industry or PI from a developed country
- Did you notice some change in data management practices in recent years?

Topic 5: Interactions between roles

Other than yourself, who else in involved in data management? [In what way?]

- Medical doctors/Nurses, data entry clerks, statisticians...

Are these always the same persons? [Why or why not?]

- Does it matter? For example, for the quality of resulting data

Beyond data management (processing), in which other aspects of clinical studies are you involved? [In what way?]

- Planning: Do you have a say in the creation of data collection tools?
- Analysis (knowledge of statistics)

What do you think about your overall involvement in clinical studies?

- Do the possibilities to be involved correspond to your expectations?
- If you could chose an additional research activity to take part in, what would it be? For example, conceiving CRFs, writing publications

[If applicable] What prevents you from becoming more involved?

- Hierarchy, knowledge gap, time constraints...

Topic 6: Role recognition

What do you think about the recognition you get for your work as a < job role > in your research team?

- Lack of recognition/gratitude
- It depends – on the study type, on the investigator

[If applicable] How would you like to be (better) recognised for this work?

- And how would this impact on your current work?

Topic 7: Training opportunities

What training did you receive to prepare you for your specific work?

- Generic research training *vs.* role-specific tasks *vs.* study-specific procedures
- On-site – supervision from more experienced personnel
- Formal training – face-to-face/online...

Do you know about other training opportunities? [If applicable] why did you follow < training X > rather than another one?

- Issues with cost/accessibility
- Time pressure, authorisation from line-manager

Topic 8: Challenges

What is the easiest [then, ask about hardest] part/activity in your work?

- What makes it easy/hard?

What makes your work as a < job role > easier/harder in general? [OR: What makes the planning/execution of data difficult?]

- Think about concrete examples from your experience
- Resources (time, equipment) and tools
- Interactions between team members and with line-managers or above

How do you think your job/data management in general could be (further) facilitated?

- Specific tools
- Better training/guidance/supervision

Tell me about a study when data management went well. What happened?

- What lessons do you take away from this experience? Have you applied them/found them useful in other studies?

And a study when data management went wrong?

- What do you wish you did differently?

Topic 9: Local capacity development

In your experience, what works (or does not) to strengthen the capacity for data management?

- Specific examples from personal experience: situations, successes/issues encountered, underlying reasons
- Different levels: individual, team, institute, region/country

What is missing at your institute to facilitate data management?

- Which tools/resources do you wish you had access to?
- What can (should) organisations at different levels (local, regional/national, international) do?

Topic 10: Data preservation

What happens to data at the end of a study?

- Data preservation format, associated documentation, transfer/sharing...

Do you think they are preserved in an appropriate way to enable long-term use?

- Why those (in)adequate choices?
- Do you think data sharing has a role to play in data preservation? [Why/How?]

C.2 NVivo coding tree and other charts produced during analysis

The final coding tree comprised three major branches: 'Organisational context (structure and resources)', 'People (behaviour and knowledge)' and 'Data (procedures)'. The sub-nodes and contents of these are briefly described in sections C.2.1 to C.2.3. Chapters and sections of the qualitative work (reported in Part III) that were most informed by contents coded under each node are also specified.

C.2.1 Organisational context (structure and resources)

Four children nodes branched under the 'Organisational context (structure and resources)' parent node, as described below. Letter codes associated to each child node reflect the main keyword describing the node's data content.

Node (E): National research Environment, institutions' remit of activities

Node (E) captures 103 quotes referring to existing research institutes at case study sites, and to usual research activities of participants and their organisations. The node also covers aspects of structural organisation, especially regarding what the structure may enable or prevent individuals to do in terms of research and data management. This node's contents support the general description of case study sites and the context for clinical research they are embedded in (Chapter 8, section 8.2).

The following sub-nodes branched under Node (E):

- (E0) Possible research projects within local context (study types)
- (E1) Organisations' mission (goal), activities and competencies
- [descr*] General study procedures (*versus* data procedures)

*descriptive sub-node, whose contents did not play a role in the analysis

Node (R): Limiting Resources (mostly time/finances) affecting processes

Node (R) captures 149 quotes referring to the ('non-people') resources whose availability or scarcity influence clinical research and data management practices. The consequences of limited resources are a cross-cutting theme, hence data coded at this node contributed to the understanding of multiple challenges and coping strategies put in place by participants. However, contents of this node mostly support the general description of case study sites, their context and underlying data manage-

ment systems (Chapter 8, section 8.2 and Chapter 10, section 10.3), as well as the joint interpretation of qualitative and quantitative findings (Chapter 12).

The following sub-nodes branched under Node (R):

- (R0) Financial limitations and funding mechanisms
- (R1) Infrastructure (buildings) and equipment (incl. consumables)
- (R2) Time constraints, delays and competing tasks
- (R3) External factors influencing implemented systems and procedures

Node (S): Structure & organisational arrangements in institution/team

Node (S) captures 160 quotes referring to structural constraints and informal arrangements that affect the distribution of data management and other research tasks among staff. This node's contents greatly informed sections 10.4 and 10.5 (Chapter 10), where I discuss managerial challenges and opportunities at the sites.

Of note, the sub-nodes (S0) and (S2) (and their parent Node (S)) are amongst those that were most inspired by the pre-existing Aragón framework (deductive coding of the data); by contrast with other nodes (and sub-nodes) that were primarily data-driven (inductive coding).

The following sub-nodes branched under Node (S):

- (S0) Structural and organisational hard components (esp. determining tasks distribution)
- (S1) Contracts, recruitment, staff turnover

- (S2) Hierarchy, project coordination and oversight, supervision, team dynamics and competition
- (S3) Perennity, sustainability

Node (P): Procedures or Policies (written or acknowledged)

Node (P) has no sub-node, and captures 36 quotes in which participants directly referred to procedures, policies or rules, whether formal or informal. Many of the quotes under this Node (P) were also coded at Node (S3) 'Perennity, sustainability', thus highlighting the significance of (formal) procedures in facilitating the establishment of a sustainable institution and DMS.

C.2.2 People (behaviour and knowledge)

Five children nodes branched under the 'People (behaviour and knowledge)' parent node, as described below. Letter codes associated to each child node reflect the main keyword describing the node's data content.

Node (N): Collaborations, Network, interlinks, relationships

Node (N) contains 102 quotes across two sub-nodes, and captures interviewees' perspectives on collaborative work and working as a team. Collaborations occur at

different levels, from institutional (micro-level), local or national, up to international (macro-level). These quotes highlight the challenges and opportunities of working together, and particularly contribute to sections 10.4 and 10.5 (Chapter 10).

The following sub-nodes branched under Node (N):

- (N0) Integrative teamwork, knowledge exchange, symbiotic collaboration and recognition (internal, micro-level)
- (N1) Institutional collaborations and involved parties (external, macro-level)

Node (A): Raising Awareness, teaching, creating change and continuity

Node (A) has 78 quotes and two sub-nodes, which capture elements related to the necessity of raising awareness and training people into specific areas ranging from clinical research in general to data management and statistics more specifically. The quotes also highlight the challenges and opportunities of acting as a teacher or driver of change, which are particularly reported in sections 10.4 and 10.5 (Chapter 10).

The following sub-nodes branched under Node (A):

- (A0) Teaching, educating and raising awareness (esp. challenges)
- (A1) Change agents, pioneers (vision)

Node (T): Training, Teaching, learning and career development

Node (T) has 144 quotes under five sub-nodes, some of which capturing data with similar meaning as found in Node (C) – see below. These interview quotes relate to participants' perspectives on the available or missing training opportunities, as well as on the strategies and preferable formats for capacity strengthening of individuals. The contents of Node (T) particularly informed section 10.5 (Chapter 10).

The following sub-nodes branched under Node (T):

- (C1) Individual study, online training, self-learning (drivers and limitations)
- (C2) Continuous Professional Development, short modules, capacity strengthening
- (T0) Formal university/private school education (structure and opportunities)
- (T1) Research students' training
- (T2) Clinical study project-specific and -related training

Node (C): Career path, work motivations

Node (C) contains 69 rather descriptive quotes, which relate to participants' career trajectories and motivations for their job. Information coded under the two sub-nodes contributed to the first two sections of the qualitative results' chapter; sections 10.1 and 10.2 (Chapter 10).

Node (Ind0): Individual competencies

Node (Ind0) has no sub-node, and captures 97 quotes suggestive of the knowledge and skills that need to be covered by some member(s) of staff so that the intended job can be properly done. Eventually, I found that this rather descriptive node did not bring any new information regarding the challenges and opportunities of data management capacity strengthening, hence I did not write a section focussed around this theme.

C.2.3 Data (procedures)

After trying out different ways of breaking down the qualitative data coded under this parent node (for example, by pre-defined step of the data lifecycle, and/or by type of threats to data quality and coping strategies), I found it more helpful to work directly from the larger parent node. I re-read and synthesised into charts (presented hereafter, first for institutes in the Congo, then for Senegal) the 390 quotes coded under the 'Data (procedures)' node. These tables particularly informed the write-up of Chapter 10, section 10.3; and further cover the case of National Control Programmes.

DM STEP	Congolese Foundation for Medical Research (Fondation Congolaise pour la Recherche Médicale, **FCRM**)	Malaria National Control Programme (Programme National de Lutte contre le Paludisme, **PNLP**) **Independent PI** – frequent PNLP consultant	National Public Health Laboratory (Laboratoire National de Santé Publique, **LNSP**)	Hospital (paediatricians and microbiologist/ statistician – Independent PIs) **IRSSA + IRSEN**
Pre-study/Planning (i) Data collection tool	• **Pilot-testing of questionnaires/data collection forms on 'real' patients** – especially to ensure participants understand well (informed consent) + to ensure *"norms"* are respected (AM21)	• Surveys are planned, rolled out and analysed by the National Statistics (demography) Institute, then results are communicated to the PNLP – collection tools include questions about malaria (LLITNs use, availability of health services and medicines…) • Therapeutic efficacy studies of antimalarials (delegated to consultant) rely on WHO data collection forms	• Therapeutic efficacy studies of anti-malarials using WHO data collection forms • Regular surveys always rely on the same questionnaire (updated before each cross-sectional survey) – design led by LNSP biologists with specific research objectives + students have opportunities to add questions; validated by external clinicians in contact with patients/participants	• Students elaborate their own survey form; validated by PI (clinical doctor) • **Pilot-testing of questionnaires/data collection forms on 'real' patients** – especially to check for incoherencies (pilot data included in analysis if forms remain unchanged)
Pre-study/Planning (ii) IT system(s) development, DM soft/hardware	• Unreliable internet and inaccessibility of hosting server (to receive data from the field) acts as a barrier to EDC adoption • Each researcher works on their laptop using MS Excel® (or EpiData/Epi Info™) – no server and network, hence no *"dynamic"* database such as SQL or OpenClinica (although SA21 expressed the wish for such a system, with offline work functionality) • Two desktop computers (at FCRM lab) utilised for data entry in the past, for a specific study; removed upon study completion due to space constraints (unclear what happened to these computers)	• Uses a *"BGCL"* database (implemented with the support of an IT consultant), where to compile all data at the central (programme) level, including data from all departments (and other diseases) • Currently moving to District Health Information System 2 (DHIS-2)	• *"Data entry mask"* (in fact, MS Excel® file template) created by PI or expert statistician • Master (old) database copied-pasted for each student; former values erased, headers (variables) kept as is – adding variables as necessary to account for form's update • MS Excel® chosen because easiest for students and *"non-savvy"* PIs, but the biostatistician *"complains, because often when we work with MS Excel® … he takes a lot of time to transform [harmonise]"* (PI23) • PI22 tried MS Access® *"out of curiosity"* but then reverted to MS Excel®	• Data entry mask (Epi Info™) prepared by student in collaboration with PI/statistician • International collaborations: use of a data entry mask (Epi Info™) prepared by WHO reference centre based in France – for all countries performing WHO-sponsored studies • SA22 learns and programmes using new tools as needed (= as requested by sponsor) – SPSS®, Epi Info™, CSPro • SA22 includes codes to limit errors, because database *"cleaning is difficult"* – favour Multiple Choice Questions; impose value range; limit Yes/No questions, as errors on these can easily occur but are hard to spot

Pre-study/Planning (iii) Other DM documents + Training	• Informal (not written?) procedure for data entry, explained to clerk during recruitment interview – take one form at time; go through it page by page; move the pages you have already entered to the side; do not go back to past pages; if modifying an entered value, signal it… • For *"larger"* studies (meant by participant as studies involving external collaborators), training must be delivered at least to collaborating PI – how to report data on collection forms, with emphasis on the necessity to fill in mandatory fields (yet, high proportions of missing data are an issue)	• Monitoring & Evaluation plan, communicated to each level (district and health facilities) – how to collect data and with which tool; how to compile data, when to send it and to whom (up until data reaches central, programmatic level)	• Training delivered to technicians and clinicians collecting data – especially how to prepare thick blood smears and how to fill in forms according to study protocol • Informed consent procedure – training to communicate the project rapidly and gain consent, while ensuring the participant has understood
Collection (i) Means and sources (collection process *per se*)	• Using paper	• Collection tools for malaria have different age ranges, etc. depending on who collects (SNIS – National Health Information System/Système National d'Information Sanitaire, PNLP, other…) and who to report to – greatly increases the workload of health-workers collecting aggregated data; effort should be reduced with deployment of DHIS-2 (entry at sites of semi-aggregated/disaggregated data?) • Consultant (PI21) uses paper – *"I am a bit old-fashioned"*	• Retrospective studies: data collection from (old) registries maintained by LNSP • Epidemiological studies: recurring collection of patients' samples from hospitals in big Congolese cities; punctual (cross-sectional) collection in other places • Using paper for epidemiological and clinical studies (WHO forms in latter case) – different sheets for collection of demographic/clinical data (including Rapid Diagnostic Test, if available at time of study) in the field; then for reporting of laboratory analyses in Brazzaville (thick blood smears)

Collection (ii) On-site monitoring	• For *"larger"* studies: on-site monitoring (not performed by the data manager, but dedicated staff hired for that activity)	• Quarterly/Biannual supervisions from central level to districts (at least when there was funding from The Global Fund) – drawing at random a representative sample of data and health facilities; surprise on-site visits to check data in the database matches data in registries • Monthly/Quarterly (unclear?) supervisions from districts to health facilities	• Supervision visits by PI – especially when noticing many missing data during entry (in database), going to the field is necessary to improve the situation ahead of next survey
Collection (iii) Tracking	• For *"larger"* studies: *"the return is always difficult"* (SA21) – a lot of missing data needs to be chased back	• *"lengthy…heavy"* official circuit – from health facilities to districts to central level (SNIS) • Parallel circuits put in place by PNLP for faster access to malaria-related data; to enable timely reports and reaction in case of bad practices at collection sites • N.B. Parallel database (on MS Excel®) for quick data tracking and cross-checking also mentioned by trial monitor at the PNT in Senegal	• Paper forms from hospitals/field arrive in folders • Data entry start promptly after arrival of forms – to quickly contact clinicians whose forms have missing data, or parents (of enrolled children) directly	• SA22 flags that missing data is a real issue – *"we don't have a culture of demand on missing data"*; some studies have up to 50-70% missing data (whole pages of collection forms are blank), especially studies collecting data from registries
Processing (i) Data entry	• Studies with FCRM biologists: those analysing samples at the bench enter data themselves in the database (copy of MS Excel® master-file) • For *"larger"* studies: two dedicated data entry staff – unfortunately, these are usually not *"equipped"* to do the job (often, students); they sometimes need continuous supervision • SA21 wishes for unsupervised, individual data entry, with centralisation from different computers to a server – hampered by lack of competent data entry staff and absence of a server/network	• With DHIS-2 system, the mask and application can be deployed for direct data entry (locally, instead of centrally) – already in place in Senegal, unclear in the Congo • For therapeutic efficacy studies of antimalarials (and other investigator-led studies as well): data entry in WHO spreadsheet, *"which is quite practical even if not specifically for data handling"* (PI21)	• Data entered in MS Excel® data files by students mostly – often, one student per site entering forms collected in the field; lab data later added by PI/students • For studies beyond the scope of a student's thesis, data entered by LNSP secretaries • Personal data from participants kept on paper only, not entered in database – because it is not necessary for analysis (rather than for confidentiality concerns)	• Data entered in MS Excel® at time of analysis only, with one new entry episode for each analysis/parameter – *"you create a new table, that is the difficulty … if software are designed, this is the direction to look at; somewhere you compile all data … you click on what you want, it provides [you] what you want"* (PI25) • Data entered in Epi Info™ by student, ideally on a daily basis in parallel to data collection to allow for timely analysis start • SA22 uses data entry clerks + double-entry, if budget allows • Data entry clerks have a document where to log the difficulties they encounter (e.g. 'real' out-of-range value not allowed by system)

Processing (ii) Review/Validation	• Data manager *"looks"* at MS Excel® files provided by biologists after each data entry episode, to flag aberrant values (e.g. off-range); then sits with biologist who entered data to ask them to cross-check within their lab book • For *"larger"* studies: two additional persons hired to control/validate data – in these cases, control is difficult (a lot of mistakes are found) *"it is as if you attempt to redo the same work … that they [data entry clerks] already had to do"* (SA21); look at forms one by one for comparison with MS Excel® data contents (on daily basis, about 10-20 forms are entered and reviewed per day)	• Phone calls from central level to districts, asking questions to clarify whether communicated data is trustworthy – acknowledged as insufficient to guarantee data quality, but only solution in absence of funding for on-site supervision/monitoring • Matching of reported information on medicine used versus reported patients • Data validation meetings, gathering all health facilities from one district, for data review before passing it on to central level • Confronting reports from SNIS (official circuit) and parallel circuit • For therapeutic efficacy studies, the *"pre-programmed"* WHO spreadsheet has built-in controls	• Before export to analysis software, confrontation form by form of source data with entered data – lengthy process, *"it's very, very, very hard … it takes time, and I don't know what you can suggest us"* (PI23); but enables spotting some data entry typos (letters in numerical value, aberrant ages…) • Complete review (not only a percentage of the database) – especially checking for completeness of entered data • Data entered by students is validated by secretaries or PIs; data entered by secretaries is validated by PI	• Using means/curves computed in MS Excel®, to spot out-of-range values that *"shift"* curves – if a mistake is spotted, going backwards to all data and original forms to track the mistake • *"Fastidious"* confrontation form by form, data by data, to *"visually"* check that source data and database contents match – *"it's not very efficient, but it enables correction of certain things"* • On top of data entry controls, SA22 *"juggles with"* SPSS® (boxplot) to spot outliers – Epi Info™ (preferred for DM) does not allow spotting outliers easily • Depending on sponsor, SA22 also hires a supervisor to control/validate data entry (incoherence of frequencies on each variable, aberrant values…)
Processing (iii) Database harmonisation, compatibility	• Data manager and head of lab sometimes come at weekends to compare versions of the MS Excel® master-file (compile independent files where more data has been entered)	• Move to DHIS-2 expected to facilitate harmonisation, and increase completeness/timeliness of collected routine data	• MS Excel® dataset exported to Statistica (Stata®?) for analysis • **No issue with centralising data from different sites (different files) in one MS Excel® master-file**, because templates (headers/variables) are similar – one file for all sites, with one spreadsheet per site • Issues with harmonising/homogenising dataset contents, in absence of systems imposing specific codes at data entry – *"Students were so tired, to the point where [they] put a lot of abbreviations, which say absolutely nothing when you don't know the meaning … [I tried to homogenise]"*	• SA22 recommends Epi Info™ for DM, owing to easiness of file conversion/compatibility – *"I try to flee [avoid] software that can only do one format"*

Analysis (i) Data manipulations during statistical analyses			
• Remaining aberrant values necessitate to come back to source data (stored far away – burdensome) • External statisticians hired on a per-project basis, depending on funding availability – *"[when we] had funding from EDCTP … we had a statistician … who analysed these data, [using some] appropriate software"*	• For therapeutic efficacy studies: performed within WHO spreadsheet – automated calculations (pre-programmed in downloaded file) • Complemented by analyses with Epi Info™ – associations, odds ratios/relative risks (calculated by PI) • Overall, limited interest for detailed data and analyses (PNLP programme managers want an answer to a policy problem) – SA22 gives the example of an (obvious) issue not spotted by programme managers in reports, when their data manager had mistakenly masked some data in MS Excel®, resulting in aberrant sums	• Pls do basic analyses with MS Excel®, e.g. to prepare conference abstracts or quick reports – *"one will never say it enough, and often one doesn't realise it, but MS Excel® has prodigious capabilities … it's a simple name, one thinks it is also simplistic [compared to] other names that are a bit bling-bling and really fancy such as Stata, such as EpiData – one thinks all problems will be solved [by these]"* • For multivariate and other more complex analysis, MS Excel® dataset sent to statistician (external to LNSP and often the one who helped to set up the database, except if mathematics students are currently working within the institute) – analyses using whichever software they prefer	• Described as *"data manipulation"*, with MS Excel® (IRSSA investigator) – to see if there is pharmacological activity/therapeutic efficacy • Moment when a computer becomes useful, to do statistical tests – although *"it is good to understand manually what is done … to be sure the device did not fail"* • One new MS Excel® table created per parameter analysed, i.e. back to source data file for every new analysis – Student's test on means; difficulties analysing qualitative variables, e.g. increase/decrease in diarrhoea or other symptoms • Using SAS® (IRSEN investigator) – version obtained from a colleague who had a licence from his company; or MS Excel®, which is seen as powerful and enough for Master's students theses; *"we ask [the FCRM biostatistician to give] to our students a really sharp training on MS Excel®. Someone may say 'I know MS Excel®', but MS Excel® is vast"* • Using Epi Info™ – classical epidemiology (Chi-2, correlation tests, Schwarz formula); some external statisticians involved in projects may use other software (e.g. SA22 also uses SPSS®, or MS Excel® for survival curves)

Analysis (ii) Reporting	• Mostly to WHO, whether for antimalarial efficacy studies (outsourced to consultants) or national (routine) data • Data from HIS becomes part of an annual report, communicated to WHO for global malaria report – including indicators of morbidity, mortality, prevalence, antimalarial coverage… • As consultant, PI21 reports to sponsor/funder – i.e. to WHO (national and international offices) and PNLP (but rather sceptical the PNLP has the capacity to store reports sustainably) • Scientific publications	• Students' theses are an occasion to receive feedback from jury about proposed analyses • Participation to international conferences – especially in West Africa/Dakar and Marseille • Publications (not so frequently yet, but further efforts on-going) – mostly French journals (Path Exo, Médecine Tropicale)	
Preservation (i) Storage during study, Archiving	• For *"larger"* studies, data collection forms first stored at the FCRM lab; then archived at the administration (WHO campus), because the volume is too large for the available space in the lab	• Data preserved in aggregated form – within database for health information system data, within reports only for antimalarial efficacy studies • Dedicated computer at PNLP offices for storage (password-protected) – database can also be accessed (via network) on other computers; programme manager rarely accesses it, hence lost his password; database handled by data manager solely • Consultant keeps data on his personal computer and shares with student working as co-PI – should provide data to the PNLP *"but they don't have a unit … there's a problem … there's nothing, absolutely nothing"* • SA22 suggested programmes for diseases other than malaria are more *"serious"* and reliable in storing their data/reports • All completed WHO data files sent back to WHO (i.e. raw data), including for studies not sponsored/funded by WHO – also seen as a way to preserve (archive) data	• Paper forms archived at LNSP in folders – sorted by site, survey and year • Each PI may also keep at LNSP the files from their own studies – *"in my office [on the side] there's a big pile (laugh)"*; paper format seen as an essential storage format • Electronic database on dedicated computer, *"which no one touches unless you have data to provide/contribute"* (PI23) • LNSP electronic storage system has limited capacity (mostly devoted to yellow fever data storage) – described by PI22 as unlikely to cope with the volume of data • Each PI also keeps their study data on their own laptop	• Paper forms archived in folders on PI's office shelf – volume is an issue + unstable format, because have to fight against mice eating paper • Long-term storage not seen as critical – *"I don't do this type of studies anymore. If there's too much paper I can throw away … that is not necessary anymore"* • Storage on personal computer – until it breaks down (suggesting no backup plan) • Storage on personal computer and possibly hard drive, no centralised system for research data management in hospitals • SA22 only uses CDs/DVDs (also to communicate data to sponsor/funder) and does not accept USB sticks (viruses) + keeps copy on personal hard drive

Preservation (ii) Backups, security			
• External hard drive	• Backup on DropBox (by data manager) and external hard drives – everyone from PNLP brings a copy at home	• At LNSP level, storage is *"unfortunately"* on computer only (though head of labs is aware that having a copy on external discs would be good practice, as he learnt at IRD in Dakar) – *"that's now I have to answer the question that I realise that we commit the foolishness to not have an external backup"* (PI24) • Each PI has their backup strategy to address the lack of LNSP system: database copy on USB stick – recognised as not durable (external hard drive are desired, and needed to share data with collaborators when travelling); documents shared *"widely"* to all students and other involved in project; database sent via e-mail to oneself • Preservations in one single place (LNSP) considered insufficient – *"we are not politically stable, it can burn over there, and if there is no backup somewhere … we have to diversify the storage …"* (PI23)	• Copies on personal hard drives

DM STEP	UCAD Parasitology (Data Centre)	UCAD Public Health
Pre-study/Planning (i) Data collection tool design	• PIs share draft data collection forms with data managers (and other involved clinical research staff) for comments, review and validation – exchange via e-mails and modifications using track changes + face-to-face meetings for discussion if urgent/necessary • For *"DSS"* (HDS survey), data collection forms' contents depend on villages' organisation (district, village, neighbourhood…) – questionnaires preparation and printing necessitate collaboration with locals, described as *"fastidious"* • Having SA11 as a permanent member of staff (rather than as a consultant, which was formerly the case) enables higher quality in DM/data collection forms, because others can come and ask questions at any time of study	• Data collection tools are designed with intended analyses in mind; hence implication of the whole study team (especially data analyst – statistics/epidemiology expert) • Questionnaires' structure and questions' constraints are adapted based on field testing/deployment – described within the context of a study using EDC • Many data collection forms/tools already exist – part of the expertise of investigators is knowing how to adapt them according to project/sponsor
Pre-study/Planning (ii-a) IT system(s) dev., DM software	• Former *"database engineer"* (data manager) and DM12: use an Access data entry mask, linked to a SQL server – upgrade from past use of Access as the database management system, which was too slow for the volume of data entered and managed • DM11 favours open-source software (MySQL) and EDC (Android) • General agreement that MS Excel® is not a data management software • Database structure updated regularly for DSS (survey with continuous data collection) – a challenging aspect of maintaining such long-term/large-scale databases • Consideration in database design and issue that a data manager can solve if appropriately involved in study design/monitoring: *"when I went to the field to train surveyors and [supervise collection], I noticed some entities do not exist in databases. So I had to modify tables in the background"* (DM11) • Investigators (and data managers too, according to one participant) use illegal software versions because licences are considered too expensive – few controls in Africa, but heads of research units should be incentivised to invest in proper software (and hardware); recently, SA11 convinced to buy the latest MS Office® version, bought with MARCAD funds	• Data entry masks prepared by (or in consultation with) IT technician or by investigators themselves – using EpiData/Epi Info™ or SPSS® (including built-in constraints for data entry) • Tools designed based on paper (in which case the IT technician may support DMS planning and data entry mask design) or EDC (recently – still lacking of in-house expertise to program in Open Data Kit, ODK)

Pre-study/Planning (ii-b) IT system(s) dev., DM hardware	• Too few and not robust enough servers according to both data managers + SA11: a *"powerful"* server is needed, which investigators are *"not ready"* to invest in – would come with personal storage space for each team member + shared space for data, procedures, documentation, etc. (necessitates overcoming the *"fear"* that private data becomes visible to others within the team, by explaining how to select files/folders to sync and how to create restricted access rights) • Connectivity and availability of (in-house) data hosting server act as a barrier to the deployment of EDC – servers rented from ICT providers are costly + unclear security/reliability in data transfer • Internet speed recently upgraded by SA11, to facilitate interactions with international collaborators and access to online videos (training material) – using part of the funding from a clinical study grant • UCAD Parasitology now has ten 9-inch tablets available – bought Nov. 2016	• Internet connection (optic fibre) seen as the responsibility of central university IT services; local network of the faculty/department is the responsibility of UCAD Public Health IT technician • Availability of a hosting server within the country acts as a barrier for EDC deployment; data hosting services for the few UCAD Public Health projects using EDC are outsourced to university partner in the USA – along with a lack of willingness to invest in this new technology, the unavailability of a server really slows down the adoption of EDC tools at UCAD Public Health
Pre-study/Planning (iii-a) Other DM documents/procedures	• Unanimously, participants expressed that data managers should be involved in study design (and all steps of the study) – a practice that is not yet anchored in habits of all clinical researchers in Africa: *"The big problem in Africa … it's that the data manager, he's involved once data has been collected … it's the sinews of war … you can't convince people if you don't have quality data. And the data manager, he must be there, from project conception up to the end, to ensure that all data, all the data flow has been well managed. But you can't limit data management from database conception to clean-up"* (DM12) • Data managers' involvement in study has increased at UCAD Parasitology in the past 2 years • Data managers' expertise enables them to design appropriate DMPs, collection tools and databases, which reflect the study protocol and objectives (especially through appropriate questionnaires/databases' structure and coding dictionaries). Both data managers shared experiences of getting involved midway through a study, when they encountered issues that they could have easily solved at the study onset, but could not always correct for an on-going study. • Example: In the past + externally to UCAD Parasitology, a data manager was asked to create a database for a study where questionnaires (and samples) had already been collected. The DM team noticed duplicate IDs on different collection forms. Provided with the questionnaires but not informed of the existence of matching samples, they created new IDs (without keeping good track of changes they made), and many samples could not be matched with their clinical data.	• Similarly to UCAD Parasitology data managers, PI12 notices that colleagues from other departments/external to UCAD do not commonly involve experts (statisticians or epidemiologists) in study design; they then approach UCAD Public Health for support with analyses that are impossible to do (e.g. calculating a population age mean when age has been collected as a factor/range variable) • IT technician may be involved in budgeting studies – possible DMS will vary depending on resources, and study budgeting should conversely take account of necessary DMS cost

Pre-study/Planning (iii-b) Training	• Critical point for high-quality data: fieldworkers collecting data must be very well trained • A procedure is written to guide fieldworkers' training, which is directed by 2 members of staff (PhD student/project managers) • Training covers the study procedures, which detail how the study should enfold (i.e. what fieldworkers should do) – including informed consent procedure, participant enrolment process, questionnaire completion using data collection guide… • Usually, training lasts 48-72h (depending on study/tools complexity) for experienced fieldworkers hired in Dakar ahead of going to the field – tools are also finalised with fieldworkers on this occasion, before printing forms (paper) or deploying the template onto tablets (EDC) • Fieldworkers (hired locally, e.g. health-workers) are not yet accustomed to EDC – familiarisation with tablets and training is critical, especially on how to save data until its transfer in areas where connectivity for real-time transmission is poor	• Some fieldworkers are not accustomed to EDC – it happened that data was lost because a fieldworker did not save/transfer it properly (this issue remains of limited scope, as this happened for one fieldworker in a national-scale study)
Collection (i) Means and sources	• Depending on study scope and needs, data is collected by local health-workers and/or by experienced fieldworkers recruited in Dakar (who have experience working with UCAD or IRD, or other research institutes) • Former "*database engineer*" (data manager) and DM12: usually rely on paper • DM11 and SA11 try to introduce EDC, which is also used by more and more "*people in the private sector*" • For DSS, transcription from health centres' registries onto notebooks provided by investigators/UCAD team, according to clearly defined instructions • For lab data (thick blood smears), 2 readings made by 2 technicians in different institutes – a third one involved if disagreement beyond 25% (otherwise, the mean of 2 measurements is recorded)	• Collection tool mostly depends on sponsor: "*if the sponsor imposes tablets, we take [some], if the sponsor has nothing we have to work on paper*" (PI12) • Using paper, EDC or scanners – example of one study (that did not work so well), when all pages of health registries at study sites were scanned • EDC relying on ODK and Samsung Galaxy tablets (android technology), which PI12 found "*very practical*" – example of one study sponsored by the World Bank
Collection (ii) On-site monitoring	• Data managers should go in the field and interact with fieldworkers – enables solving issues: "*when you stay in the office, there are issues that you can't solve here … You're supposed to be on the ground to be able to bring solutions*" (DM11) + "*when you are in the field, there you understand the origin of errors; the nature of data you're managing. That's what's important, and that's what people didn't understand initially*" (DM12) • For clinical trials, data is checked (monitored) in the field before centralisation for data processing; for DSS too • QC procedure for validation supervisors (DSS case): verify that the whole period of study has been transcribed, and that the totals of registries/pages/entries are correct… • Check matching of a specified fraction of data between copied registry and source registry (at health centre); generate random numbers to select which pages of registries should be checked – if >30-50% errors (?), then draw another sample; if still >30-50% errors (?) in second sample, then re-do the whole transcription	• Supervisors check questionnaires on-site (in the field) before centralisation for data processing

Collection (ii) On-site monitoring (continued)	• Monthly supervisions by UCAD Parasitology clinical research team (students, research assistants/project manager and/or data manager – study zones to visit are distributed among members); to check whether fieldworkers understood the study process + whether questionnaires are properly and completely filled… • In addition, a local supervisor is recruited on-site (above fieldworkers in the hierarchy, e.g. district medical doctors) and weekly visits the fieldworkers (health-workers); to check they have understood the study procedures and inquire about issues – can phone UCAD Parasitology clinical research team (including PI or statistician too) in case of issue they cannot resolved • Weekly/Daily calls from UCAD Parasitology clinical research team (study zones to call are distributed among members); to try and solve problems remotely in-between 2 visits; possibility to involve regional medical doctors	
Collection (iii) Tracking/Transfer	• Paper questionnaires sent back and forth for corrections, which may result in data loss (or delays) and is difficult because of the large distance between collection site(s) and DM centre – transfer/tracking forms were established *"a bit late"* to facilitate the process • EDC allows (almost) direct transfer to DM centre and verification by data manager – can feed back 'real-time' on issues to solve, before fieldworkers leave collection sites • However, data may get lost during transfer with EDC… • Hybrid system (decentralised data entry): data collection on paper, data entry on tablets (in the field) – a local supervisor is in charge of checking forms and entering data at the end of the collection day, hence can still find the fieldworker and correct data if needed	• QC at data collection point: EDC enables constraints on source data entered – as opposed to paper-based collection, where similar errors are spotted (later) during data entry, and require sending forms back to the field for correction • Tracking is easier with EDC, as the data manager can check the database (almost) in real-time and communicate issues before fieldwork is over – rather than having to wait until data entry and validation for this feedback from the data manager • No need to track forms with EDC, but data may get lost during transfer
Processing (i) Data entry	• Still the most frequent system: centralisation of paper forms at UCAD data centre for data entry by professional clerks + students integrating lab data to existing database where questionnaires have been entered (?) • Data entry is supervised by data/project managers, so that data entry clerks can come and ask questions at any time if encountering issues • Errors occur when data entry clerks get too tired, or do not ask for clarifications about unclear handwriting/spelling • Unanimously described as a lengthy process – including because of subsequent data cleaning and validation + necessity to chase back correct information from the field (although those should be minimal thanks to the on-site verification steps) • Double data entry takes even more time – not so commonly used	• Performed by data entry clerks – errors usually occur because of tiredness • Lengthy process – study may go *"wrong"* (= be extremely delayed) when this is under-estimated, as happened for the study where data was scanned in the field • Using specific software: Epi Info™, R, SPSS®, Stata®

Processing (ii) Validation	• QC procedure for data entry – first, accuracy is controlled (for each clerk, 10% of entered data checked against source data); then, consistency between variables and questions' structures/constraints + completeness of data (does the number of records into the database correspond to the number of collection forms, etc.) • If over 10% of errors found, the whole data entry must be redone – "*difficult*" moment, as it further increases data entry time	• Post-entry validation is performed (e.g. using consistency checks on applicable variables) • Double-entry preferable, may be used depending on budget allowance
Processing (iii) Database harmonisation	• Harmonisation of data from recurring study is challenging (e.g. for DSS – demography and health survey database = 600,000+ individuals, i.e. very large database) • Many inconsistencies may occur (e.g. duplicate villages' names in different districts, or different spelling of one same village; missing individuals in one household, or duplicate IDs…) – necessitate careful data codes and fields for district, village, neighbourhood, etc. + sufficient staff • Example of one year, when only 2-3 DM staff were employed to handle the whole DM process (data entry and validation) for the DSS: inconsistencies were found in all data extractions, either due to data collection or to data entry errors, and a lot of back and forth was necessary before database lock (let alone analyses)	• No dedicated DM unit (as opposed to UCAD Parasitology); everyone from UCAD Public Health is involved, according to area of expertise
Processing (iv) Extraction	• Data exploitation (analysis) format is different from data entry format (file) – when converting or transferring data, there may be issues: "*it can happen that formats change, that we can't figure out anymore [which file is which]*" (SA11) • Data sent to statistician suggested poor version control – hence SA11 requested that data is extracted and sent for analysis by data manager (not by investigators)	• Almost all databases are transferred/preserved in an MS Excel® format, because most software have .xls(x) as an export option
Analysis (i) Data manipulations/Statistics	• Analyses may be very delayed: large volume of data vs. limited resources (statisticians) – one statistician (hired as project manager) in charge of most analyses + external collaborators (especially in the UK) • Inconsistencies noticed by UK collaborators (not due to English/French format compatibility issues) – prompted late consistency checks on database as a whole, which highlighted that children took medicine before being born; not all inconsistencies can be resolved at such late stages of the study (hence data may become unusable for analysis) • Datasets should be accessible on a server (shared space), to enable prompt correction of errors – including those spotted at late stages (analyses) • SA11 software preference: R – free, open-source software; easier/faster to download, install and use than Stata®	• Everyone at UCAD Public Health department has a field of expertise, and is accordingly able to analyse and manage data • Using Stata®, R, Epi Info™, SPSS®… For logistic or linear regressions, Epi Info™ is not adapted, but it is easiest to use (because it has French translations) • Students and colleagues from other departments (and external to UCAD) turn to UCAD Public Health for support with analyses – sometimes too late, because the way data is coded makes the intended analysis impossible (e.g. deriving an age mean from data collected as age range)

Preservation (i) Storage during study, Archiving	• Archiving of paper questionnaires/registries is problematic (because it takes a lot of space), but considered useful (because repeated survey rounds may be carried out and require checking back old data) – for DSS, data is stored *"in boxes … by district, by village"* and by registry/questionnaire type • Recently, SA11 has introduced an online shared space for centralised storage of key data (MS OneDrive®, *"the Platform"*): datasets + procedures and study documentation, dictionary, other metadata (e.g. study and SMC date) – access and understanding of data then becomes independent of researchers who come and leave (i.e. sustainable re-use of data) • SA11 likes Stata® format for storage, because this software allows dataset annotation (metadata) • Data from studies conducted in the past 4-5 years has now been cleaned and archived on MS OneDrive® with its documentation – may become useful for meta-analyses • Storage system should include version control (audit-trail, which is possible in MS OneDrive®) – investigators are not always sure of what data they have, hence the necessity for an intermediary (project manager) to oversee the process • An online, centralised storage system is also useful for sharing and data access from anywhere (provided access rights are granted) – facilitates the establishment of new collaborations, as PIs can easily download and show raw data to colleagues (e.g. when travelling to conferences); formerly relying on e-mails, not ideal because files may be too large to be uploaded and attached • SA11 (and DM team as a whole) currently unclear about ethical requirements, but he wishes to address this knowledge gap: how long should we/is it allowed to store data from participants? – At present, everything that is printed (paper format) is kept	• Preservation seen as useful for potential re-analyses/meta-analyses • IT technician oversees data storage only during study lifecycle (but may also retain a copy); afterwards the data is archived by investigators on their computers + reference data is copied on the department's hard drive (shared between Public Health and Biotoxicology departments) – *"once we've entered data, reviewed information … each one of us [here] manages the database"* PI12 • Some sponsors request only a report, some request also the data
Preservation (ii) Backups, security	• Online server (MS OneDrive®), i.e. online backup + Automatic backup on local server + Manual copies on external hard drives – SA11 also saves copies on his personal computer at home • Automatic backup used to be (only) on a computer within UCAD Parasitology; over time, data managers realised that backups should also be on a remote computer – hence diversification of copies, including outside university walls. • (Daily) downloads from shared MS OneDrive® folder; backup copies should be created before modifying master files (otherwise old data may get lost or hard to locate, because data files are overwritten) • Security issue: People do not always have appropriate antivirus software	

Appendix D

First-author articles

D.1 First publication: Systematic Review of Antischistosomal Trials

The following article [166] is reproduced from PLoS Neglected Tropical Diseases (creative commons license), and describes part of the results covered in Chapters 3 to 6. I also presented these results at a workshop organised by WHO and TDR in September 2015, when schistosomiasis stakeholders were consulted about the feasibility and future directions for an antischistosomal efficacy data-sharing platform [133].

PLOS | NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

The Schistosomiasis Clinical Trials Landscape: A Systematic Review of Antischistosomal Treatment Efficacy Studies and a Case for Sharing Individual Participant-Level Data (IPD)

Amélie M. Julé[1,2]*, Michel Vaillant[3], Trudie A. Lang[1,2], Philippe J. Guérin[2,4], Piero L. Olliaro[2,5]

1 The Global Heath Network, University of Oxford, Oxford, United Kingdom, **2** Centre for Global Health and Tropical Medicine, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom, **3** Competence Centre for Methodology and Statistics (CCMS), Luxembourg Institute of Health, Luxembourg, Luxembourg, **4** WorldWide Antimalarial Resistance Network (WWARN), University of Oxford, Oxford, United Kingdom, **5** UNICEF/UNDP/World Bank/WHO Special Programme on Research and Training in Tropical Diseases (TDR), Geneva, Switzerland

* amelie.jule@gtc.ox.ac.uk

CrossMark click for updates

OPEN ACCESS

Citation: Julé AM, Vaillant M, Lang TA, Guérin PJ, Olliaro PL (2016) The Schistosomiasis Clinical Trials Landscape: A Systematic Review of Antischistosomal Treatment Efficacy Studies and a Case for Sharing Individual Participant-Level Data (IPD). PLoS Negl Trop Dis 10(6): e0004784. doi:10.1371/journal.pntd.0004784

Editor: David Joseph Diemert, The George Washington University School of Medicine and Health Sciences, UNITED STATES

Received: January 19, 2016

Accepted: May 27, 2016

Published: June 27, 2016

Copyright: © 2016 Julé et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: AMJ holds a Nuffield Department of Medicine (NDM) Prize Studentship, jointly covered by the Medical Research Council UK and the NDM at the University of Oxford. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Background

Schistosomiasis control mainly relies on preventive chemotherapy with praziquantel (PZQ) distributed through mass drug administration. With a target of 260 million treatments yearly, reliably assessing and monitoring efficacy is all-important. Recommendations for treatment and control of schistosomiasis are supported by systematic reviews and meta-analyses of aggregated data, which however also point to limitations due to heterogeneity in trial design, analyses and reporting. Some such limitations could be corrected through access to individual participant-level data (IPD), which facilitates standardised analyses.

Methodology

A systematic literature review was conducted to identify antischistosomal drug efficacy studies performed since 2000; including electronic searches of the Cochrane Infectious Diseases Group specialised register and the Cochrane Library, PubMed, CENTRAL and Embase; complemented with a manual search for articles listed in past reviews. Antischistosomal treatment studies with assessment of outcome within 60 days post-treatment were eligible. Meta-data, i.e. study-level characteristics (*Schistosoma* species, number of patients, drug administered, country, etc.) and efficacy parameters were extracted from published documents to evaluate the scope of an individual-level data sharing platform.

PLOS | NEGLECTED TROPICAL DISEASES

Competing Interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: PLO is a staff member of the World Health Organization (WHO); the authors alone are responsible for the views expressed in this publication and it does not necessarily represent the decisions, policy, or views of their organizations.

Principal findings

Out of 914 documents screened, 90 studies from 26 countries were included, enrolling 20,517 participants infected with *Schistosoma spp*. and treated with different PZQ regimens or other drugs. Methodologies varied in terms of diagnostic approaches (number of samples and test repeats), time of outcome assessment, and outcome measure (cure rate or egg reduction rate, as an arithmetic or geometric mean), making direct comparison of published data difficult.

Conclusions

This review describes the landscape of schistosomiasis clinical research. The volume of data and the methodological and reporting heterogeneity identified all indicate that there is scope for an individual participant-level database, to allow for standardised analyses.

Author Summary

Schistosomiasis is a parasitic disease prevalent in tropical climates, especially Africa. To control morbidity and eventually eliminate the disease, the World Health Organisation recommends that populations living in endemic areas be preventively treated with praziquantel. This means hundreds of million treatments over time, which calls for careful monitoring of drug efficacy. Meta-analyses of aggregated data from clinical trials support this policy, but conclusions are limited because studies use diverse methodologies, and detailed analyses (e.g. by age-groups) are not possible. Access to primary data generated by trials would strengthen the evidence and allow answering key questions, such as what the optimum paediatric dose is. This review assesses the number of trials conducted on antischistosomal drugs and their characteristics (where studies were conducted, in which population, what drug(s)/dose(s) were administered, and the methodologies applied), to determine the merit and scope of a common database of antischistosomal efficacy data.

Introduction

Despite its heavy burden worldwide, schistosomiasis remains a neglected tropical disease (NTD) which fails to attract enough attention from research and funding bodies to generate novel drugs and diagnostics [1,2]. Currently, only one drug, praziquantel (PZQ), is available to treat schistosomiasis [3]. The World Health Organization (WHO) strategic plan for schistosomiasis control aims at controlling morbidity due to schistosomiasis by 2020; eliminating schistosomiasis as a public health problem by 2025; and interrupting transmission of schistosomiasis in selected countries in Africa and all other countries in the rest of the world by 2025 [4]. To reduce morbidity and transmission, the WHO recommends the use of PZQ, given as preventive chemotherapy to endemic populations [3]. Along with evidence from pre-clinical studies [5,6] and individual clinical trials, the efficacy and safety of PZQ is further supported by systematic reviews and meta-analyses of clinical trials [7,8]. However, aggregated data meta-analysis suffers from the heterogeneity in the design and reporting of the published studies [9,10]. This lack of standardisation is a common problem across several NTDs; specifically, in the case of schistosomiasis, it concerns diagnosis and test of cure (diagnostic tests tend to be performed on variable numbers of samples per individual), outcome measures (cure rate (CR),

PLOS | NEGLECTED TROPICAL DISEASES

or egg reduction rate (ERR) calculated by arithmetic or geometric mean), and time of assessment [11]. Access to detailed, individual-level participant data (IPD) for meta-analyses would circumvent some of those limitations and allow harmonisation of relevant variables and analyses [9,10], thus enabling more meaningful analyses and comparisons across individuals or studies. This could unravel yet unknown aspects of drug efficacy and safety, in specific populations and over time. Such a model for systematic review and secondary use of clinical trials' data has brought valuable evidence for other disease areas, whether non-communicable (as reviewed in [10]) or infectious (e.g. malaria [12]).

Monitoring PZQ efficacy is crucial, considering the size of the problem and the inherent risks of resistance [13]. In 2013, of the some 260 million people requiring treatment for schistosomiasis (of whom, 46% school-aged children), 13% did receive preventive chemotherapy with PZQ. While still far from the intended programme target, this is close to 40 million treatments (covering two thirds of the school-aged children at need) [14].

Recognising the potential benefits of a shared IPD database, the UNICEF/UNDP/World Bank/WHO Special Programme on Research and Training in Tropical Diseases (TDR) and the WHO programme for NTDs are currently working with a range of stakeholders to build a dedicated, sustainable data-sharing platform for antischistosomal efficacy studies [15]. This initiative also aims to respond to growing requirements for researchers to share their data, coming notably from funding bodies, journal editors, WHO and to some extent, the public [16–21].

The present work was performed to lay the groundwork for the development of the schistosomiasis data-sharing platform, and had three main objectives. Firstly, we sought to identify efficacy studies of antischistosomal drugs performed since 2000, and published between January 2001 and May 2015 (either as an article or as a conference abstract). We followed PRISMA guidelines [22] (S1 Checklist), so that the resulting list of studies could form the basis for future collection and meta-analyses of IPD. Focus on recent studies is for practical reasons (datasets less likely lost, corrupted or unreadable), but we would encourage investigators to share data from earlier studies too, if available. Secondly, we aimed to assess the volume of such IPD and its inherent characteristics, e.g. species, treatment, participants' age-group, and study site. Such an assessment gives an overview of the expected range of data to be potentially assembled, and can help design the standardised database by highlighting key variables and meta-data to consider in its structure. Thirdly, we investigated further the heterogeneity in study design and methodologies that has already been highlighted by past systematic reviews and aggregated-data meta-analyses [11].

Methods

Literature search

To identify relevant references, the following journal databases were interrogated: the Cochrane Infectious Diseases Group specialised register and the Cochrane Library, PubMed, CENTRAL and Embase. Bibliographic references of previous meta-analyses of antischistosomal trials [7,8,11,23,24] were also manually examined. No restriction on publication language was used. Where possible, search results were filtered for 'human' as the species of study.

Keywords belonged to three main themes reflecting the purpose of the search: (1) schistosomiasis [(1.1.) schisto*; (1.2.) 'schistosoma hematobium' OR 'schistosomiasis haematobia' OR 'schistosom* haematobi*' OR 'urinary schistosom*' OR 'urogenital schistosom*'; (1.3.) 'schistosoma japonicum' OR 'schistosomiasis japonica' OR 'schistosom* japonic*' OR 'oriental schistosom*'; (1.4.) 'schistosoma mansoni' OR 'schistosomiasis mansoni' OR 'schistosom* mansoni' OR 'intestinal schistosom*' OR 'bilharz*']; (2) clinical study [treatment OR trial OR random* OR double-blind* OR cohort OR prospective]; and (3) specific drugs [(3.1.)

praziquantel OR artesunate OR artemether OR metrifonate OR albendazole; (3.2.) oxamniquine OR mefloquine].

Of note, although albendazole has not shown efficacy against *Schistosoma spp*., it was included in the keywords for search in order to better identify studies performed by the soil-transmitted helminthiases (STHs) community which may have looked at schistosomiasis aside of their major study.

Four subsequent searches were performed between 25 May 2015 and 12 June 2015. The three first focussed on one major pathogenic species of *Schistosoma* each (*S. haematobium*, *S. japonicum* and *S. mansoni*), and thus used keywords (1.2., 1.3. or 1.4.) AND (2) AND (3.1.). As they constituted an update of previous (published) Cochrane systematic searches, the two searches for literature on *S. haematobium* and *S. mansoni* were limited to records published from 2000 up to the date of search (25 May 2015). The search for *S. japonicum* used no such restriction on publication date, as no prior Cochrane search had been performed. The fourth, complementary search looked at all species altogether and was restricted to publications from 2001, but focussed on different drugs that were missing in the initial searches, therefore combining keywords (1.1.) AND (2) AND (3.2.).

Eligibility screening

Only abstracts published from 2001, and whenever necessary, the corresponding full texts, were eventually screened for eligibility. In line with the primary objective of estimating the size of a global database of antischistosomal efficacy data, and as the likelihood of retrieving datasets dating back before 2000 is much more limited [25], it was decided to exclude articles reporting on studies that had been completed before 1 January 2000. Abstracts of conferences run before 1 January 2014 were also excluded, owing to the difficulty in establishing whether earlier abstracts were duplicate records of articles published within the considered time period. Finally, all case reports or case series, as well as non-primary research studies (meta-analyses, reviews, textbook chapters, opinion papers and comments) were excluded.

Regardless of their design and associated risk of bias, remaining studies where antischistosomal (registered or experimental) drugs had been administered to human subjects in an endemic setting were eligible, provided that: (i) at least a subset of participants was positively diagnosed for infection with at least one of the three *Schistosoma* species of interest, prior to receiving treatment; and (ii) a post-treatment diagnostic test was performed on the same participants to assess outcome within 60 days, where day 0 is the day of administration of the first (and in most cases, single) treatment dose.

The WHO recommends assessing the outcome within 21 days of therapy, in order to distinguish between cases of treatment failure and cases of early reinfection [26]. However, the maturation time of *Schistosoma* worms in the human host induces a delay of at least 4–5 weeks between the moment of infection and the excretion of eggs [27]. Within the 60 days post-treatment window, individuals who were initially cured but got re-infected are still likely to excrete low numbers of eggs (if at all), and to appear negative upon diagnosis by egg count. The 60-day limit was thus retained, to avoid excluding too many efficacy studies which performed an outcome assessment beyond the recommended 21-day timeframe.

Data extraction and analysis

Relevant information from eligible studies published in English or French was extracted by one researcher, using a detailed variable dictionary to facilitate consistent extraction. Help from co-authors and from native Chinese speakers was sought, as required. Extracted information was recorded using a spreadsheet document (Microsoft Excel 2000), and analysed using R (version

3.1.3) [28]. The source data used for analysis, the associated dictionary and the R code are all available as supporting documentation in S1 Dataset (Zip).

Estimation of the number of participants

Some of the eligible studies had a broad scope, and were looking simultaneously at schistosomiasis and at other parasitic infections, in particular STHs. Articles on such studies did not systematically report on the exact number of participants with a *Schistosoma* infection at baseline, thus the need to estimate this number for the purpose of this review.

Whenever possible, the number of *Schistosoma*-positive participants was re-calculated from the specified prevalence of infection among enrolled participants. If the prevalence amongst enrolled participants was unknown, the reported prevalence in the study area was used instead. Finally, in the rare case when there was no mention of prevalence at all, half of the cohort was arbitrarily considered infected with *Schistosoma spp*. at baseline (i.e. assumed prevalence of 50% among enrolled participants).

Of note and unless more specific information was provided in the article, the retained frequency of infection (exact prevalence in study population, reported prevalence in study area, or 0.5) was then consistently applied as a correction factor when estimating the number of relevant participants in subgroups (e.g. treatment group, group followed-up at later time-points): this implies that the subgroup characteristic and the initial infection status of the participant would be independent.

Quality control

After extraction and coding of all information as per the variable dictionary, the datasets were checked for potential errors. To do so, each article was systematically re-examined for variables considered most 'impactful'. These were primarily the variables recording numbers of relevant participants (*Schistosoma*-positive at baseline, and who successfully received the assigned treatment regimen) at both enrolment and follow-up, and in each assignment arm; because those data directly impact on the total number of IPD, whose estimation was our primary objective. The species and drug(s) under study were also double-checked, as the information coded in those variables was used to filter articles by category (*S. haematobium* vs. *S. mansoni*, PZQ vs. other drugs, etc.) and to create specific datasets for analyses by sub-category. Finally, impactful variables also included the time-points of post-treatment assessment of outcome, as they influenced inclusion/exclusion of the article from the analysis.

This double-checking led to the correction of 3 to 12 errors per variable, mainly due to ambiguity in the original articles (e.g. unclear flowchart of included, treated and followed-up participants), and their correction did not significantly change the total of estimated IPD.

A subset of articles was also randomly selected for full, independent re-analysis. This step unravelled 3 typos, which, spread across a total of 45 variables (excluding the core variables previously double-checked) and 12 articles, suggest an error rate of 0.56% in the data as a whole, and of $>1\%$ per variable.

Results

Search results, number of eligible studies and corresponding IPD

The three species-specific searches of the literature in electronic resources yielded 545, 440 and 272 study abstracts for *S. mansoni*, *S. haematobium* and *S. japonicum*, respectively. The later search for additional drugs gave only 11 additional results, of which 6 abstracts reported on studies of *S. mansoni*, and 5 on *S. haematobium*. In parallel, 190 references were identified in

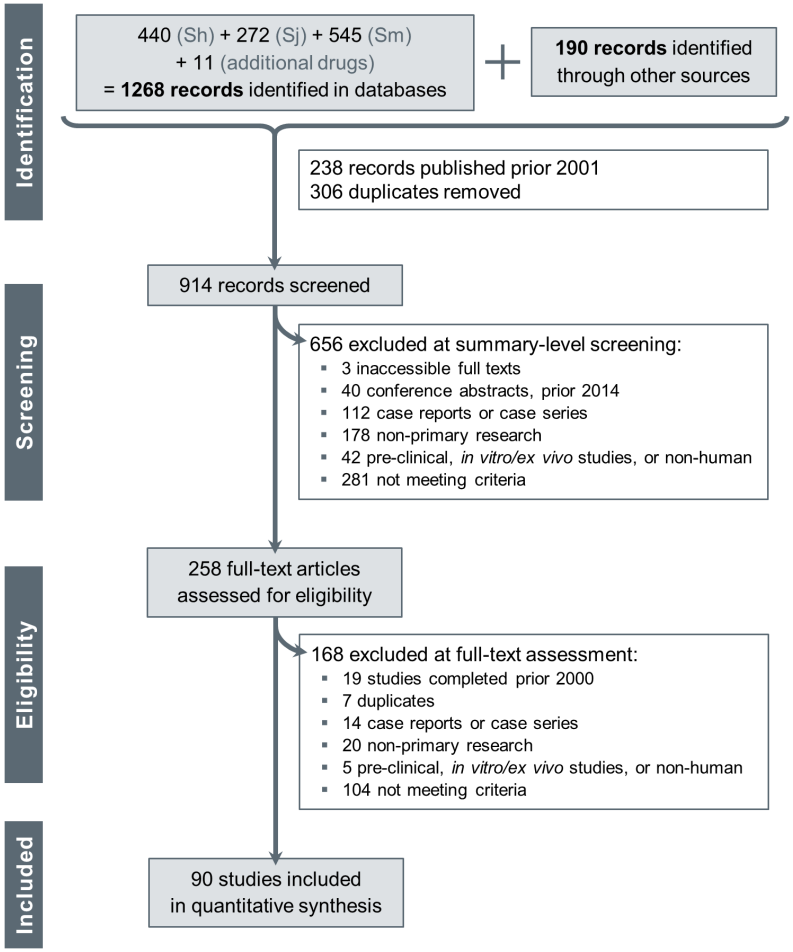


Fig 1. PRISMA diagram. Flow of the study references from identification during the search, through screening and eligibility and up until inclusion. The search strategy and eligibility criteria are detailed in the methods section. Briefly, all clinical studies of antischistosomal drugs, completed in 2000 or later and performed in an endemic setting, were included if at least some of the recruited participants were diagnosed for infection with *Schistosoma spp*. before and within 60 days after treatment.

doi:10.1371/journal.pntd.0004784.g001

the bibliography of published meta-analyses [7,8,11,23,24]. Overall, those amounted to 914 unique abstracts published since 1 January 2001, of which 90 studies were retained after eligibility screening (Fig 1).

Although most studies could be thoroughly evaluated, the case of three abstracts [29–31] remained unclear, and we could not access the corresponding article. In addition, the eligibility of two other studies [32,33] could not be ascertained despite access to full text. At this stage, and as this review is a preliminary phase prior to possible collection of IPD and performance of clinically relevant meta-analyses, no attempt was made to contact the authors and clarify missing information. This step shall be performed later, along with a public call addressed to the

schistosomiasis community to help identify more studies that might have been missed by this scoping review (e.g. studies that were not published, or published in non-indexed journals).

The 90 included studies correspond to 104 cohorts, where a 'cohort' is the largest possible subset of study participants from the same country, which followed the same protocol and thus share the same study-level characteristics (except for the assigned intervention in multi-arm studies). Most studies (79) involved a single cohort, but some looked at different groups such as infants and adults who were followed up at different time-points, or involved sites across several countries with different parasite species and diagnostic approaches. Those study groups were thus considered to be separate cohorts, and the analysis by cohort is preferred whenever comparing the specific methodologies and protocols applied.

The 90 studies enrolled an estimated total of 20,517 participants of relevance for schistosomiasis treatment efficacy meta-analyses. One conference abstract did not provide any information to estimate the number of participants: accordingly, this record did not contribute to analyses by number of participants, but was still considered for analyses by cohort or study. Throughout this article, whenever mentioning a number of participants, it is derived from the number of enrolled participants infected with *Schistosoma spp*. at baseline, and who received the full treatment regimen they were assigned. The number of participants refers to the initial cohorts, as losses to follow-up were inconsistently reported and therefore could not be accounted for here (estimates suggest that the total of participants with complete data is about 17,500).

IPD characteristics of relevance for clinical or epidemiological analyses

A major advantage of statistical meta-analyses based on IPD pooled from multiple studies is to provide large numbers. With this method, specific sub-populations whose outcomes are usually 'hidden' in aggregated data could be individually considered for eligibility in re-analyses on a case-by-case basis, depending on the objectives. Here for instance, 26 cohorts are formed of participants infected with several possible parasites, which made it difficult to extract information on *Schistosoma*-infected participants using published information only (Fig 2). Depending on the purpose of the meta-analysis, it could be interesting to retrieve data about these

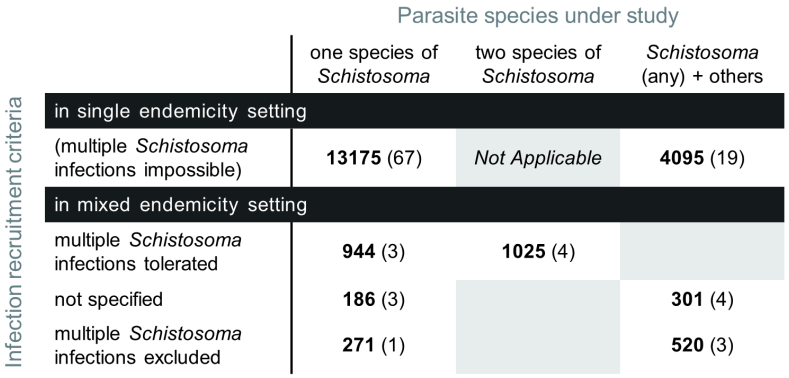


Fig 2. Type of *Schistosoma* infection required for recruitment in the study, depending on the endemic setting and parasite of interest. Numbers in bold are totals of participants (out of 20,517) recruited in the specified setting (single or mixed endemicity), applying the corresponding criteria (multiple *Schistosoma* infections tolerated or strictly excluded) and for the study of the corresponding parasite(s) (one or two *Schistosoma* species, and possibly other parasites as well). Numbers in brackets are corresponding totals of cohorts (out of 104).

doi:10.1371/journal.pntd.0004784.g002

TREATMENT	N
praziquantel (PZQ)	**18302**
Standalone	
PZQ (40 mg/kg)	16294
PZQ (50 mg/kg)	150
PZQ (60 mg/kg)	1125
PZQ (80 mg/kg)	156
PZQ (120 mg/kg)*	51
PZQ (180 mg/kg)*	36
in combination	
PZQ (40 mg/kg) + anthelmintic	73
PZQ (40 mg/kg) + antimalarial	229
PZQ (40 mg/kg) + ARA	90
PZQ (60 mg/kg) + antimalarial	53
PZQ (120 mg/kg)* + antimalarial	45
other antischistosomal	**329**
Oxamniquine	118
Myrrh	211
other anthelmintic	**259**
Ivermectin	66
Triclabendazole	49
Moxidectin	144
antimalarial therapy	**1206**
artemisinin-based combination therapy	
Artemether-Lumefantrine	31
Artesunate-Amodiaquine	12
Artesunate-Mefloquine	18
Artesunate-SP/SMP	583
other antimalarial drug	
Artesunate	478
Mefloquine	67
Sulfadoxine-Pyrimethamine (SP)	17
control and other	**421**
Placebo	330
Arachidonic Acid (ARA)	91

Fig 3. Antischistosomal treatments tested in the included studies. N: Number of participants, infected with *Schistosoma spp*. at baseline, and who successfully received the complete assigned regimen. * Treatment dose delivered in split doses of 20 mg/kg every 4 hours (PZQ 120 mg/kg), or 60 mg/kg every 24 hours (PZQ 180 mg/kg).

doi:10.1371/journal.pntd.0004784.g003

estimated 4,916 participants infected with schistosomiasis, and treated with drugs having anti-schistosomal activity as part of larger studies focussing on other parasites (Fig 2).

The 104 cohorts were also recruited in varied endemic settings and countries, and tested several therapies—mostly various doses and regimens of PZQ, but also a dozen other therapies (Fig 3). Considering the large amount of participants having received PZQ at the dose of 40

AFRICA	Cohorts	N
Eastern Africa	40	7625
Uganda	15	1998
Kenya	4	1876
Zimbabwe	11	1823
Ethiopia	6	803
Tanzania	2	560
Mozambique	1	556
Malawi	1	9
Western Africa	30	5669
Niger	5	1872
Cote d'Ivoire	10	1063
Mali	1	800
Nigeria	5	732
Senegal	4	667
Mauritania	2	337
Ghana	1	70
Liberia	1	70
Togo	1	58
Northern Africa	10	2161
Egypt	8	2045
Sudan	2	116
Middle Africa	7	2026
Cameroon	3	1266
Gabon	2	365
Angola	1	325
D.R.C.*	1	70
Southern Africa	1	29
South Africa	1	29

OTHERS	Cohorts	N
Eastern Asia	10	1675
China	10	1675
South America	4	680
Brazil	4	680
South-Eastern Asia	2	652
Philippines	2	652

Fig 4. Number of cohorts and corresponding participants recruited by country. Country names and world macro-regions are based on the United Nations Statistics Division (UN data) classification [34]. Countries that account for more than 8% of the total IPD are highlighted in yellow. N: Number of participants recruited. * D.R.C.: Democratic Republic of the Congo.

doi:10.1371/journal.pntd.0004784.g004

mg/kg of body weight (the WHO-recommended regimen for preventive chemotherapy [3]), pooled data could help better characterize responses and identify subgroups with sub-optimal responses, and evaluate efficacy over time and in different settings (the 104 cohorts were recruited in 26 different countries (Fig 4) and in various contexts (Fig 5)). In addition, the

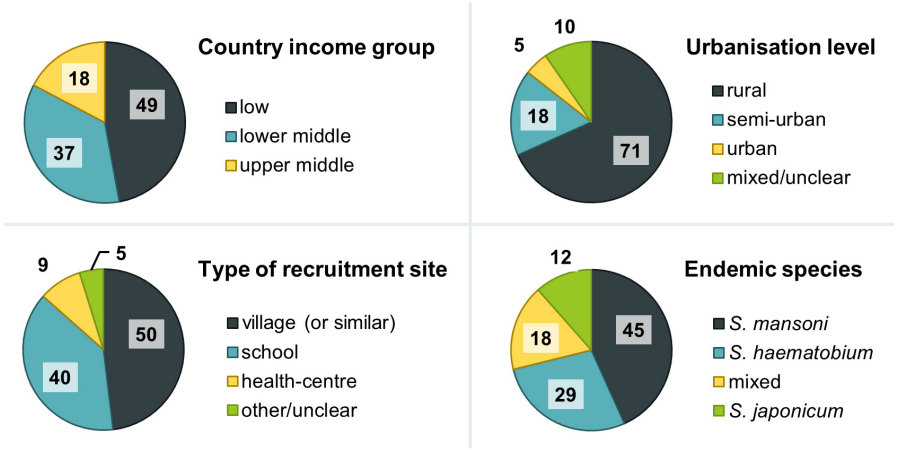


Fig 5. Variety of settings in which cohorts were recruited. Counts correspond to the number of cohorts (total per pie chart, n = 104) recruited at a site with the corresponding characteristic. Country income groups are based on World Bank classification [35].

doi:10.1371/journal.pntd.0004784.g005

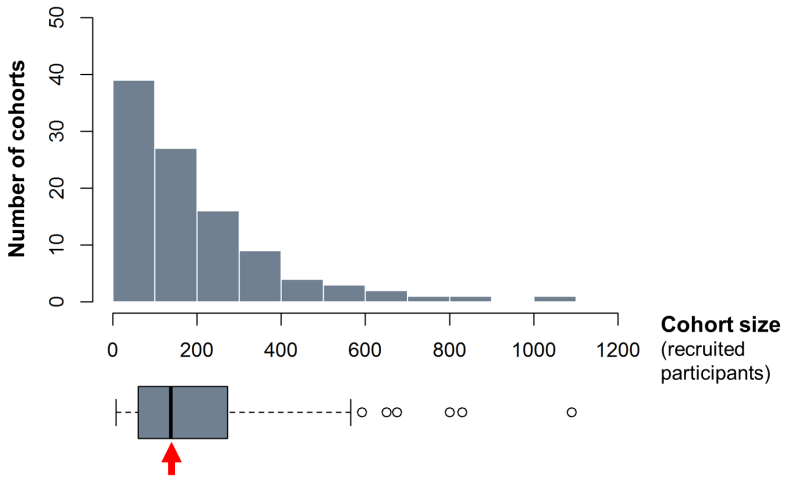


Fig 6. Cohort size distribution. 103 cohorts contribute to the histogram, as the number of recruited participants remains unknown for one cohort. The box covers values ranging between the first (60.5 participants) and third (273 participants) quartiles. The red arrow points to the median cohort size (138 participants). Whiskers extend to values up to 1.5 times the interquartile range; dots show outliers (up to a maximum of 1,090 participants).

doi:10.1371/journal.pntd.0004784.g006

20,517 IPD are likely to represent different sub-populations of interest and that remain understudied, such as preschool-aged children (30 of the 104 cohorts clearly recruited at least one participant aged strictly less than 6 years), or pregnant women (2 cohorts focussed on this group, i.e. a total of 452 participants and possibly more, as several cohorts seemed not to specifically exclude those women). Most cohorts were relatively small, with 62.5% of them involving less than 200 participants with schistosomiasis at baseline (Fig 6), thus the need to pool data together to reach the sample size giving enough statistical power to unravel drug effects in those smaller populations.

Heterogeneity in study design hampering meta-analyses of aggregated data

Not all of the studies included in the present review would meet more stringent eligibility criteria as required for inclusion in Cochrane systematic reviews [36]. No judgement on the type of study design was performed prior to inclusion. Potentially all good-quality drug efficacy data, even if not obtained from a randomised controlled trial (RCT), can contribute to IPD meta-analyses (comparison of continued drug efficacy in specific areas, modelling and model validation, trial methodology questions, etc.), and thus should be eligible for inclusion in a global database. Less than half (41) of the studies included here were comparative, and a third of them (30 studies) delivered a comparator intervention to the control arm (Fig 7). Of these comparative studies, 26 studies were randomised, out of which 12 mentioned a computer-generated randomisation sequence (Fig 7). Access to IPD would facilitate the appraisal of study design, for example by enabling to check the reliability of the randomisation method [9,10].

Owing to those loose eligibility criteria regarding study design, the studies included in the present review did not necessarily have the assessment of efficacy as their primary objective. Some rather focussed on specific immune responses following drug administration (15

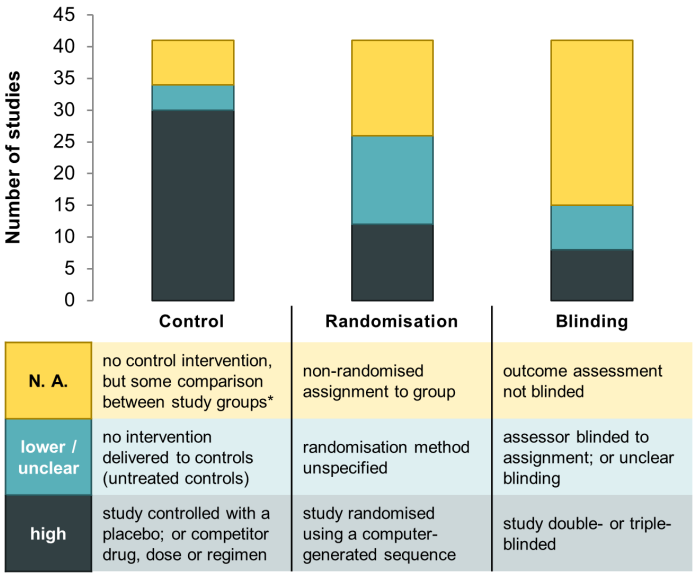


Fig 7. Characteristics of the study design among multiple-arm studies (n = 41), with quality assessment where applicable. Independently for each of the three design elements (control, randomisation and blinding), the quality as compared to expected 'gold standards' of treatment efficacy trials was estimated as 'high', 'lower/unclear', or 'N.A.' (not applicable), as detailed in the table accompanying the graph. * For example, comparison of response to the same treatment within different endemic settings.

doi:10.1371/journal.pntd.0004784.g007

studies), or on the search for specific biomarkers of infection and new approaches to diagnosis (14 studies). This means that not all studies are expected to report a drug efficacy outcome such as a CR or an ERR. The CR, or proportion of participants who were infected at baseline and negative at follow-up [37], was the most frequently reported measure of drug efficacy, with 66 of the 90 studies providing such information (Fig 8).

As the diagnosis of schistosomiasis is mainly performed by counting *Schistosoma* eggs in excreta (urine or stool for urinary and intestinal schistosomiasis, respectively), drug efficacy can also be assessed as ERR, or variation in the mean number of eggs found in the excreta of the population before and after treatment [26]. Studies however tend to compute the ERR in different manners, and in particular, to derive it from an arithmetic or a geometric mean (here 7 and 39 articles, respectively, Fig 8). Such variability in efficacy reporting is a major issue in meta-analysis of aggregated data, which could be solved by recalculating measures using the original IPD, and ideally, using the original raw egg count from each of several measurements done for each participant.

Other causes of heterogeneity

In all cohorts, the primary method to diagnose infection at screening was egg detection and counting in excreta. Eight cohorts also diagnosed participants by quantifying levels of schistosomiasis circulating anodic or cathodic antigen in body fluids (urine or serum). Of those 8 cohorts, 3 used only the latter circulating antigen method to assess the outcome after treatment. The most frequent method to diagnose infection with *Schistosoma spp*. is thus counting eggs; however, differences exist in how samples are prepared (different techniques may be used to

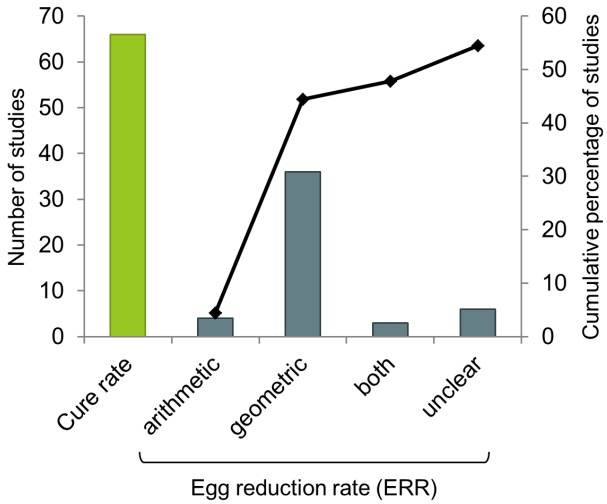


Fig 8. Outcome measure reported in published article or conference abstract. The histogram bars express the total of studies reporting on the primary outcome using the corresponding computation (CR or ERR expressed as arithmetic, geometric mean or both). The curve expresses the cumulative percentage of studies reporting an ERR. All studies reporting an ERR also reported a CR, except for one.

doi:10.1371/journal.pntd.0004784.g008

process samples prior to egg count under microscopy, especially for the preparation of stools, Fig 9), and how many samples are analysed (numbers of samples and/or test repeats, Fig 9). Since data are reported as means, and the sensitivity of the method depends on the number of independent egg counts for each subject, access to pooled IPD enables better to quantify the extent of this variation in sensitivity, and how it impacts on reported treatment outcomes [24].

The time at which outcome is assessed post-treatment also varies across studies (Fig 10): 48% of the cohorts (50 cohorts corresponding to 8,144 participants) have at least one data point at 3 to 4 weeks post-treatment, and 20 of those cohorts also have later assessments, sometimes much beyond the analysed 60-day limit. Such cohorts with data points at both 3–4 weeks and later provide opportunities for analyses distinguishing between initial cure and potential reinfection patterns.

Discussion

This review provides an overview of the landscape of schistosomiasis clinical research, focused on drug efficacy trials conducted over the past 15 years. We identify a total of approximately 20,000 participants enrolled in antischistosomal efficacy studies, covering the three main *Schistosoma* species, and a range of drugs and regimens, study participants, methods (diagnosis and treatment outcomes), countries and contexts. This supports the value of collating data within a global database, as the opportunities for analyses of clinical relevance would be plenty. At best, IPD from schistosomiasis studies should be shared at the level of the raw data, including for instance egg counts for each sample taken.

Limitations

We are cognisant that not all the data identified here may be available (e.g., electronic dataset non-existent, corrupted or lost), or be made available (some data generators might be reluctant

(a) Participants whose infection was diagnosed by performing x x y egg counts



(b) Primary technique to process stool samples

	S. mansoni	*S. japonicum*
Kato-Katz		
41.7, 42 or 42.7 mg	5045	717
50 mg	33	205
modified	215	
unspecified	3995	1260
Kato		
25 mg	154	
41.7 or 43.7 mg	104	
50 mg	155	
unspecified	892	
Katz		
unspecified	128	
Other		
concentration/hatching	9	145
unspecified	60	

Fig 9. Approach for the diagnosis of infection with *S. haematobium*, *S. mansoni* or *S. japonicum* by egg count. (a) For each species, participants were diagnosed by counting ova excreted in urine (*S. haematobium*) or stools (*S. mansoni* or *S. japonicum*). The diagnostic test was performed on *y* independent samples, and repeated *x* times on each sample. Values are numbers of participants. (b) While preparation of urine samples was standard (10 mL urine filtration and microscopy slide), techniques used to prepare stool samples varied. Values are numbers of participants whose stools were processed using the corresponding technique (total: 10,790 and 2,327 participants for *S. mansoni* and *S. japonicum*, respectively).

doi:10.1371/journal.pntd.0004784.g009

to share) [25,38], but also that more datasets might be available than those captured here. Those include national control programmes surveys (most of which are generally not published), and studies still on-going and/or awaiting publication. The delay between study completion and publication is sometimes very long, ranging from 5 months to 8 years in the case of the articles analysed here (S1 Fig); and only 12 of 90 studies were found to be registered in ClinicalTrials.gov or Controlled-Trials.com. In addition, several studies published locally in languages other than English could also have been missed. No restriction on language was used in the search, but regional registries such as LILACS (South American journals) [39] and CNKI (Chinese journals) [40] were not interrogated. Only 4 of the 90 included articles were in French, and none was in Portuguese or Chinese despite Brazil and China having reputed control programmes and productive research institutes.

Few to no data on oxamniquine (OXA) and metrifonate (MTF) could be identified in the post-2000 timeframe. A search of PubMed articles published in the 1986–2000 period identified at least 2000 and 500 more IPD for OXA and MTF, respectively. Though experience suggests that such 'older' data will be more difficult to retrieve [25], efforts to collect past data

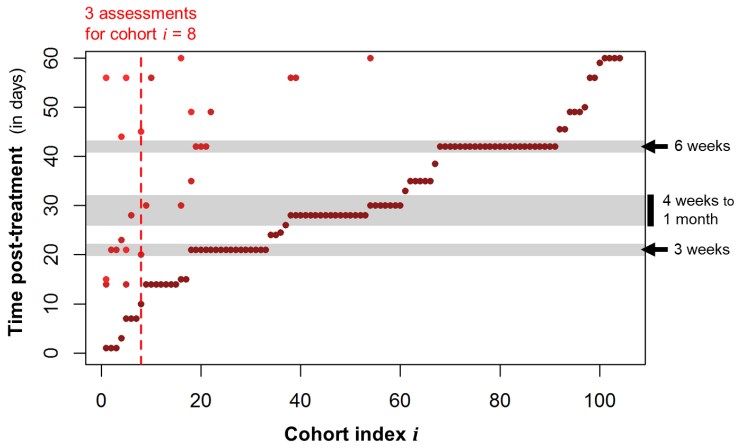


Fig 10. Time of outcome assessment after treatment. For each cohort (indexed from 1 to 104, in increasing order of first time of outcome assessment), all the time-points (within 60 days) at which an assessment of outcome occurred after treatment are represented by one dot. Darker dots: earlier time-point. Lighter dot: later time-point. As an illustration, the dotted line highlights cohort i = 8, for which 3 assessments of outcome were performed at Day 10, 20 and 45. The weekly distribution of the times of outcome assessment seem to depend on species (Fischer's Exact Test, p-value = 0.01835), with *S. haematobium* infection preferably re-assessed at later time-points.

doi:10.1371/journal.pntd.0004784.g010

from Brazil on OXA are under way and should be promoted, in order to provide a more complete picture of the efficacy of several antischistosomal drugs, and its variations over time. Similarly, the collection of data from China should yield additional data on *S. japonicum*, which is currently under-represented (2,327 IPD) as compared to *S. haematobium* (9,833 IPD) and *S. mansoni* (10,790 IPD).

Moreover, local variations in the preferred post-treatment time of outcome assessment may have led to the exclusion of yet relevant studies. When consulted on the development of an antischistosomal data-sharing platform, some investigators suggested to extend the boundary of acceptable time of first assessment to 3 months (90 days), instead of the 2 months (60 days) used as a cut-off for inclusion in the present review: 104 articles were excluded at screening on the grounds of the outcome assessment occurring too late, of which many had performed it at 3 months post-treatment. The preferred timeframe for study inclusion should also depend on the purpose of the subsequent meta-analysis, whether strictly restricted to drug efficacy assessment, or aiming at estimating long-term effects on infection intensity in the population.

Significance

Despite its limitations, the present review confirms and extends the problem of the variability in clinical research methods and reporting [11], which makes data summaries difficult to re-analyse and undermines the strength of evidence. Some sources of variability could be corrected with access to IPD and application of standardised analyses, as demonstrated by the example of efficacy outcomes: with IPD, it is possible to recalculate ERRs from distinct studies in a homogeneous manner using arithmetic means (as recommended by the WHO [26]), and to explore and compare alternative ways of expressing efficacy, as already done on an initial dataset of up to 4,740 individuals [24,41]. Other sources of variability will be more difficult to correct, but a global database would allow to pool data from studies with consistent diagnostic

approaches or, as shown in Olliaro *et al.* [24], to analyse the entire database using the WHO standard operating procedures (one single test of one single sample [26]).

Based on the present analysis and on our past experience of performing IPD meta-analyses [24,41], we find that the following variables would be most useful: site(s) and country of study, *Schistosoma* species under study, year of study start and stop, diagnostic method (number of samples and test repeats, technique of samples' preparation), regimen (drug name, dose, formulation and treatment regimen, if applicable) and time of follow-up(s). Dates of different events (recruitment, treatment and follow-up(s)) and associated outcomes (e.g. diagnostic result, adverse events) should also be collected for each study participant.

As a final consideration, although 20,000 (and possibly more) IPD is significant and justifies efforts for the creation of an IPD shared database, this amount of data for all the published schistosomiasis treatment trials of the past 15 years is disappointingly low, and reflects the lack of investments in clinical research in this aptly-named NTD.

Conclusion

The present scoping exercise supports the interest of an antischistosomal efficacy data-sharing platform, as more than 20,000 IPD could contribute to extensive meta-analyses, using a common analytical plan. This would improve our knowledge of how effective treatments are in the overall population as well as in subgroups (age, location, etc.) and over time; and of how to analyse data in a meaningful way by exploring and comparing different approaches. The need for such a database is further emphasised by the push from funders and journals to make research data more widely accessible and optimally used. With this in mind, schistosomiasis stakeholders, researchers and policymakers have been invited to discuss their views on this matter at a WHO-convened meeting held in September 2015 [15]. Providing that a clear data governance structure is defined, especially regarding how data which are shared will be accessed and used, and that measures are taken to address the need for data management and analysis capacity-building, there was general support to the establishment of such a database towards producing a stronger evidence-base for schistosomiasis treatment.

Supporting Information

S1 Checklist. PRISMA checklist for the systematic review of antischistosomal trials. The template PRISMA checklist was downloaded from the PRISMA statement website and used to guide the reporting of our work. For each applicable item, we specify the sub-section or paragraph of this article where the information can be found.
(PDF)

S1 Dataset. Source data supporting the systematic review of antischistosomal trials. The datasets (CSV files) contain the complete information extracted from the eligible articles or conference papers. An associated dictionary provides a detailed description of each dataset and each variable it comports. The scripts written to run the data analysis (using R 3.1.3) are also available.
(ZIP)

S1 Fig. Time elapsed between the end of data collection and the dissemination of results. Represented data is for a total of 70 cohorts, as the date of completion of data collection was unknown for 34 cohorts. Each bar of the histogram represents the number of cohorts for which study outcomes were made publicly available (article publication or presentation at conference) within the corresponding 6-month period after the end of the data (biological samples) collection for that cohort. The red arrow points to the median reporting time (30

months = 2.5 years). The box covers values ranging between the first (18 months) and third (50 months) quartiles, whiskers extend to values up to 1.5 times the interquartile range.
(TIF)

Acknowledgments

We wish to thank the Cochrane Infectious Diseases Group and in particular Vittoria Lutje for conducting the electronic search of the literature.

Author Contributions

Analyzed the data: AMJ. Wrote the paper: AMJ PLO. Conceived the systematic review protocol and data extraction form: AMJ PJG MV PLO TAL. Screened abstracts for eligibility: AMJ. Extracted information from eligible publications and collated the analysis datasets: AMJ.

References

1. Adams J, Gurney KA, Pendlebury D. Thomson Reuters Global Research Report: Neglected Tropical Diseases. Leeds; 2012.
2. Pedrique B, Strub-Wourgaft N, Some C, Olliaro P, Trouiller P, Ford N, et al. The drug and vaccine landscape for neglected diseases (2000–11): a systematic assessment. Lancet Glob Heal. 2013; 1: e371–e379. doi: 10.1016/S2214-109X(13)70078-0
3. World Health Organization. Preventive chemotherapy in human helminthiasis. Geneva: WHO Press; 2006.
4. World Health Organization. Schistosomiasis: Control Report 2001–2011 and Strategic Plan 2012–2020. WHO Press. Geneva; 2013.
5. Andrews P. A summary of the efficacy of praziquantel against schistosomes in animal experiments and notes on its mode of action. Arzneimittelforschung. 1981; 31: 538–41. PMID: 7016122
6. Gönnert R, Andrews P. Praziquantel, a new board-spectrum antischistosomal agent. Zeitschrift für Parasitenkd (Berlin, Ger. 1977; 52: 129–50.
7. Kramer CV, Zhang F, Sinclair D, Olliaro PL. Drugs for treating urinary schistosomiasis (review). The Cochrane Library. John Wiley & Sons, Ltd.; 2014. doi: 10.1002/14651858.CD000053.pub2
8. Danso-Appiah A, Olliaro PL, Donegan S, Sinclair D, Utzinger J. Drugs for treating Schistosoma mansoni infection (review). The Cochrane Library. John Wiley & Sons, Ltd.; 2013.
9. Stewart L a, Tierney JF. To IPD or not to IPD? Eval Health Prof. 2002; 25: 76–97. PMID: 11868447
10. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010; 340: c221. doi: 10.1136/bmj.c221 PMID: 20139215
11. Danso-Appiah A, Garner P, Olliaro PL, Utzinger J. Treatment of urinary schistosomiasis: methodological issues and research needs identified through a Cochrane systematic review. Parasitology. 2009; 136: 1837. doi: 10.1017/S0031182009005939 PMID: 19493363
12. World Health Organization. Guidelines For The Treatment of Malaria. WHO Press. Geneva; 2015.
13. Doenhoff MJ, Cioli D, Utzinger J. Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis. Curr Opin Infect Dis. 2008; 21: 659–667. doi: 10.1097/QCO.0b013e328318978f PMID: 18978535
14. World Health Organization. Schistosomiasis: number of people treated worldwide in 2013. Wkly Epidemiol Rec. 2015; 5: 25–32.
15. WHO, TDR. Schistosomiasis Data Platform Stakeholders' Meeting. Geneva; 2015.
16. Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risks. Washington DC: The National Academies Press; 2015.
17. Bill & Melinda Gates Foundation. Bill & Melinda Gates Foundation Open Access Policy [Internet]. 2015 [cited 6 Nov 2015]. Available: http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy
18. PLOS. Data Availability [Internet]. [cited 6 Nov 2015]. Available: http://journals.plos.org/plosone/s/data-availability
19. Research Councils UK. RCUK Common Principles on Data Policy [Internet]. 2014 [cited 6 Nov 2015]. Available: http://www.rcuk.ac.uk/research/datapolicy/

20. The BMJ. Open Data [Internet]. 2015 [cited 6 Nov 2015]. Available: http://www.bmj.com/open-data
21. Pisani E, Abou-Zahr C. Sharing health data: good intentions are not enough [Internet]. 2010 [cited 27 Nov 2015]. Available: http://www.who.int/bulletin/volumes/88/6/09-074393/en/
22. Moher D, Liberati A, Tetzlaff J, Altman DG, Grp P. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement (Reprinted from Annals of Internal Medicine). Phys Ther. 2009; 89: 873–880. doi: 10.1371/journal.pmed.1000097 PMID: 19723669
23. Zwang J, Olliaro PL. Clinical efficacy and tolerability of praziquantel for intestinal and urinary schistosomiasis-a meta-analysis of comparative and non-comparative clinical trials. PLoS Negl Trop Dis. 2014; 8: e3286. doi: 10.1371/journal.pntd.0003286 PMID: 25412105
24. Olliaro PL, Vaillant M, Diawara A, Coulibaly JT, Garba A, Keiser J, et al. Toward Measuring Schistosoma Response to Praziquantel Treatment with Appropriate Descriptors of Egg Excretion. PLoS Negl Trop Dis. 2015; 9: e0003821. doi: 10.1371/journal.pntd.0003821 PMID: 26086551
25. Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, et al. The availability of research data declines rapidly with article age. Curr Biol. 2014; 24: 94–97. doi: 10.1016/j.cub.2013.11.014 PMID: 24361065
26. World Health Organization. Assessing the Efficacy of Anthelminthic Drugs Against Schistosomiasis and Soil-Transmitted Helminthiases. Geneva: World Health Organization; 2013.
27. Loker ES. A comparative study of the life-histories of mammalian schistosomes. Parasitology. Cambridge University Press; 2009; 87: 343. doi: 10.1017/S0031182000052689
28. The R Foundation. The R Project for Statistical Computing [Internet]. [cited 27 Nov 2015]. Available: https://www.r-project.org/
29. Buowari YD. Clinical presentation and management of schistosomiasis at a hospital in a rural area in Nigeria. Niger J Med J Natl Assoc Resid Dr Niger. 2011; 20: 61–64.
30. Banwat ME, Ogbonna C, Daboer JC, Chingle MP, Envuladu EA, Audu S, et al. Prevalence of urinary schistosomiasis in school-aged children in Langai, Plateau State: pre- and post-intervention. Niger J Med. 2012; 21: 146–149. PMID: 23311180
31. Zinyowera S, Muchaneta-Kubara CE, Mutapi F, Midzi N, Ndlovu PD, Mduluza T. Changes in the humoral immune responses after chemotherapy in single and co-infected individuals with Schisosoma haematobium and Plasmodium falciparum. Cent Afr J Med. 2006; 52: 104–111. PMID: 20353134
32. García N, Isturiz G, Aular S, Incani RN. The efficacy of human schistosomicide treatment may depend on the rate of transmission. Parasitol Res. 2006; 98: 545–9. doi: 10.1007/s00436-005-0047-1 PMID: 16418873
33. Zeng QR, Hou JW, He YK, Luo XS, Zhang SK, Shu HP, et al. Analysis on morbidity and chemotherapy effects of Schistosoma japonicum infection in fishermen on Dongting Lake (Chinese). Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi. 2004; 22: 199–203. PMID: 15587151
34. United Nations Statistics Division. Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings [Internet]. 2013 [cited 9 Nov 2015]. Available: http://unstats.un.org/unsd/methods/m49/m49regin.htm
35. The World Bank. Countries and Economies [Internet]. 2015 [cited 9 Nov 2015]. Available: http://data.worldbank.org/country
36. Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration. John Wiley & Sons, Ltd.; 2008. doi: 10.1002/9780470712184
37. Montresor A. Cure rate is not a valid indicator for assessing drug efficacy and impact of preventive chemotherapy interventions against schistosomiasis and soil-transmitted helminthiasis. Trans R Soc Trop Med Hyg. 2011; 105: 361–363. doi: 10.1016/j.trstmh.2011.04.003 PMID: 21612808
38. Swan A, Brown S. To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. 2008;
39. BIREME, PAHO/WHO. LILACS [Internet]. 2015 [cited 9 Nov 2015]. Available: http://lilacs.bvsalud.org/en/
40. Tsinghua Tongfang Knowledge Network Technology Co. Ltd. China Knowledge Resource Integrated Database [Internet]. 2014 [cited 9 Nov 2015]. Available: http://oversea.cnki.net/
41. Walker M, Mabud TS, Olliaro PL, Coulibaly JT, King CH, Raso G, et al. New approaches to measuring anthelminthic drug efficacy: Parasitological responses of childhood schistosome infections to treatment with praziquantel. Parasites and Vectors. Parasites & Vectors; 2016; 1–15. doi: 10.1186/s13071-016-1312-0

D.2 Second publication: TDR Global Competency Framework for Clinical Research

The following article [276] is reproduced from BMJ Global Health (creative commons attribution non-commercial license). The reported work was commissioned by TDR, and although not directly contributing to my thesis, it greatly informed my thinking on clinical research data management capacity strengthening, and first pointed me to the gap in a clearly defined role for clinical research data managers. Together with colleagues from The Global Health Network, I drove the development of the TDR Global Competency Framework for Clinical Research, which we presented in a draft version at a TDR workshop in September 2015. Following input from the clinical research experts in attendance, an extent TDR report was published [274, 275], and a set of beta online tools was released on `tghn.org`.

Analysis

BMJ Global Health

Developing a globally applicable evidence-informed competency framework to support capacity strengthening in clinical research

Amélie Julé,[1] Tamzin Furtado,[1] Liam Boggs,[1] Francois van Loggerenberg,[1] Victoria Ewing,[1] Manhaz Vahedi,[2] Pascal Launois,[2] Trudie Lang[1]

To cite: Julé A, Furtado T, Boggs L, *et al*. Developing a globally applicable evidence-informed competency framework to support capacity strengthening in clinical research. *BMJ Global Health* 2017;**2**:e000229. doi:10.1136/bmjgh-2016-000229

AJ and TF contributed equally.

Received 31 October 2016
Revised 10 February 2017
Accepted 16 February 2017



[1]The Global Health Network, Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK
[2]Special Programme for Research and Training in Tropical Diseases (TDR), Research capacity Strengthening and Knowledge Management (RCS-KM), World Health Organization, Geneva, Switzerland

Correspondence to
Tamzin Furtado;
tamzin@globalhealthtrials.org

ABSTRACT

Capacity development for clinical research is held back by a lack of recognition for the skills acquired through involvement in clinical trials and in other varied types of global health research studies. Although some competency frameworks and associated recognised career pathways exist for different clinical research roles, they mostly apply to a single role or study setting. Our experience supports the need for an integrated approach, looking at the many roles in parallel and at all types of clinical research beyond trials. Here, we propose a single, flexible framework which is applicable to the full global health research team, and can be used for recognising staff by highlighting acquired skills and possible progression between various roles. It can also illuminate where capacity needs strengthening and contribute to raising research engagement. Through systematic analysis of existing competency frameworks and current job descriptions covering 11 distinct, broad clinical research roles, we identified and defined 50 key competencies required by the team as a whole and throughout the study life cycle. The competencies are relevant and adaptable to studies that differ in design, geographical location or disease, and fall in five main areas—(1) Ethics, Quality and Risk Management; (2) Study and Site Management; (3) Research Operations; (4) Scientific Thinking; and (5) Professional Skills. A pilot framework and implementation tools are now available online and in paper format. They have the potential to be a new mechanism for enabling research skills development and career progression for all staff engaged in clinical research globally.

Key questions

What is already known about this topic?

- Regardless of the size, place and type of clinical research, for a project to be realised, certain steps and tasks remain the same, and must be achieved to a high standard by the clinical research team.
- Those involved in research require adequate, targeted training to acquire the necessary competencies and conduct their role accordingly.
- Competency frameworks help to identify the knowledge, skills and tasks that clinical research staff need to master, but are usually narrow in their focus.

What are the new findings?

- This study combines existing clinical research frameworks with real-life data to produce a globally applicable framework which is adaptable for different types of research, different global settings, team sizes and disease types.
- Detailed analysis of the desired traits for different clinical research roles highlights that competencies cut across roles.
- The novelty of this work lies in its broad applicability to the research team as a whole, thus highlighting continuity between roles and seeking to recognise all contributions to the clinical research endeavour.

Recommendations for policy

- This framework can be used in multiple ways, to develop curricula or inform career development for individuals working in clinical research, or to identify gaps in existing research capacity.
- These applications will assist in the delivery of clinical research, while increasing the recognition of all staff for their contribution.
- Career pathways need to be developed to enable all staff to progress in their role as well as in others, in light of their demonstrated competencies rather than their job title.

INTRODUCTION

In its 2013 World Health Report (WHR),[1] WHO urged the global research and policy communities to take integrated efforts to strengthen the capacity of all countries to generate evidence for the improvement of their health systems. The need for action towards health research capacity development had already been taken up at the regional level by the Pan American Health Organization (PAHO), which first issued a policy document on the matter in 2009.[2] These two documents particularly emphasise

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

BMJ Global Health

the need to develop research capacity in low and middle income countries (LMICs), so that these countries can take ownership in addressing local health issues. Research agendas and practices which are set in high-income settings indeed often fail to adequately target the disease burdens that affect the majority of the world's populations.[3 4] To make real changes to global health outcomes, LMICs must therefore equip themselves with the skills to carry out high-quality research studies that address locally relevant research questions.[5]

The capacity needed to conduct global health studies is multifaceted, and ranges from the funding and physical facilities (ward space, laboratories), through to managerial support and personnel training.[5] While all building blocks of a well-functioning research system are necessary and interlinked, arguably the key to developing sustainable local capacity comes from having well-trained teams, and individuals who can demonstrate leadership in research, whatever their role. This concept of a capable workforce is central to many capacity development models, which stress the importance of training individuals and ensuring that they are situated within environments of organisational support, mentoring and supervision.[2 6 7]

Although on the rise, the availability and accessibility of adequate training programmes remain limited for many research staff;[8 9] and particularly for those who are not qualified as medical doctors.[9] Technological developments now make it possible to reach new populations through various models of online training.[10–13] However, with a lack of understanding of the clinical research skills that need to be developed, it is not clear whether existing 'clinical trials' curricula (whether face-to-face, or online and distance learning) cover the full range of topics to be mastered by global health research staff. In addition, many core roles within clinical research are little understood and still relatively unrecognised as viable career paths.[9 14] This is the feedback we received during discussions with local partners and when we evaluated training exercises we regularly conduct in LMICs, especially on the role of research nurses, trial managers and data staff.[8 15]

Furthermore, sustainable capacity building must enable research groups to independently plan, obtain funding for and operate high-quality, locally led studies that address most pressing issues. Tackling those require a whole range of clinical research; from epidemiology, to observational studies and cross-sectional surveys, through to interventional trials. When defining a set of competencies for clinical research, we therefore consider that an integrative approach which aims to capture information about a spectrum of clinical research will be most successful and practicable, as it will best help staff determine how to transfer their advanced research skills to different types of project.

Several groups have worked towards clarifying the competencies for specific clinical research roles, and the 'Multi Regional Clinical Trials Center (MRCT) Harmonized Core Competency Framework'[16–18] makes an excellent move towards combining the information about several roles (principal investigator, clinical research coordinator and clinical research associate) into one common set of knowledge domains. Here, we further this integration effort to many more roles, including essential though less recognised ones in more operational or technical positions (eg, laboratory technician, data entry clerk). We also take particular account of LMIC settings.

CREATING THE FRAMEWORK

To create the *Special Programme for Research and Training in Tropical Diseases (TDR) Global Competency Framework for Clinical Research*, we looked at the complete set of activities needed for a study to take place, regardless of which individual may perform them. We then examined this information independently of the initial job role it pertained to, to determine the shared competencies (tasks, skills and knowledge) required to support the research endeavour as a whole. The rationale for this approach is that each research project is unique, and the exact distribution of roles and responsibilities within the team are likely to vary depending on the size and demands of the study, as well as on the resources and team available to conduct the trial on the ground. Yet, in any clinical research project there are core, invariant aspects that need to be covered: a protocol must be written, ethics approval must be obtained, data must be collected, verified and analysed, etc. Therefore, the research process and the team undertaking it can be envisioned as a continuum for the definition of cross-cutting 'competency areas'.

This systematic examination and categorisation of clinical research activities involved qualitative content analysis of 28 past competency frameworks and 116 current job descriptions from partners around the world, altogether covering the following 11 broad research roles: data staff (from entry to managerial roles), laboratory scientist (from technician to head of laboratory), trial pharmacist, community engagement staff, research nurse, study physician, principal investigator, trial manager or project coordinator, quality control monitor, ethics review board member and sponsor. The data-driven, evidence-based framework resulting from this work was then appraised by a panel of 20 expert trialists, and accordingly refined—detailed methods are provided in the *TDR Global Competency Framework for Clinical Research* development report.[19] In the remainder of this article, we describe the framework (accessible at https://globalhealthtrials.tghn.org/global-competency-framework-clinical-research/), as well as its associated implementation tools; we share the lessons learnt while developing this global competency framework, and we highlight its potential as a mechanism to strengthen clinical research capacity.

CLINICAL RESEARCH ACTIVITIES REVOLVE AROUND 50 COMPETENCIES

After several rounds of iterative, thematic categorisation of the textual data that we gathered from multiple

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

sources and partners around the world, we found that information from the records included in the analysis crystallised to around 50 distinct competencies, which can be grouped into five balanced categories (figure 1): Ethics, Quality and Risk Management; Study and Site(s) Management; Research Operations; Scientific Thinking; and Professional Skills. The 'Ethics, Quality and Risk Management' area encompasses all the competencies related to putting safeguards into place, so as to ensure the risks to participants and of not reaching a scientific conclusion from the research work are kept as low as possible. The 'Study and Site(s) Management' area focuses on research management, from project coordination through to finances and human resources management, and also includes the ability to communicate to relevant parties; an essential part of large, multinational clinical research. The 'Research Operations' area contains more action-based competencies, and covers most of the practical requirements of interacting with participants, delivering interventions or analysing biological samples to collect data on a daily basis during the study. The 'Scientific Thinking' area concentrates on the planning and interpretation aspects of research, which may require various levels of competencies in key disciplines (eg, medicine, statistics, pharmacology, public health), as well as an understanding of the 'science' of doing research itself and the methods that are applicable to different research questions. Finally, the 'Professional Skills' are generalisable, essential competencies, such as leadership, interpersonal skills or organisational skills.

While the day-to-day tasks and behaviours suggesting competency in the corresponding area will differ between team members and vary according to their specific role, the thorough examination of employers' requirements alongside relevant literature suggests that the underlying competencies are remarkably similar and cross-cutting.

A competency is defined as the knowledge or skill required to carry an activity out, not the activity itself.[20 21] Our framework distinguishes between the following three types of competencies: (1) the theory, or knowledge-based competency, which reflects a level of theoretical understanding of a particular topic, and which can be acquired through learning about a field; (2) the capability, or task-based competency, which necessitates the application of technical or practical knowledge to the performance of a specific task, and which is better learnt by doing and (3) the trait, or skill-based competency, which corresponds to the demonstration of appropriate behaviours in various situations, and which is better developed by experiencing and reflecting on the experience.

SIMILARITY BETWEEN COMPETENCIES ALLOWS FOR A UNIFYING FRAMEWORK

The finalised *TDR Global Competency Framework for Clinical Research* consists primarily of an encompassing 'Competency Wheel' (figure 1), which is applicable to all research roles. It visually presents the framework with its 50 competencies, distributed into the five categories. The core 'Professional Skills' category is surrounded by the four thematic, clinical research-specific categories (Ethics, Quality and Risk Management; Study and Site(s) Management; Research Operations; Scientific Thinking); and the latter are each further divided into 3–4 areas of competencies. While figure 1 provides a paper version of the framework, an interactive version allows navigating it more easily and is available online (https://globalhealthtrials.tghn.org/competencywheel/).

A 'Competency Dictionary' has been created as a direct companion to the Competency Wheel, and provides detail on what is meant by each of the 50 competencies (figure 2). The Competency Dictionary thus explains the overall, generic framework by suggesting a definition, further clarified with keywords, for each competency. Essential skills and knowledge required to master the competency are also listed along with the competencies that are related to it and that appear elsewhere in the framework. Usual abilities associated with the competency are given and directly derived from analysed documents; without those illustrative examples being job specific. The full Competency Dictionary is downloadable at https://globalhealthtrials.tghn.org/site_media/media/medialibrary/2016/11/TDR_Framework_Competency_Dictionary.pdf.

THE FRAMEWORK CAN FACILITATE TRAINING AND CAREER DEVELOPMENT

The core, unifying wheel and the dictionary are supported by practical implementation tools, particularly meant to support assessment and follow-up of an individual's competencies.

Drawing on previously developed guidance on scoring and evaluating one's skill level,[22] a 'Competency Radar' (figure 3) has been developed, and can be used to grade individuals on each of the areas of competency defined within the TDR Global Competency Framework for Clinical Research. Consistent grading should be used as much as possible, in line with the scale developed for the Professional Membership Scheme (PMS),[22] and to facilitate comparisons of competency levels, over time and between individuals. In the future, this consistent grading process will be further facilitated through the release of role-specific frameworks mapped to the generic one, to provide more detailed, targeted examples of tasks and expected levels for each role and competency.

Once filled in, the Competency Radar (figure 3) readily shows the areas of strengths of the concerned individual based on their past experience, and outlines which of their skills could be transferred to a different study or setting. It is indeed important to consider one role's activities within the global picture of competencies, so as to highlight the significance of their

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

BMJ Global Health

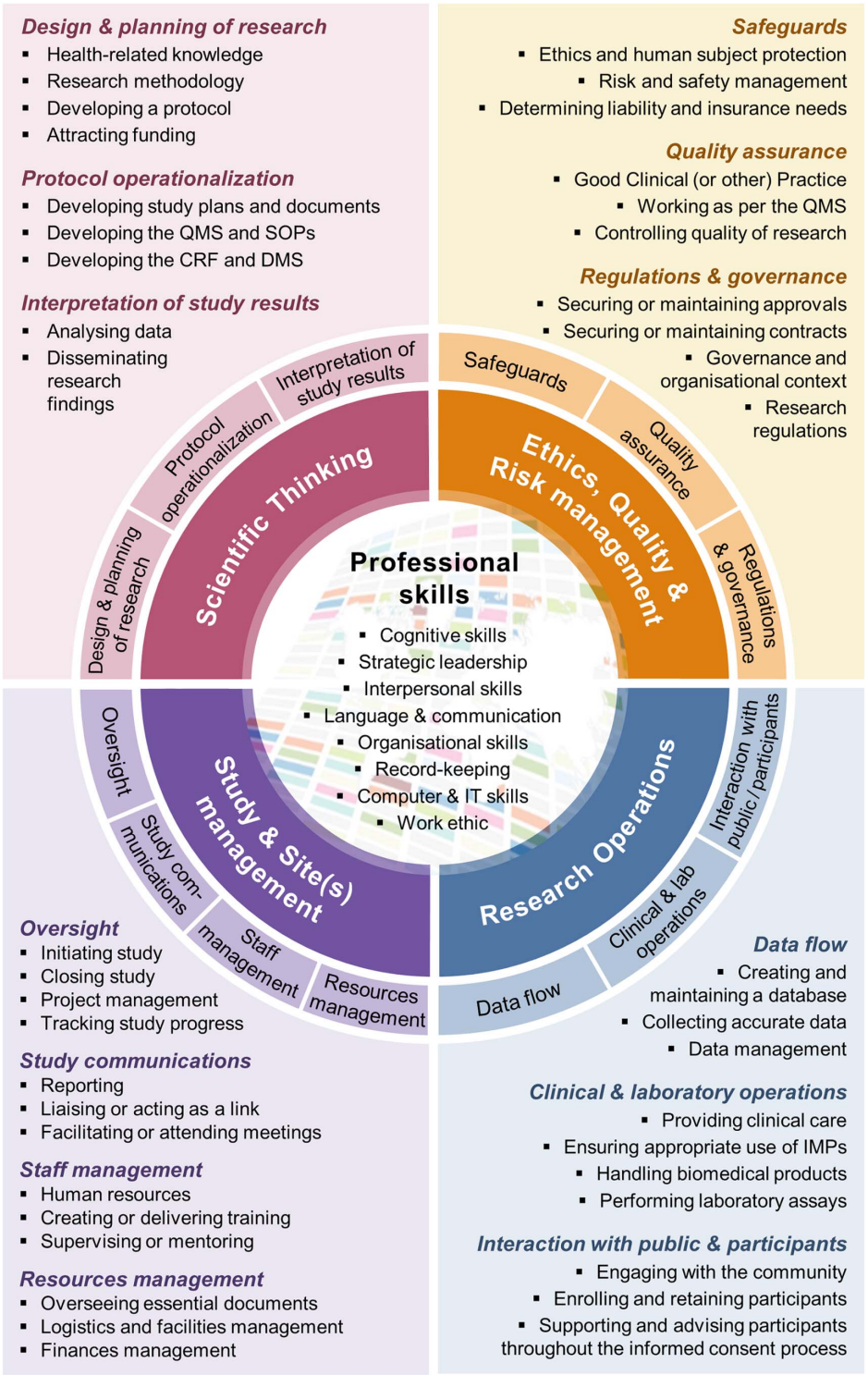


Figure 1 The Competency Wheel or one-page overview of the framework. All 50 competencies are visible in this overall representation of the framework. CRF, Case Record Form; DMS, Data Management System; IMPs, Investigational Medical Products; IT, Information Technology (computing); QMS, Quality Management System; SOPs, Standard Operating Procedures.

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

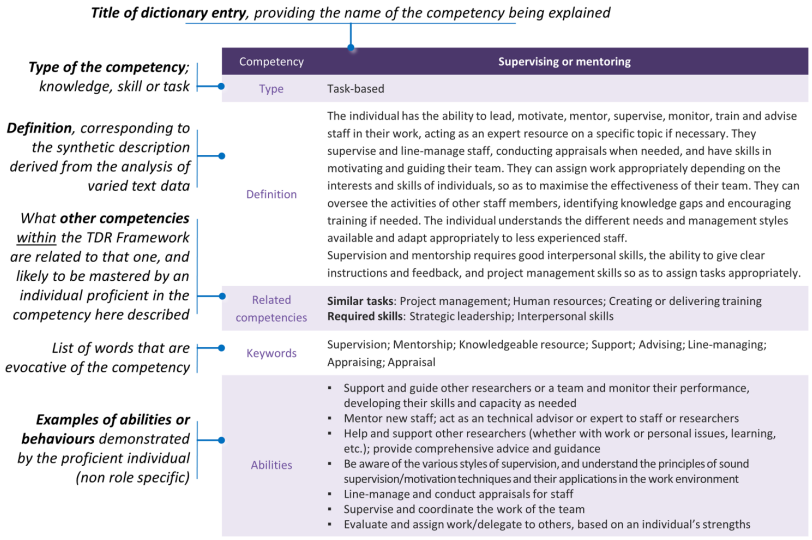


Figure 2 Annotated sample 'Competency Dictionary' entry. The 'Supervising or mentoring' competency from the Study and Site(s) Management area of competency is used as an example to illustrate how competencies have been defined. Each of the 50 competencies listed above have its specific entry in the companion dictionary, which can be downloaded at https://globalhealthtrials.tghn.org/site_media/media/medialibrary/2016/11/TDR_Framework_Competency_Dictionary.pdf.

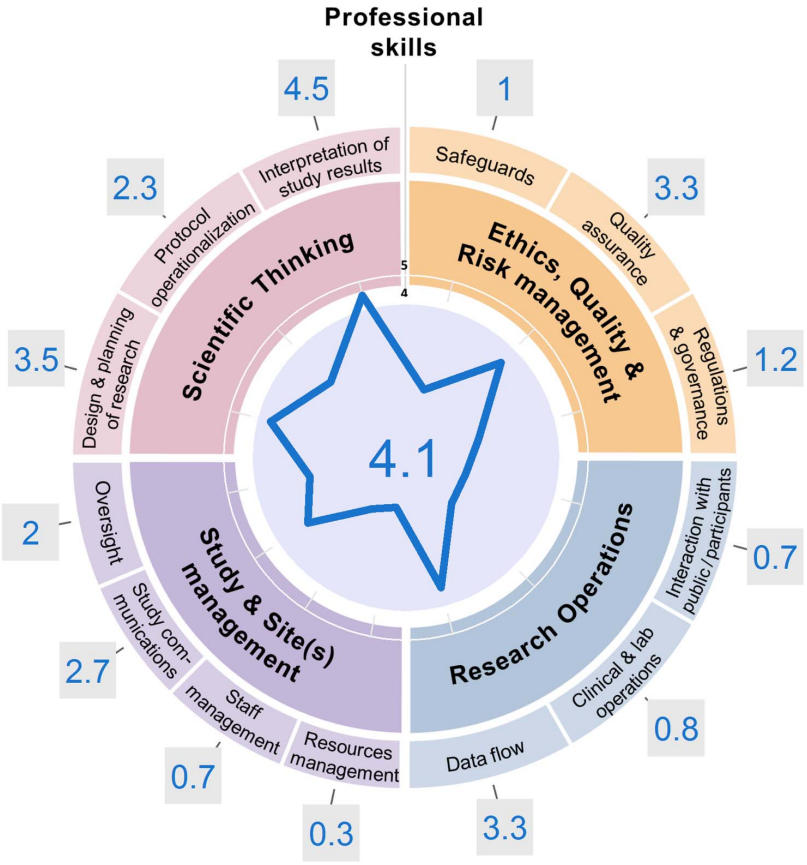


Figure 3 Example of completed 'Competency Radar'. The grading scheme accompanying the Special Programme for Research and Training in Tropical Diseases (TDR) Global Competency Framework was carefully applied to assess the competency level of a junior researcher in global health. The radar shape readily highlights areas of strengths for this young investigator (who has been involved in designing, planning and interpreting study results), as opposed to major areas where further training and experience should be sought, in line with career plans (eg, the present researcher still lacks confidence in the performance of managerial tasks). The image was generated using the online App for the Competency Radar.

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

contribution to the study, and their possible career pathways. Past experience of TDR and The Global Health Network suggests there is demand for such a flexible and endorsed system, so as to assist individuals with their career progression. The PMS[22] on which the present grading scheme is derived was initially built by The Global Health Network and TDR as a means of providing a free, standardised system for trialists of all roles to track and demonstrate their progression as they build their careers in research; and it currently counts 2500 members as of February 2017.

As a complementary supporting tool, a 'Piloting Protocol' (available at https://globalhealthtrials.tghn.org/site_media/media/medialibrary/2016/11/TDR_Framework_User_Guide.pdf) has been developed to provide further guidance about how to pilot test and use the framework in practice. This protocol also contains a feedback questionnaire, which the early adopters are encouraged to use to let the authors know of their experience using this beta version.

FURTHER DEVELOPMENT TO REINFORCE THE FRAMEWORK

While the framework is designed to evolve as we learn, it is currently being released in a beta version, a necessary step to facilitate its further development and testing. Continuous evaluation is indeed envisioned, so as to iteratively improve and validate the work and tools in line with users' needs. This learning process should offset the limitations discussed below.

First, job descriptions have been used as an easily accessible means of determining the roles, tasks and ideal attributes of applicants. These provide a list of the activities each individual undertakes, the training basis required for that role, and which key skills or personal attributes are desirable or essential. However, we recognise there could be a discrepancy in what a job description describes and the actual day-to-day tasks undertaken by an individual. This problem is likely to have limited impact on the overall framework's picture, but we recommend additional research and different methods to further the development of role-specific frameworks. This will be particularly needed for roles where we had limited access to job descriptions, such as the role of the study sponsor, which is critical to clarify to enable locally led research in LMICs.

Second, our framework aims to cover all types of clinical research and thus job descriptions from many diverse locations and trial types were taken into consideration. While this provides the widest possible range of competencies and highlights the connections between apparently different roles, it also means that some of the tasks and roles will not be applicable to every situation. Similarly, we strived to derive broadly applicable and adaptable competency definitions; these may, however, appear too generic in some cases. The generic picture of competencies is nonetheless beneficial to promote career development, as it enables to emphasise bridges between roles and to assess the skills and experience research staff have acquired in a comparable manner. The adoption of this whole-encompassing view is anticipated to lead to better career recognition and capacity building, provided that there is institutional buy-in with creation of adequate training and job opportunities on the ground.

In line with Lucia and Lepsinger's recommendations for the validation of competency frameworks,[21] the pilot testing will seek to: (1) show face validity, by ensuring that individuals who are currently working in each role understand the competencies suggested, and how they apply to their role; (2) determine whether key tasks have been identified and are clearly covered within one competency or another; (3) determine what the 'predictors of success' are for each role, that is, which behaviour demonstrates that the individual has mastered a competency.

Data from this testing phase will first serve to clarify how users interpret the proposed definitions, to address the face-validity question. A feedback questionnaire has been developed, allowing for in-depth comments on the easiness of understanding of different competencies. Where necessary, changes will be made to the framework or the wording of competencies and/or their definition, to improve the comprehensiveness of the framework.

Data from different users of the framework will also enlighten the key skills and tasks that all job roles require for the successful conduct of a study, including but beyond the investigator. We therefore encourage research teams to adopt and pilot test this framework to capture whether competencies are present or missing at the individual and at the larger, site level. Research teams should benefit from using the framework, as their progress can be tracked and demonstrated over time using the proposed tools; hence their efforts can be better directed towards addressing most pressing training or recruitment needs.

CONCLUSION

The Global Health Network and TDR have created a unifying framework which makes use of pre-existing, more role-specific or setting-specific frameworks, as well as of real-life data. The TDR Global Competency Framework for clinical research is broadly applicable and has been successfully appraised by multiple experts; in addition, stakeholders based in LMICs have particularly emphasised the need for such a scheme. In light of this support, the TDR framework has been launched along with its tools to enable refining this resource in the future, based on users' experience of applying it in context. This framework is laying the ground for future work in the field of clinical research capacity strengthening; and we encourage teams to get involved and use this scheme. Their feedback will be very valuable in refining and widely implementing this valuable new tool for research capacity development.

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

Handling editor Seye Abimbola.

Acknowledgements The authors would like to thank the following individuals for their reviews and feedback on the framework during the workshop on the development of a global core competency framework for clinical research held at the WHO head quarter, Geneva, Switzerland on 29–30 September 2015: Barbara E. Bierer, Christian Burri, Núria Casamitjana, C. Padma Chandrasekaran, Roma Chilengi, Helen Demarest, Michäel Kaeser, David Lalloo, Kamal Mansinho, Renata Mendizabal Sole de Cabrera, Wilfried Mutombo Kalonji, Jean Nachega, Raffaella Ravinetto, Morven Roberts, Nidia Rizzo, Nandi Siegfried, Nathalie Strub Wourgaft, Alfred Tiono, Jeremy Whitty. The authors would also like to thank Elizabeth Allen and Cordelia Reddy of the University of Cape Town, and Barbara Farrell of the UK Trial Managers Network, for their invaluable support and feedback in the initial stages of development. We also thank Alexander Mietke, Martin Hadley and Alexis Sergt for support with developing the online version of the framework.

Contributors All authors contributed to designing the study. TF and AJ collected and analysed the data with particular support and comments from VE and FvL. AJ created the design of the framework and its online App. TF and AJ wrote the first draft. FvL, VE, TL, LB, MV and PL reviewed and approved the manuscript. TL is the guarantor of the work.

Funding The Special Programme for Research and Training in Tropical Diseases (TDR).

Competing interests All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare: financial support for the submitted work from TDR; The Global Health Network is funded by the Bill and Melinda Gates Foundation; AJ holds a Nuffield Department of Medicine (NDM, University of Oxford) Prize Studentship, jointly covered by the Medical Research Council (UK) and NDM; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The study brought together existing data, mainly publically available data, which are referenced in the online reports, and from data obtained upon request and subject from a number of different sources. Full details how these data were obtained are available in the documentation available at https://globalhealthtrials.tghn.org/articles/tdr-global-competency-framework-clinical-research-set-tools-help-develop-clinical-researchers/.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES

1. World Health Organization. *The World Health Report 2013: Research for universal health coverage*. Luxembourg: WHO Press, 2013.
2. Pan American Health Organization/World Health Organization (PAHO/WHO). PAHO's Policy on Research for Health: 49th Directing Council, 61st Session of the Regional Committee (CD49/10). 10 July 2009. http://www2.paho.org/hq/index.php?option=com_content&view=article&id=1414%3A2009-policy-on-research-health&catid=2502%.
3. Yozwiak NL, Happi CT, Grant DS, *et al*. Roots, not parachutes: research collaborations combat outbreaks. *Cell* 2016;166:5–8.
4. Miranda JJ, Zaman MJ. Exporting "failure": why research from rich countries may not benefit the developing world. *Rev saúde pública* 2010;44:185–9.
5. Franzen SRP, Chandler C, Lang T. Health research capacity development in low and middle income countries: reality or rhetoric? A systematic meta-narrative review of the qualitative literature. *BMJ Open* 2017;7:e012332.
6. Potter C, Brough R. *Systemic capacity building: a hierarchy of needs. Heal Policy Plan*. Vol. 19. Cardiff, UK: Department of Epidemiology, Statistics and Community Medicine, University of Wales College of Medicine, 2004:336–45.
7. ESSENCE on Health Research. *Seven principles for strengthening research capacity in low- and middle-income countries: simple ideas in a complex world*. World Health Organization, 2014.
8. Franzen SRP, Chandler C, Enquselassie F, *et al*. Understanding the investigators: a qualitative study investigating the barriers and enablers to the implementation of local investigator-initiated clinical trials in Ethiopia. *BMJ Open* 2013;3:e003616.
9. Lang TA, White NJ, Tran HT, *et al*. Clinical research in resource-limited settings: enhancing research capacity and working together to make trials less complicated. *PLoS Negl Trop Dis* 2010;4:4–7.
10. The Global Health Network. The Global Health Network: enabling research by sharing knowledge. 2015 (cited 7 August 2015). https://tghn.org/
11. Coursera Inc. Coursera. 2015. https://www.coursera.org/.
12. edX Inc. edX [Internet] (cited 15 January 2015). https://www.edx.org/
13. London School of Hygiene and Tropical Medicine. Masters in Clinical Trials (Distance Learning) (cited 15 January 2015). http://www.lshtm.ac.uk/study/masters/dmsct.html
14. Ajay S, Bhatt A. Training needs of clinical research associates. *Perspect Clin Res* Clinical Operations, Feasibility and Site ID, Quintiles India, 5th Floor, Leela Business Park, M. V. Road, Andheri (E), Mumbai—400 059, India. 2010;1:134–8.
15. Furtado T, Franzen S, van Loggerenberg F, *et al*. Strengthening neglected tropical disease research through enhancing research-site capacity: an evaluation of a novel web application to facilitate research collaborations. *PLoS Negl Trop Dis* 2014;8:e3225.
16. Multi-Regional Clinical Trials Initiative. MRCT Resources. 2016 (cited 10 January 2016). http://mrctcenter.org/resources/
17. Thomas Jones C, Browning S, Gladson B, *et al*. Defining competencies in clinical research: Issues in clinical research education. *Res Pract* 2012;3:99–107.
18. Sonstein SA, Seltzer J, Li R, *et al*. Moving From Compliance to Competency: A Harmonized Core Competency Framework for the Clinical Research Professional. *Clin Res* 2014:17–23.
19. WHO/TDR. *Development of the TDR global competency framework for clinical research*. Geneva: World Health Organization, 2016.
20. Royal College of Nursing Research Society. Research Nurse Competency Framework—version 2. 2011. https://www2.rcn.org.uk/__data/assets/pdf_file/0019/201466/Research_Nurse_Competency_Framework_-_Version_2_-_Full_-_Oct_2011.pdf.
21. Lucia AD, Lepsinger R. *The art and science of competency models: pinpointing critical success factors in organizations*. San Francisco: John Wiley & Sons, Inc., 1999.
22. TDR, The Global Health Network. Global Health Network Professional Membership Scheme. 2015 (cited 15 January 2015). https://globalhealthtrainingcentre.tghn.org/cpd/

BMJ Glob Health: first published as 10.1136/bmjgh-2016-000229 on 30 March 2017. Downloaded from http://gh.bmj.com/ on 31 May 2018 by guest. Protected by copyright.

Bibliography

[1] A. Gertel, "CDISC Clinical Research Glossary," *Applied Clinical Trials (Special resource issue)*, pp. 11–43, Dec 2011. Available at: `https://www.cdisc.org/system/files/all/standard_category/application/pdf/act1211_011_043_gr_glossary.pdf` [Accessed 30 May 2018].

[2] E. Pisani, J. Whitworth, B. Zaba, and C. Abou-Zahr, "Time for fair trade in research data," *The Lancet*, vol. 375, pp. 703–5, Feb 2010.

[3] K. Ashley, "Developing Skills for Managing Research Data and Software in Open Research," London: Wellcome Trust, 2016. Available at: `https://dx.doi.org/10.6084/m9.figshare.4133916` [Accessed 29 May 2018].

[4] L. M. Federer, Y.-L. Lu, and D. J. Joubert, "Data literacy training needs of biomedical researchers," *Journal of the Medical Library Association*, vol. 104, pp. 52–57, Jan 2016.

[5] S. Bull, N. Roberts, and M. Parker, "Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research," *Journal of Empirical Research on Human Research Ethics*, vol. 10, pp. 225–238, Jul 2015.

[6] B. Fecher, S. Friesike, and M. Hebing, "What drives academic data sharing?," *PLOS ONE*, vol. 10, p. e0118053, Feb 2015.

[7] L. M. Federer, Y.-L. Lu, D. J. Joubert, J. Welsh, and B. Brandys, "Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff," *PLOS ONE*, vol. 10, p. e0129506, Jun 2015.

[8] Institute of Medicine, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, D.C.: The National Academies Press, 2015.

[9] P. Vallance, A. Freeman, and M. Stewart, "Data Sharing as Part of the Normal Scientific Process: A View from the Pharmaceutical Industry," *PLoS Medicine*, vol. 13, p. e1001936, Jan 2016.

[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo,

R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, Mar 2016.

[11] International Consortium of Investigators for Fairness in Trial Data Sharing, "Toward Fairness in Data Sharing," *New England Journal of Medicine*, vol. 375, pp. 405–407, Aug 2016.

[12] B. Lo and D. L. DeMets, "Incentives for Clinical Trialists to Share Data," *New England Journal of Medicine*, vol. 375, pp. 1112–1115, Sep 2016.

[13] F. Rockhold, P. Nisen, and A. Freeman, "Data Sharing at a Crossroads," *New England Journal of Medicine*, vol. 375, pp. 1115–1117, Sep 2016.

[14] B. Lo and S. N. Goodman, "Sharing Clinical Research Data: Finding the Right Balance," *JAMA Internal Medicine*, vol. 177, p. 1241, Sep 2017.

[15] H. Pampel, P. Vierkant, F. Scholze, R. Bertelmann, M. Kindling, J. Klump, H.-J. Goebelbecker, J. Gundlach, P. Schirmbacher, and U. Dierolf, "Making research data repositories visible: the re3data.org Registry," *PLOS ONE*, vol. 8, p. e78080, Nov 2013.

[16] "re3data.org — Registry of Research Data Repositories." `https://doi.org/10.17616/R3D` [Accessed 3 Jul 2018].

[17] P. McQuilton, A. Gonzalez-Beltran, P. Rocca-Serra, M. Thurston, A. Lister, E. Maguire, and S.-A. Sansone, "BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences," *Database: the Journal of Biological Databases and Curation*, vol. 2016, p. baw075, May 2016.

[18] "FAIRsharing.org — standards, databases, policies," 2009. `https://fairsharing.org/` [Accessed 4 Feb 2018].

[19] D. B. Taichman, J. Backus, C. Baethge, H. Bauchner, P. W. de Leeuw, J. M. Drazen, J. Fletcher, F. A. Frizelle, T. Groves, A. Haileamlak, A. James, C. Laine, L. Peiperl, A. Pinborg, P. Sahni, and S. Wu, "Sharing Clinical Trial Data: A Proposal from the International Committee of Medical Journal Editors," *PLoS Medicine*, vol. 13, p. e1001950, Jan 2016.

[20] Higher Education Funding Council for England, Research Councils UK, Universities UK, and Wellcome Trust, "Concordat on Open Research Data," 2016. Available at: `http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/` [Accessed 4 Feb 2018].

[21] R. Kiley, T. Peatfield, J. Hansen, and F. Reddington, "Data Sharing from Clinical Trials — A Research Funder's Perspective," *New England Journal of Medicine*, vol. 377, pp. 1990–1992, Nov 2017.

[22] Springer Nature, "Research Data Policy Types," 2018. Available at: `http://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096` [Accessed 4 Feb 2018].

[23] S. Bonini, H.-G. Eichler, N. Wathion, and G. Rasi, "Transparency and the European Medicines Agency – sharing of clinical trial data," *New England Journal of Medicine*, vol. 371, pp. 2452–5, Dec 2014.

[24] World Health Organization, "International Clinical Trials Registry Platform (ICTRP) — Joint statement on public disclosure of results from clinical trials," 2017. Available at: `http://www.who.int/ictrp/results/jointstatement/en/` [Accessed 4 Feb 2018].

[25] J. E. Kratz and C. Strasser, "Researcher perspectives on publication and peer review of data," *PLOS ONE*, vol. 10, no. 2, p. e0117619, 2015.

[26] E. Warren, "Strengthening Research through Data Sharing," *New England Journal of Medicine*, vol. 375, pp. 401–403, Aug 2016.

[27] K. Hate, S. Meherally, N. Shah More, A. Jayaraman, S. Bull, M. Parker, and D. Osrin, "Sweat, Skepticism, and Uncharted Territory," *Journal of Empirical Research on Human Research Ethics*, vol. 10, pp. 239–250, Jul 2015.

[28] A. Swan and S. Brown, "To share or not to share: Publication and Quality Assurance of Research Data Outputs," tech. rep., Research Information Network, 2008. Available at: `http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs` [Accessed 25 Oct 2017].

[29] M. Rani and B. S. Buckley, "Systematic archiving and access to health research data: rationale, current status and way forward.," *Bulletin of the World Health Organization*, vol. 90, pp. 932–9, Dec 2012.

[30] S. R. P. Franzen, C. Chandler, F. Enquselassie, S. Siribaddana, J. Atashili, B. Angus, and T. Lang, "Understanding the investigators: a qualitative study investigating the barriers and enablers to the implementation of local investigator-initiated clinical trials in Ethiopia," *BMJ Open*, vol. 3, p. e003616, Nov 2013.

[31] S. R. P. Franzen, C. Chandler, S. Siribaddana, J. Atashili, B. Angus, and T. Lang, "Strategies for developing sustainable health research capacity in low and middle-income countries: a prospective, qualitative study investigating the barriers and enablers to locally led clinical trial conduct in Ethiopia, Cameroon and Sri Lanka," *BMJ Open*, vol. 7, p. e017246, Oct 2017.

[32] A. Rosala-Hallas, A. Bhangu, J. Blazeby, L. Bowman, M. Clarke, T. Lang, M. Nasser, N. Siegfried, K. Soares-Weiser, M. R. Sydes, D. Wang, J. Zhang, and P. R. Williamson, "Global health trials methodological research agenda: results from a priority setting exercise," *Trials*, vol. 19, p. 48, Dec 2018.

[33] E. Pisani and C. AbouZahr, "Sharing health data: good intentions are not enough," *Bulletin of the World Health Organization*, vol. 88, pp. 462–6, Jun 2010.

[34] M. Walport and P. Brest, "Sharing research data to improve public health," *The Lancet*, vol. 377, pp. 537–9, Feb 2011.

[35] K. Modjarrad, V. S. Moorthy, P. Millett, P.-S. Gsell, C. Roth, and M.-P. Kieny, "Developing Global Norms for Sharing Data and Results during Public Health Emergencies," *PLoS Medicine*, vol. 13, p. e1001935, Jan 2016.

[36] S. Bull, P. Y. Cheah, S. Denny, I. Jao, V. Marsh, L. Merson, N. Shah More, L. N. T. Nhan, D. Osrin, D. Tangseefa, D. Wassenaar, and M. Parker, "Best Practices for Ethical Sharing of Individual-Level Health Research Data From Low- and Middle-Income Settings," *Journal of Empirical Research on Human Research Ethics*, vol. 10, pp. 302–313, Jul 2015.

[37] L. Bezuidenhout and E. Chakauya, "Hidden concerns of sharing research data by low/middle-income country scientists," *Global Bioethics = Problemi di Bioetica*, vol. 29, pp. 39–54, Jan 2018.

[38] World Health Organization, "Health Topics — Clinical trials," 2015. Available at: `http://www.who.int/topics/clinical` [Accessed 29 May 2018].

[39] C. De Angelis, J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. M. Overbeke, T. V. Schroeder, H. C. Sox, M. B. Van Der Weyden, and International Committee of Medical Journal Editors, "Clinical trial registration: a statement from the International Committee of Medical Journal Editors," *Canadian Medical Association Journal*, vol. 171, pp. 606–7, Sep 2004.

[40] C. Laine, R. Horton, C. D. DeAngelis, J. M. Drazen, F. A. Frizelle, F. Godlee, C. Haug, P. C. Hébert, S. Kotzin, A. Marusic, P. Sahni, T. V. Schroeder, H. C. Sox, M. B. Van der Weyden, and F. W. A. Verheugt, "Clinical trial registration: looking back and moving ahead," *Canadian Medical Association Journal*, vol. 177, pp. 57–8, Jul 2007.

[41] V. Manja and S. Lakshminrusimha, "Epidemiology and Clinical Research Design, Part 1: Study Types," *NeoReviews*, vol. 15, pp. e558–e569, Dec 2014.

[42] V. Manja and S. Lakshminrusimha, "Epidemiology and Clinical Research Design, Part 2: Principles," *NeoReviews*, vol. 16, pp. e94–e108, Feb 2015.

[43] T. A. Lang, N. J. White, T. T. Hien, J. J. Farrar, N. P. J. Day, R. Fitzpatrick, B. J. Angus, E. Denis, L. Merson, P. Y. Cheah, R. Chilengi, R. Kimutai, and

K. Marsh, "Clinical Research in Resource-Limited Settings: Enhancing Research Capacity and Working Together to Make Trials Less Complicated," *PLoS Neglected Tropical Diseases*, vol. 4, p. e619, Jun 2010.

[44] N. L. Watson, C. Prosperi, A. J. Driscoll, M. M. Higdon, D. E. Park, M. Sanza, A. N. DeLuca, J. O. Awori, D. Goswami, E. Hammond, L. Hossain, C. Johnson, A. Kamau, L. Kuwanda, D. P. Moore, O. Neyzari, U. Onwuchekwa, D. Parker, P. Sapchookul, P. Seidenberg, A. Shamsul, K. Siazeele, P. Srisaengchai, M. Sylla, O. S. Levine, D. R. Murdoch, K. L. O'Brien, M. Wolff, and M. Deloria Knoll, "Data Management and Data Quality in PERCH, a Large International Case-Control Study of Severe Childhood Pneumonia," *Clinical Infectious Diseases*, vol. 64, pp. S238–S244, Jun 2017.

[45] World Health Organization and Special Programme for Research and Training in Tropical Diseases (TDR), "Global report for research on infectious diseases of poverty," Geneva: WHO, 2012. Available at: `http://www.who.int/tdr/publications/global_report/en/` [Accessed 11 Jan 2018].

[46] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), "ICH Guidelines." Available at: `http://www.ich.org/products/guidelines.html` [Accessed 14 Jun 2018].

[47] U.S. Department of Health and Human Services Food & Drug Administration (FDA), "Regulations: Good Clinical Practice and Clinical Trials," 2018. Available at: `https://www.fda.gov/scienceresearch/specialtopics/runningclinicaltrials/ucm155713.htm` [Accessed 3 Jul 2018].

[48] C. L. Ward, D. Shaw, D. Sprumont, O. Sankoh, M. Tanner, and B. Elger, "Good collaborative practice: reforming capacity building governance of international health research partnerships," *Globalization and Health*, vol. 14, p. 1, Dec 2018.

[49] World Health Organization, "Malaria — Tools for monitoring antimalarial drug efficacy," 2018. Available at: `http://www.who.int/malaria/areas/drug_resistance/efficacy-monitoring-tools/en/` [Accessed 20 Jun 2018].

[50] World Health Organization, "Malaria — Antimalarial drug efficacy and drug resistance," 2018. Available at: `http://www.who.int/malaria/areas/treatment/drug_efficacy/en/` [Accessed 20 Jun 2018].

[51] S. B. Johnson, F. J. Farach, K. Pelphrey, and L. Rozenblit, "Data management in clinical research: Synthesizing stakeholder perspectives," *Journal of Biomedical Informatics*, vol. 60, pp. 286–293, Apr 2016.

[52] UK Data Archive, "Create & Manage Data: Research Data Lifecycle," 2016. Available at: `http://www.data-archive.ac.uk/create-manage/life-cycle` [Accessed 13 Jun 2016].

[53] UK Data Service, "Manage Data: Research data lifecycle," 2018. Available at: https://www.ukdataservice.ac.uk/manage-data/lifecycle [Accessed 26 Jun 2018].

[54] US Department of Health and Human Services Food & Drug Administration (FDA), "Data Integrity and Compliance With CGMP: Guidance for Industry," 2016. Available at: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm [Accessed 29 May 2018].

[55] Z. A. Bhutta, J. Sommerfeld, Z. S. Lassi, R. A. Salam, and J. K. Das, "Global burden, distribution, and interventions for infectious diseases of poverty," *Infectious Diseases of Poverty*, vol. 3, p. 21, 2014.

[56] M. Marmot, "Social determinants of health inequalities," *The Lancet*, vol. 365, pp. 1099–1104, Mar 2005.

[57] J. C. Semenza, "Strategies to intervene on social determinants of infectious diseases," *Euro Surveillance*, vol. 15, pp. 32–9, Jul 2010.

[58] World Health Organization, "Neglected Tropical Diseases — Diseases," 2018. Available at: http://www.who.int/neglected_diseases/diseases/en/ [Accessed 29 May 2018].

[59] J. Adams, K. A. Gurney, and D. Pendlebury, *Thomson Reuters Global Research Report: Neglected Tropical Diseases*. Leeds: Evidence, 2012.

[60] A. F. Cowman, J. Healer, D. Marapana, and K. Marsh, "Malaria: Biology and Disease," *Cell*, vol. 167, pp. 610–624, Oct 2016.

[61] WHO Department of Control of Neglected Tropical Diseases, "Working to overcome the global impact of neglected tropical diseases: first WHO report on neglected tropical diseases," Geneva: World Health Organization, 2010.

[62] Y. Nuyens, "No Development Without Research: A challenge for capacity strengthening," Geneva: Global Forum for Health Research, 2005.

[63] N. L. Yozwiak, C. T. Happi, D. S. Grant, J. S. Schieffelin, R. F. Garry, P. C. Sabeti, and K. G. Andersen, "Roots, Not Parachutes: Research Collaborations Combat Outbreaks," *Cell*, vol. 166, pp. 5–8, Jun 2016.

[64] The World Bank, "Money and microbes: strengthening clinical research capacity to prevent epidemics," Washington, D.C.: World Bank Group, 2018.

[65] World Health Organization, "Research for universal health coverage: World health report 2013," Geneva: WHO Press, 2014.

[66] N. Shafiq, P. Pandhi, and S. Malhotra, "Investigator-initiated pragmatic trials in developing countries–much needed but much ignored," *British Journal of Clinical Pharmacology*, vol. 67, pp. 141–2, Jan 2009.

[67] N. A. Vasilevsky, J. Minnier, M. A. Haendel, and R. E. Champieux, "Reproducible and reusable research: are journal data sharing policies meeting the mark?," *PeerJ*, vol. 5, p. e3208, Apr 2017.

[68] M. A. Lansang and R. Dennis, "Building capacity in health research in the developing world," *Bulletin of the World Health Organization*, vol. 82, pp. 764–70, Oct 2004.

[69] E. E. Vasquez, J. S. Hirsch, L. M. Giang, and R. G. Parker, "Rethinking health research capacity strengthening," *Global Public Health*, vol. 8, pp. S104–S124, Jul 2013.

[70] C. Potter and R. Brough, "Systemic capacity building: a hierarchy of needs," *Health Policy and Planning*, vol. 19, pp. 336–45, Sep 2004.

[71] S. R. P. Franzen, C. Chandler, and T. Lang, "Health research capacity development in low and middle income countries: reality or rhetoric? A systematic meta-narrative review of the qualitative literature," *BMJ Open*, vol. 7, p. e012332, Jan 2017.

[72] J. Jesson, L. Matheson, and F. M. Lacey, *Doing your literature Review: traditional and systematic techniques*. London: SAGE Publications, 2011.

[73] C. AbouZahr and T. Boerma, "Health information systems: the foundations of public health," *Bulletin of the World Health Organization*, vol. 83, pp. 578–83, Aug 2005.

[74] M. Chan, M. Kazatchkine, J. Lob-Levyt, T. Obaid, J. Schweizer, M. Sidibe, A. Veneman, and T. Yamada, "Meeting the demand for results and accountability: a call for action on health data from eight global health agencies," *PLoS Medicine*, vol. 7, p. e1000223, Jan 2010.

[75] NHS Health Research Authority (UK), "Defining research: NRES guidance to help you decide if your project requires review by a Research Ethics Committee," 2013. Available at: `https://researchsupport.admin.ox.ac.uk/sites/default/files/researchsupport/documents/media/defining-research.pdf` [Accessed 14 Jun 2018].

[76] K. Bommakanti, I. Feldhaus, G. Motwani, R. A. Dicker, and C. Juillard, "Trauma registry implementation in low- and middle-income countries: challenges and opportunities," *The journal of Surgical Research*, vol. 223, pp. 72–86, Mar 2018.

[77] M. L. Willcox, W. Peersman, P. Daou, C. Diakité, F. Bajunirwe, V. Mubangizi, E. H. Mahmoud, S. Moosa, N. Phaladze, O. Nkomazana, M. Khogali, D. Diallo, J. De Maeseneer, and D. Mant, "Human resources for primary health care in sub-Saharan Africa: progress or stagnation?," *Human Resources for Health*, vol. 13, p. 76, Dec 2015.

[78] M. Nyikuri, B. Tsofa, E. Barasa, P. Okoth, and S. Molyneux, "Crises and Resilience at the Frontline—Public Health Facility Managers under Devolution in a Sub-County on the Kenyan Coast," *PLOS ONE*, vol. 10, p. e0144768, Dec 2015.

[79] S. D. Makombe, M. Hochgesang, A. Jahn, H. Tweya, B. Hedt, S. Chuka, J. K.-L. Yu, J. Aberle-Grasse, O. Pasulani, C. Bailey, K. Kamoto, E. J. Schouten, and A. D. Harries, "Assessing the quality of data aggregated by antiretroviral treatment clinics in Malawi," *Bulletin of the World Health Organization*, vol. 86, pp. 310–4, Apr 2008.

[80] W. Mphatswe, K. Mate, B. Bennett, H. Ngidi, J. Reddy, P. Barker, and N. Rollins, "Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa," *Bulletin of the World Health Organization*, vol. 90, pp. 176–182, Mar 2012.

[81] J. H. Ledikwe, J. Grignon, R. Lebelonyane, S. Ludick, E. Matshediso, B. W. Sento, A. Sharma, and B.-w. Semo, "Improving the quality of health information: a qualitative assessment of data management and reporting systems in Botswana," *Health Research Policy and Systems*, vol. 12, p. 7, Dec 2014.

[82] M. Kumar, D. Gotz, T. Nutley, and J. B. Smith, "Research gaps in routine health information system design barriers to data quality and use in low- and middle-income countries: A literature review," *The International Journal of Health Planning and Management*, vol. 33, pp. e1–e9, Jan 2018.

[83] K. Herbst, S. Juvekar, T. Bhattacharjee, M. Bangha, N. Patharia, T. Tei, B. Gilbert, and O. Sankoh, "The INDEPTH Data Repository: An International Resource for Longitudinal Population and Health Data From Health and Demographic Surveillance Systems," *Journal of Empirical Research on Human Research Ethics*, vol. 10, pp. 324–33, Jul 2015.

[84] P. Bocquier, C. Ginsburg, K. Herbst, O. Sankoh, and M. A. Collinson, "A training manual for event history data management using Health and Demographic Surveillance System data," *BMC Research Notes*, vol. 10, p. 224, Dec 2017.

[85] C. B. Aranda-Jan, N. Mohutsiwa-Dibe, and S. Loukanova, "Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa," *BMC Public Health*, vol. 14, p. 188, Feb 2014.

[86] J. Brinkel, A. Krämer, R. Krumkamp, J. May, and J. Fobil, "Mobile Phone-Based mHealth Approaches for Public Health Surveillance in Sub-Saharan Africa: A Systematic Review," *International Journal of Environmental Research and Public Health*, vol. 11, pp. 11559–11582, Nov 2014.

[87] S. Agarwal, H. B. Perry, L.-A. Long, and A. B. Labrique, "Evidence on feasibility and effective use of mHealth strategies by frontline health workers in developing countries: systematic review," *Tropical Medicine & International Health*, vol. 20, pp. 1003–1014, Aug 2015.

[88] K. Shirima, O. Mukasa, J. A. Schellenberg, F. Manzi, D. John, A. Mushi, M. Mrisho, M. Tanner, H. Mshinda, and D. Schellenberg, "The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania," *Emerging Themes in Epidemiology*, vol. 4, p. 5, Jun 2007.

[89] M. Ali, J. L. Deen, A. Khatib, G. Enwere, L. von Seidlein, R. Reyburn, S. M. Ali, N. Y. Chang, V. Perroud, F. Marodon, A. A. Saleh, R. Hashim, A. L. Lopez, J. Beard, B. N. Ley, K. Thriemer, M. K. Puri, B. Sah, M. S. Jiddawi, and J. D. Clemens, "Paperless registration during survey enumerations and large oral cholera mass vaccination in Zanzibar, the United Republic of Tanzania," *Bulletin of the World Health Organization*, vol. 88, pp. 556–9, Jul 2010.

[90] D. Paudel, M. Ahmed, A. Pradhan, and R. Lal Dangol, "Successful use of tablet personal computers and wireless technologies for the 2011 Nepal Demographic and Health Survey," *Global Health, Science and Practice*, vol. 1, pp. 277–84, Aug 2013.

[91] T. Homan, A. Di Pasquale, I. Kiche, K. Onoka, A. Hiscox, C. Mweresa, W. R. Mukabana, W. Takken, and N. Maire, "Innovative tools and OpenHDS for health and demographic surveillance on Rusinga Island, Kenya.," *BMC Research Notes*, vol. 8, p. 397, Sep 2015.

[92] P. Byass, S. Hounton, M. Ouédraogo, H. Somé, I. Diallo, E. Fottrell, A. Emmelin, and N. Meda, "Direct data capture using hand-held computers in rural Burkina Faso: experiences, benefits and lessons learnt," *Tropical Medicine & International Health*, vol. 13 Suppl 1, pp. 25–30, Jul 2008.

[93] S. Style, B. J. Beard, H. Harris-Fry, A. Sengupta, S. Jha, B. P. Shrestha, A. Rai, V. Paudel, M. Thondoo, A.-M. Pulkki-Brannstrom, J. Skordis-Worrall, D. S. Manandhar, A. Costello, and N. M. Saville, "Experiences in running a complex electronic data capture system using mobile phones in a large-scale population trial in southern Nepal," *Global Health Action*, vol. 10, p. 1330858, Jan 2017.

[94] A. Raza, I. Raza, T. M. Drake, A. B. Sadar, M. Adil, F. Baluch, A. U. Qureshi, and E. M. Harrison, "The efficiency, accuracy and acceptability of smartphone-delivered data collection in a low-resource setting – A prospective study," *International Journal of Surgery*, vol. 44, pp. 252–254, Aug 2017.

[95] P. Memiah, T. Ah Mu, J. Penner, K. Owour, C. Ngunu-Gituathi, K. Prevot, V. Mochache, P. Wekesa, J. Oyore, S. Muhula, and P. Komba, "Bridging the Gap in Implementation Science: Evaluating a Capacity-Building Program in Data Management, Analysis, Utilization, and Dissemination in Low- and Middle-Income Countries," *Population Health Management*, p. pop.2017.0093, Sep 2017.

[96] M. Newman Owiredu, N. B. Bellare, C. C. Chakanyuka Musanhu, T. A. Oyelade, E. M. Thom, F. Bigirimana, C. Anyaike, M. T. Eliya, A. Mushavi, E. A. Nyarko,

D. O. Okello, and F. R. Zawaira, "Building Health System Capacity Through Implementation Research," *Journal of Acquired Immune Deficiency Syndromes*, vol. 75, pp. S240–S247, Jun 2017.

[97] L. Perrier, E. Blondal, A. P. Ayala, D. Dearborn, T. Kenny, D. Lightfoot, R. Reka, M. Thuna, L. Trimble, and H. MacDonald, "Research data management in academic institutions: A scoping review," *PLOS ONE*, vol. 12, p. e0178261, May 2017.

[98] P. Y. Cheah and N. P. J. Day, "Data sharing: experience from a tropical medicine research unit," *The Lancet*, vol. 390, p. 1642, Oct 2017.

[99] FOSTER (Fostering the practical implementation of Open Science in Horizon 2020 and beyond), "FOSTER — The future of science is Open." `https://www.fosteropenscience.eu/` [Accessed 13 Jun 2018].

[100] D. K. Babre, "Clinical data interchange standards consortium: A bridge to overcome data standardisation," *Perspectives in Clinical Research*, vol. 4, pp. 115–6, Apr 2013.

[101] M. N. Zozus, A. Lazarov, L. R. Smith, T. E. Breen, S. L. Krikorian, P. S. Zbyszewski, S. K. Knoll, D. A. Jendrasek, D. C. Perrin, D. N. Zambas, T. B. Williams, and C. F. Pieper, "Analysis of professional competencies for the clinical research data management profession: implications for training and professional certification," *Journal of the American Medical Informatics Association*, vol. 24, pp. 737–745, Jul 2017.

[102] L. Lyon and A. Brenner, "Bridging the Data Talent Gap: Positioning the iSchool as an Agent for Change," *International Journal of Digital Curation*, vol. 10, pp. 111–122, Mar 2015.

[103] Y. Demchenko, A. Belloum, and T. Wiktorski, "EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS), Release 2," EDISON, 2017. Available at: `http://edison-project.eu/data-science-competence-framework-cf-ds` [Accessed 3 Jul 2018].

[104] Education for Data Intensive Science to Open New science frontiers (EDISON), "EDISON Data Science Framework (EDSF)," 2017. `http://edison-project.eu/edison/edison-data-science-framework-edsf` [Accessed 13 Jun 2018].

[105] Research Data Alliance (RDA), "Education and Training on handling of research data Interest Group." Available at: `https://rd-alliance.org/groups/education-and-training-handling-research-data.html` [Accessed 13 Jun 2018].

[106] C. Ijsselmuiden, D. L. Marais, F. Becerra-Posada, and H. Ghannem, "Africa's neglected area of human resources for health research – the way forward," *South African Medical Journal = Suid-Afrikaanse Tydskrif vir Geneeskunde*, vol. 102, pp. 228–33, Mar 2012.

[107] B. Ene-Iordache, S. Carminati, L. Antiga, N. Rubis, P. Ruggenenti, G. Remuzzi, and A. Remuzzi, "Developing Regulatory-compliant Electronic Case Report Forms for Clinical Trials: Experience with The Demand Trial," *Journal of the American Medical Informatics Association*, vol. 16, pp. 404–408, May 2009.

[108] M. A. Wildeman, J. Zandbergen, A. Vincent, C. Herdini, J. M. Middeldorp, R. Fles, O. Dalesio, E. van der Donk, and I. B. Tan, "Can an online clinical data management service help in improving data collection and data quality in a developing country setting?," *Trials*, vol. 12, no. 1, p. 190, 2011.

[109] D. H. M. Sanvictores, M. G. Lucero, H. Nohynek, V. L. Tallo, A. Tanskanen, L. T. Nillos, and G. Williams, "The data management of a phase III efficacy trial of an 11-valent pneumococcal conjugate vaccine and related satellite studies conducted in the Philippines," *BMC Research Notes*, vol. 5, no. 1, p. 274, 2012.

[110] S. Hossmann, A. G. Haynes, A. Spoerri, I. D. Diatta, B. Aboubacar, M. Egger, F. Rintelen, and S. Trelle, "Data management of clinical trials during an outbreak of Ebola virus disease," *Vaccine*, Oct 2017. [In Press, Corrected Proof].

[111] J. Y. Park, D. R. Kim, B. Haldar, A. H. Mallick, S. A. Kim, A. Dey, R. K. Nandy, D. K. Paul, S. Choudhury, S. Sahoo, T. F. Wierzba, D. Sur, S. Kanungo, M. Ali, and B. Manna, "Use of the data system for field management of a clinical study conducted in Kolkata, India," *BMC Research Notes*, vol. 9, p. 20, Dec 2016.

[112] E. Ohuma, L. Hoch, C. Cosgrove, H. Knight, L. Cheikh Ismail, L. Juodvirsiene, A. Papageorghiou, H. Al-Jabri, M. Domingues, P. Gilli, N. Kunnawar, N. Musee, F. Roseman, A. Carter, M. Wu, D. Altman, and International Fetal and Newborn Growth Consortium for the 21st Century, "Managing data for the international, multicentre INTERGROWTH-21st Project," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 120, pp. 64–70, Sep 2013.

[113] Q. S.-U. Rahman, M. S. Islam, B. Hossain, T. Hossain, N. E. Connor, M. J. Jaman, M. M. Rahman, A. S. M. N. U. Ahmed, I. Ahmed, M. Ali, S. M. I. Moin, L. Mullany, S. K. Saha, and S. E. Arifeen, "Centralized Data Management in a Multicountry, Multisite Population-based Study," *The Pediatric Infectious Disease Journal*, vol. 35, pp. S23–S28, May 2016.

[114] A. Steiner, J. Hella, S. Grüninger, G. Mhalu, F. Mhimbira, C. I. Cercamondi, B. Doulla, N. Maire, and L. Fenner, "Managing research and surveillance projects in real-time with a novel open-source eManagement tool designed for under-resourced countries," *Journal of the American Medical Informatics Association*, vol. 23, pp. 916–923, Sep 2016.

[115] P. Y. Cheah, N. P. J. Day, M. Parker, and S. Bull, "Sharing Individual-Level Health Research Data: Experiences, Challenges and a Research Agenda," *Asian Bioethics Review*, vol. 9, no. 4, pp. 393–400, 2017.

[116] Society for Clinical Data Management, "Publications — GCDMP©." Available at: https://www.scdm.org/publications/gcdmp/ [Accessed 27 Jun 2018].

[117] W. Kuchinke, C. Ohmann, Q. Yang, N. Salas, J. Lauritsen, F. Gueyffier, A. Leizorovicz, C. Schade-Brittinger, M. Wittenberg, Z. Voko, S. Gaynor, M. Cooney, P. Doran, A. Maggioni, A. Lorimer, F. Torres, G. McPherson, J. Charwill, M. Hellström, and S. Lejeune, "Heterogeneity prevails: the state of clinical trial data management in Europe - results of a survey of ECRIN centres," *Trials*, vol. 11, p. 79, Dec 2010.

[118] C. Ohmann, W. Kuchinke, S. Canham, J. Lauritsen, N. Salas, C. Schade-Brittinger, M. Wittenberg, G. McPherson, J. McCourt, F. Gueyffier, A. Lorimer, F. Torres, and ECRIN Working Group on Data Centres, "Standard requirements for GCP-compliant data management in multinational clinical trials," *Trials*, vol. 12, p. 85, Dec 2011.

[119] C. Ohmann, S. Canham, C. Cornu, J. Dreß, F. Gueyffier, W. Kuchinke, E. B. Nicolis, and M. Wittenberg, "Revising the ECRIN standard requirements for information technology and data management in clinical trials," *Trials*, vol. 14, p. 97, Apr 2013.

[120] C. Ohmann, S. Canham, J. Demotes, G. Chêne, J. Lauritsen, H. Martins, R. Mendes, E. Nicolis, A. Svobodnik, and F. Torres, "Raising standards in clinical research – The impact of the ECRIN data centre certification programme, 2011–2016," *Contemporary Clinical Trials Communications*, vol. 5, pp. 153–159, Mar 2017.

[121] L. M. Bezuidenhout, S. Leonelli, A. H. Kelly, and B. Rappert, "Beyond the digital divide: Towards a situated approach to open data," *Science and Public Policy*, vol. 44, pp. 464–475, Aug 2017.

[122] M. Kelly, M. Dowling, and M. Millar, "The search for understanding: the role of paradigms," *Nurse Researcher*, vol. 25, pp. 9–13, Mar 2018.

[123] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*. Thousand Oaks: SAGE Publications, 2nd ed., 2011.

[124] D. E. Gray, *Doing research in the real world*. London: SAGE Publications, 3rd ed., 2014.

[125] A. O. Aragón, "A Case for Surfacing Theories of Change for Purposeful Organisational Capacity Development," *IDS Bulletin*, vol. 41, pp. 36–46, May 2010.

[126] D. Burns, "A systemic perspective," in *Systemic action research: a strategy for whole system change*, ch. 2, pp. 21–40, Bristol: Policy Press, 2007.

[127] L. Lingard, M. Albert, and W. Levinson, "Grounded theory, mixed methods, and action research," *BMJ*, vol. 337, p. a567, Aug 2008.

[128] D. H. Meadows, *Thinking in systems: a primer*. Chelsea Green Publishing, 2008.

[129] Y. Claeys, J. Smedley, A. Bhattarai, H. Kalonji, P. Malaviya, R. Meester, D. Mwakazanga, M. Omello, Sayouba, Ouedraogo, S. Sok, S. Sokleng, M. Thiongo, H. Tilahun, A. Woukeu, H. van Loen, B. Faragher, and R. Ravinetto, "Overcoming the challenges of Clinical Data Management in low and middle income countries: A context-adapted Data Management Plan and Lifecycle," in *ASTMH Annual Meeting*, (New Orleans), American Society of Tropical Medicine and Hygiene, 2014. Available at: `http://www.admitnetwork.org/work-packages/astmh-2014/` [Accessed 19 Dec 2017].

[130] P. G. Smith, R. H. Morrow, and D. A. Ross, "Data management," in *Field Trials of Health Interventions: A Toolbox*, ch. 20, Oxford University Press, 3rd ed., 2015.

[131] L. Merson, O. Gaye, and P. J. Guerin, "Avoiding Data Dumpsters – Toward Equitable and Useful Data Sharing," *New England Journal of Medicine*, vol. 374, pp. 2414–5, Jun 2016.

[132] C. F. Saarnak, J. Utzinger, and T. K. Kristensen, "Collection, verification, sharing and dissemination of data: the CONTRAST experience," *Acta Tropica*, vol. 128, pp. 407–411, Nov 2013.

[133] World Health Organization and Special Programme for Research and Training in Tropical Diseases (TDR), "Schistosomiasis Data Platform Stakeholders' Meeting," 2015. Available at: `http://www.who.int/tdr/news/2015/processes-for-sharing-shisto-treat-eff-trials-data-agreed/en/` [Accessed 25 Oct 2017].

[134] B. Gryseels, K. Polman, J. Clerinx, and L. Kestens, "Human schistosomiasis," *The Lancet*, vol. 368, pp. 1106–1118, Sep 2006.

[135] Centers for Disease Control and Prevention, "DPDx - Laboratory Identification of Parasitic Diseases of Public Health Concern: Schistosomiasis Infection," 2017. Available at: `https://www.cdc.gov/dpdx/schistosomiasis/` [Accessed 19 Feb 2018].

[136] World Health Organization, "WHO Media centre – Schistosomiasis fact sheet," 2017. Available at: `http://www.who.int/mediacentre/factsheets/fs115/en/` [Accessed 19 Feb 2018].

[137] A. Danso-Appiah, P. Garner, P. L. Olliaro, and J. Utzinger, "Treatment of urinary schistosomiasis: methodological issues and research needs identified through a Cochrane systematic review," *Parasitology*, vol. 136, p. 1837, Nov 2009.

[138] J. R. Herricks, P. J. Hotez, V. Wanga, L. E. Coffeng, J. A. Haagsma, M.-G. Basáñez, G. Buckle, C. M. Budke, H. Carabin, E. M. Fèvre, T. Fürst, Y. A. Halasa, C. H. King, M. E. Murdoch, K. D. Ramaiah, D. S. Shepard, W. A. Stolk, E. A. Undurraga, J. D. Stanaway, M. Naghavi, and C. J. L. Murray, "The global burden of disease study 2013: What does it mean for the NTDs?," *PLoS Neglected Tropical Diseases*, vol. 11, p. e0005424, Aug 2017.

[139] A. Danso-Appiah, P. L. Olliaro, S. Donegan, D. Sinclair, and J. Utzinger, "Drugs for treating *Schistosoma mansoni* infection," in *Cochrane Database of Systematic Reviews*, no. 2, p. CD000528, Chichester, UK: John Wiley & Sons, Ltd, Feb 2013.

[140] C. V. Kramer, F. Zhang, D. Sinclair, and P. L. Olliaro, "Drugs for treating urinary schistosomiasis," in *Cochrane Database of Systematic Reviews*, no. 8, p. CD000053, Chichester, UK: John Wiley & Sons, Ltd, Aug 2014.

[141] World Health Organization, "Assessing the efficacy of anthelminthic drugs against schistosomiasis and soil-transmitted helminthiases," Geneva: WHO Press, 2013. Available at: `http://apps.who.int/iris/handle/10665/79019` [Accessed 4 Jul 2018].

[142] World Health Organization, "Preventive chemotherapy in human helminthiasis: coordinated use of anthelminthic drugs in control interventions – a manual for health professionals and programme managers," Geneva: WHO Press, 2006. Available at: `http://www.who.int/neglected_diseases/preventive_chemotherapy/pct_manual/en/` [Accessed 24 Feb 2018].

[143] World Health Organization, "Schistosomiasis: progress report 2001-2011, strategic plan 2012-2020," Geneva: WHO Press, 2013. Available at: `http://www.who.int/schistosomiasis/resources/9789241503174/en/` [Accessed 19 Oct 2017].

[144] WHO Department of Control of Neglected Tropical Diseases, "Schistosomiasis and soil-transmitted helminthiases: number of people treated in 2016," *Weekly epidemiological record*, vol. 49, no. 92, pp. 749–760, 2017.

[145] WHO Expert Committee on the Control of Schistosomiasis, "Prevention and control of schistosomiasis and soil-transmitted helminthiasis," tech. rep., World Health Organization, Geneva, 2002. Available at: `http://www.who.int/intestinal_worms/resources/who_trs_912/en/` [Accessed 24 Feb 2018].

[146] P. Magnussen, "Treatment and re-treatment strategies for schistosomiasis control in different epidemiological settings: a review of 10 years' experiences," *Acta Tropica*, vol. 86, pp. 243–54, May 2003.

[147] M. J. Doenhoff, D. Cioli, and J. Utzinger, "Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis," *Current Opinion in Infectious Diseases*, vol. 21, pp. 659–667, Dec 2008.

[148] S. Geerts and B. Gryseels, "Drug resistance in human helminths: current situation and lessons from livestock," *Clinical Microbiology Reviews*, vol. 13, pp. 207–22, Apr 2000.

[149] R. M. Kaplan and A. N. Vidyashankar, "An inconvenient truth: Global worming and anthelmintic resistance," *Veterinary Parasitology*, vol. 186, pp. 70–78, May 2012.

[150] E. Redman, F. Whitelaw, A. Tait, C. Burgess, Y. Bartley, P. J. Skuce, F. Jackson, and J. S. Gilleard, "The emergence of resistance to the benzimidazole anthlemintics in parasitic nematodes of livestock is characterised by multiple independent hard and soft selective sweeps," *PLoS Neglected Tropical Diseases*, vol. 9, p. e0003494, Feb 2015.

[151] F. F. Stelma, I. Talla, S. Sow, A. Kongs, M. Niang, K. Polman, A. M. Deelder, and B. Gryseels, "Efficacy and side effects of praziquantel in an epidemic focus of *Schistosoma mansoni*," *The American Journal of Tropical Medicine and Hygiene*, vol. 53, pp. 167–70, Aug 1995.

[152] S. D. Melman, M. L. Steinauer, C. Cunningham, L. S. Kubatko, I. N. Mwangi, N. B. Wynn, M. W. Mutuku, D. M. S. Karanja, D. G. Colley, C. L. Black, W. E. Secor, G. M. Mkoji, and E. S. Loker, "Reduced Susceptibility to Praziquantel among Naturally Occurring Kenyan Isolates of *Schistosoma mansoni*," *PLoS Neglected Tropical Diseases*, vol. 3, p. e504, Aug 2009.

[153] W. Wang, L. Wang, and Y.-S. Liang, "Susceptibility or resistance of praziquantel in human schistosomiasis: a review," *Parasitology Research*, vol. 111, pp. 1871–1877, Nov 2012.

[154] M. A. Babyak, "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models," *Psychosomatic Medicine*, vol. 66, pp. 411–421, May 2004.

[155] M. Albonico, B. Levecke, P. LoVerde, A. Montresor, R. Prichard, J. Vercruysse, and J. Webster, "Monitoring the efficacy of drugs for neglected tropical diseases controlled by preventive chemotherapy," *Journal of Global Antimicrobial Resistance*, vol. 3, no. 4, pp. 229–236, 2015.

[156] J. T. Bush, M. Wasunna, F. Alves, J. Alvar, P. L. Olliaro, M. Otieno, C. H. Sibley, N. Strub Wourgaft, and P. J. Guerin, "Systematic review of clinical trials assessing the therapeutic efficacy of visceral leishmaniasis treatments: A first step to assess the feasibility of establishing an individual patient data sharing platform," *PLoS Neglected Tropical Diseases*, vol. 11, p. e0005781, Sep 2017.

[157] J. B. Halder, J. Benton, A. M. Julé, P. J. Guérin, P. L. Olliaro, M.-G. Basáñez, and M. Walker, "Systematic review of studies generating individual participant data on the efficacy of drugs for treating soil-transmitted helminthiases and the case for data-sharing," *PLoS Neglected Tropical Diseases*, vol. 11, p. e0006053, Oct 2017.

[158] J. Zwang and P. L. Olliaro, "Clinical efficacy and tolerability of praziquantel for intestinal and urinary schistosomiasis-a meta-analysis of comparative and non-comparative clinical trials," *PLoS Neglected Tropical Diseases*, vol. 8, p. e3286, Nov 2014.

[159] P. L. Olliaro, M. Vaillant, A. Diawara, J. T. Coulibaly, A. Garba, J. Keiser, C. H. King, S. Knopp, A. Landouré, E. K. N'Goran, G. Raso, A. U. Scherrer, J. C. Sousa-Figueiredo, K. Stete, X.-N. Zhou, and J. Utzinger, "Toward Measuring Schistosoma Response to Praziquantel Treatment with Appropriate Descriptors of Egg Excretion," *PLoS Neglected Tropical Diseases*, vol. 9, p. e0003821, Jun 2015.

[160] T. H. Vines, A. Y. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison, "The Availability of Research Data Declines Rapidly with Article Age," *Current Biology*, vol. 24, pp. 94–97, Jan 2014.

[161] "Zotero." `https://www.zotero.org/` [Accessed 18 Oct 2017].

[162] World Health Organization, "PCT databank," 2017. Available at: `http://who.int/neglected_diseases/preventive_chemotherapy/sch/en/` [Accessed 18 Oct 2017].

[163] E. S. Loker, "A comparative study of the life-histories of mammalian schistosomes," *Parasitology*, vol. 87, pp. 343–69, Oct 1983.

[164] M. Walker, T. S. Mabud, P. L. Olliaro, J. T. Coulibaly, C. H. King, G. Raso, A. U. Scherrer, J. R. Stothard, J. C. Sousa-Figueiredo, K. Stete, J. Utzinger, and M.-G. Basáñez, "New approaches to measuring anthelminthic drug efficacy: parasitological responses of childhood schistosome infections to treatment with praziquantel," *Parasites & Vectors*, vol. 9, p. 41, Dec 2016.

[165] A. Montresor, "Cure rate is not a valid indicator for assessing drug efficacy and impact of preventive chemotherapy interventions against schistosomiasis and soil-transmitted helminthiasis," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 105, pp. 361–363, Jul 2011.

[166] A. M. Julé, M. Vaillant, T. A. Lang, P. J. Guérin, and P. L. Olliaro, "The Schistosomiasis Clinical Trials Landscape: A Systematic Review of Antischistosomal Treatment Efficacy Studies and a Case for Sharing Individual Participant-Level Data (IPD)," *PLoS Neglected Tropical Diseases*, vol. 10, p. e0004784, Jun 2016.

[167] J. Zwang and P. Olliaro, "Efficacy and safety of praziquantel 40 mg/kg in preschool-aged and school-aged children: a meta-analysis," *Parasites & Vectors*, vol. 10, p. 47, Dec 2017.

[168] Food and Agriculture Organisation of the United Nations, "Global Administrative Unit Layers (GAUL) Map," 2013. Available at: `http://ref.data.fao.org/map?entryId=f7e7adb0-88fd-11da-a88f-000d939bc5d8` [Accessed 24 Feb 2018].

[169] R. Bivand, C. Rundel, E. Pebesma, R. Stuetz, and K. O. Hufthammer, "Package 'rgeos'," 2017. Available at: `https://cran.rstudio.com/web/packages/rgeos/rgeos.pdf` [Accessed 22 Oct 2017].

[170] The R Foundation, "The R Project for Statistical Computing." `https://www.r-project.org/` [Accessed 18 Oct 2017].

[171] M. Walker, T. S. Churcher, and M.-G. Basáñez, "Models for measuring anthelmintic drug efficacy for parasitologists," *Trends in Parasitology*, vol. 30, pp. 528–537, Nov 2014.

[172] T. Crellen, M. Walker, P. H. L. Lamberton, N. B. Kabatereine, E. M. Tukahebwa, J. A. Cotton, and J. P. Webster, "Reduced Efficacy of Praziquantel Against *Schistosoma mansoni* Is Associated With Multiple Rounds of Mass Drug Administration," *Clinical Infectious Diseases*, vol. 63, pp. 1151–1159, Nov 2016.

[173] A. Zeileis, "Object-Oriented Computation of Sandwich Estimators," *Journal of Statistical Software*, vol. 16, pp. 1–16, Aug 2006.

[174] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and PRISMA Group, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," *PLoS Medicine*, vol. 6, p. e1000097, Jul 2009.

[175] L. A. Stewart and J. F. Tierney, "To IPD or not to IPD?," *Evaluation & the Health Professions*, vol. 25, pp. 76–97, Mar 2002.

[176] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: rationale, conduct, and reporting," *BMJ*, vol. 340, p. c221, Feb 2010.

[177] J. P. T. Higgins and S. Green, eds., *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. The Cochrane Collaboration, 2017. Available at: `http://handbook.cochrane.org` [Accessed 4 Jul 2017].

[178] D. Glinz, R. F. Hurrell, A. A. Righetti, C. Zeder, L. G. Adiossan, H. Tjalsma, J. Utzinger, M. B. Zimmermann, E. K. N'Goran, and R. Wegmüller, "In Ivorian school-age children, infection with hookworm does not reduce dietary iron absorption or systemic iron utilization, whereas afebrile *Plasmodium falciparum* infection reduces iron absorption by half," *The American Journal of Clinical Nutrition*, vol. 101, pp. 462–70, Mar 2015.

[179] D. De Clercq, J. Vercruysse, A. Kongs, P. Verlé, J. P. Dompnier, and P. C. Faye, "Efficacy of artesunate and praziquantel in *Schistosoma haematobium* infected schoolchildren," *Acta Tropica*, vol. 82, pp. 61–6, Apr 2002.

[180] J. Keiser, K. D. Silué, L. K. Adiossan, N. A. N'Guessan, N. Monsan, J. Utzinger, and E. K. N'Goran, "Praziquantel, Mefloquine-Praziquantel, and Mefloquine-Artesunate-Praziquantel against *Schistosoma haematobium*: A Randomized, Exploratory, Open-Label Trial," *PLoS Neglected Tropical Diseases*, vol. 8, p. e2975, Jul 2014.

[181] R. Barakat, N. E. Abou El-Ela, S. Sharaf, O. El Sagheer, S. Selim, H. Tallima, M. J. Bruins, K. B. Hadley, and R. El Ridi, "Efficacy and safety of arachidonic acid for

treatment of school-age children in *Schistosoma mansoni* high-endemicity regions," *The American Journal of Tropical Medicine and Hygiene*, vol. 92, pp. 797–804, Apr 2015.

[182] O. Ojurongbe, O. R. Sina-Agbaje, A. Busari, P. N. Okorie, T. A. Ojurongbe, and A. A. Akindele, "Efficacy of praziquantel in the treatment of *Schistosoma haematobium* infection among school-age children in rural communities of Abeokuta, Nigeria," *Infectious Diseases of Poverty*, vol. 3, no. 1, p. 30, 2014.

[183] S. Botros, H. Sayed, H. El-Dusoki, H. Sabry, I. Rabie, M. El-Ghannam, M. Hassanein, Y. A. El-Wahab, and D. Engels, "Efficacy of mirazid in comparison with praziquantel in Egyptian *Schistosoma mansoni*-infected school children and households," *The American Journal of Tropical Medicine and Hygiene*, vol. 72, pp. 119–23, Feb 2005.

[184] B. Erko, A. Degarege, K. Tadesse, A. Mathiwos, and M. Legesse, "Efficacy and side effects of praziquantel in the treatment of Schistosomiasis mansoni in schoolchildren in Shesha Kekele Elementary School, Wondo Genet, Southern Ethiopia," *Asian Pacific Journal of Tropical Biomedicine*, vol. 2, pp. 235–9, Mar 2012.

[185] R. Barakat and H. El Morshedy, "Efficacy of two praziquantel treatments among primary school children in an area of high *Schistosoma mansoni* endemicity, Nile Delta, Egypt," *Parasitology*, vol. 138, pp. 440–6, Apr 2011.

[186] P. C. Inyang-Etoh, G. C. Ejezie, M. F. Useh, and E. C. Inyang-Etoh, "Efficacy of a combination of praziquantel and artesunate in the treatment of urinary schistosomiasis in Nigeria," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 103, pp. 38–44, Jan 2009.

[187] B. Reta and B. Erko, "Efficacy and side effects of praziquantel in the treatment for *Schistosoma mansoni* infection in school children in Senbete Town, northeastern Ethiopia," *Tropical Medicine & International Health*, vol. 18, pp. 1338–43, Nov 2013.

[188] WHO Department of Control of Neglected Tropical Diseases, "Schistosomiasis and soil-transmitted helminthiases: number of people treated in 2015," *Weekly Epidemiological Record*, vol. 49/50, no. 91, pp. 585–600, 2016.

[189] WHO Department of Control of Neglected Tropical Diseases, "Summary of global update on preventive chemotherapy implementation in 2016: crossing the billion," *Weekly Epidemiological Record*, vol. 40, no. 91, pp. 589–608, 2017.

[190] G. Augusto, P. Magnussen, T. K. Kristensen, C. C. Appleton, and B. J. Vennervald, "The influence of transmission season on parasitological cure rates and intensity of infection after praziquantel treatment of *Schistosoma haematobium*-infected schoolchildren in Mozambique," *Parasitology*, vol. 136, pp. 1771–9, Nov 2009.

[191] P. L. Olliaro, M. T. Vaillant, V. J. Belizario, N. J. S. Lwambo, M. Ouldabdallahi, O. S. Pieri, M. L. Amarillo, G. M. Kaatano, M. Diaw, A. C. Domingues, T. C. Favre,

O. Lapujade, F. Alves, and L. Chitsulo, "A multicentre randomized controlled trial of the efficacy and safety of single-dose praziquantel at 40 mg/kg vs. 60 mg/kg for treating intestinal schistosomiasis in the Philippines, Mauritania, Tanzania and Brazil," *PLoS Neglected Tropical Diseases*, vol. 5, p. e1165, Jun 2011.

[192] J. Keiser, N. A. N'Guessan, K. D. Adoubryn, K. D. Silué, P. Vounatsou, C. Hatz, J. Utzinger, and E. K. N'Goran, "Efficacy and Safety of Mefloquine, Artesunate, Mefloquine-Artesunate, and Praziquantel against *Schistosoma haematobium*: Randomized, Exploratory Open-Label Trial," *Clinical Infectious Diseases*, vol. 50, pp. 1205–1213, May 2010.

[193] K. Stete, S. J. Krauth, J. T. Coulibaly, S. Knopp, J. Hattendorf, I. Müller, L. K. Lohourignon, W. V. Kern, E. K. N'goran, and J. Utzinger, "Dynamics of *Schistosoma haematobium* egg output and associated infection parameters following treatment with praziquantel in school-aged children," *Parasites & Vectors*, vol. 5, p. 298, Dec 2012.

[194] G. Raso, E. K. N'Goran, A. Toty, A. Luginbühl, C. A. Adjoua, N. T. Tian-Bi, I. I. Bogoch, P. Vounatsou, M. Tanner, and J. Utzinger, "Efficacy and side effects of praziquantel against *Schistosoma mansoni* in a community of western Côte d'Ivoire," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 98, pp. 18–27, Jan 2004.

[195] A. U. Scherrer, M. K. Sjöberg, A. Allangba, M. Traoré, L. K. Lohourignon, A. B. Tschannen, E. K. N'Goran, and J. Utzinger, "Sequential analysis of helminth egg output in human stool samples following albendazole and praziquantel administration," *Acta Tropica*, vol. 109, pp. 226–31, Mar 2009.

[196] K. D. Adoubryn, C. G. Kouadio-Yapo, J. Ouhon, N. A. D. Aka, F. Bintto, and A. Assoumou, "[Intestinal parasites in children in Biankouma, Ivory Coast (mountaineous western region): efficacy and safety of praziquantel and albendazole]," *Médecine et Santé Tropicales*, vol. 22, no. 2, pp. 170–6, 2012. (Article in French).

[197] F. Mutapi, N. Rujeni, C. Bourke, K. Mitchell, L. Appleby, N. Nausch, N. Midzi, and T. Mduluza, "*Schistosoma haematobium* treatment in 1-5 year old children: safety and efficacy of the antihelminthic drug praziquantel," *PLoS Neglected Tropical Diseases*, vol. 5, p. e1143, May 2011.

[198] F. Mutapi, N. Imai, N. Nausch, C. D. Bourke, N. Rujeni, K. M. Mitchell, N. Midzi, M. E. J. Woolhouse, R. M. Maizels, and T. Mduluza, "Schistosome infection intensity is inversely related to auto-reactive antibody levels," *PLOS ONE*, vol. 6, p. e19149, May 2011.

[199] M. Ouldabdallahi, B. Ousmane, M. Ouldbezeid, D. Mamadou, L. Konaté, and L. Chitsulo, "[Comparison of the efficacy and safety of praziquantel administered in

single dose of 40 versus 60 mg/kg for treating urinary schistosomiasis in Mauritania]," *Bulletin de la Societé de Pathologie Exotique (1990)*, vol. 106, pp. 167–9, Aug 2013. (Article in French).

[200] L.-A. T. Tchuente, D. J. Shaw, L. Polla, D. Cioli, and J. Vercruysse, "Efficacy of praziquantel against *Schistosoma haematobium* infection in children," *The American Journal of Tropical Medicine and Hygiene*, vol. 71, pp. 778–782, Dec 2004.

[201] N. Midzi, D. Sangweme, S. Zinyowera, M. P. Mapingure, K. C. Brouwer, N. Kumar, F. Mutapi, G. Woelk, and T. Mduluza, "Efficacy and side effects of praziquantel treatment against *Schistosoma haematobium* infection among primary school children in Zimbabwe," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 102, pp. 759–766, Aug 2008.

[202] M. S. Sissoko, A. Dabo, H. Traore, M. Diallo, B. Traore, D. Konate, B. Niare, M. Diakite, B. Kamate, A. Traore, A. Bathily, A. Tapily, O. B. Toure, S. Cauwenbergh, H. F. Jansen, and O. K. Doumbo, "Efficacy of artesunate + sulfamethoxypyrazine/pyrimethamine versus praziquantel in the treatment of *Schistosoma haematobium* in children," *PLOS ONE*, vol. 4, p. e6732, Oct 2009.

[203] A. Guidi, C. Andolina, S. Makame Ame, M. Albonico, D. Cioli, and H. Juma Haji, "Praziquantel efficacy and long-term appraisal of schistosomiasis control in Pemba Island," *Tropical Medicine & International Health*, vol. 15, pp. 614–618, May 2010.

[204] C. C. Jinabhai, M. Taylor, A. Coutsoudis, H. M. Coovadia, A. M. Tomkins, and K. R. Sullivan, "Epidemiology of helminth infections: implications for parasite control programmes, a South African perspective," *Public Health Nutrition*, vol. 4, pp. 1211–1219, Dec 2001.

[205] S. Borrmann, N. Szlezak, J. F. Faucher, P. B. Matsiegui, R. Neubauer, R. K. Binder, B. Lell, and P. G. Kremsner, "Artesunate and praziquantel for the treatment of *Schistosoma haematobium* infections: a double-blind, randomized, placebo-controlled study," *The Journal of Infectious Diseases*, vol. 184, pp. 1363–1366, Nov 2001.

[206] A. Mekonnen, M. Legesse, M. Belay, K. Tadesse, W. Torben, Z. Teklemariam, and B. Erko, "Efficacy of Praziquantel against *Schistosoma haematobium* in Dulshatalo village, western Ethiopia," *BMC Research Notes*, vol. 6, p. 392, Sep 2013.

[207] J. H. Kihara, E. N. M. Njagi, E. U. Kenya, M. T. Mwanje, A. E. Odek, G. van Dam, A. I. Kahama, and J. H. Ouma, "Urinary soluble egg antigen levels in *Schistosoma haematobium* infection in relation to sex and age of Kenyan schoolchildren following praziquantel treatment," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 103, pp. 1024–1030, Oct 2009.

[208] C. O. Obonyo, E. M. O. Muok, and P. N. M. Mwinzi, "Efficacy of artesunate with sulfalene plus pyrimethamine versus praziquantel for treatment of *Schistosoma man-*

soni in Kenyan children: an open-label randomised controlled trial," *The Lancet Infectious Diseases*, vol. 10, pp. 603–611, Sep 2010.

[209] G. Degu, G. Mengistu, and J. Jones, "Praziquantel efficacy against schistosomiasis mansoni in schoolchildren in north-west Ethiopia," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 96, no. 4, pp. 444–445, 2002.

[210] H. Namwanje, N. Kabatereine, and A. Olsen, "A randomised controlled clinical trial on the safety of co-administration of albendazole, ivermectin and praziquantel in infected schoolchildren in Uganda," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 105, pp. 181–188, Apr 2011.

[211] A. A. Mohamed, H. M. Mahgoub, M. Magzoub, G. I. Gasim, W. N. Eldein, A. e. A. A. Ahmed, and I. Adam, "Artesunate plus sulfadoxine/pyrimethamine versus praziquantel in the treatment of *Schistosoma mansoni* in eastern Sudan," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 103, pp. 1062–1064, Oct 2009.

[212] J. H. Kihara, N. Muhoho, D. Njomo, I. K. Mwobobia, K. Josyline, Y. Mitsui, T. Awazawa, T. Amano, and C. Mwandawiro, "Drug efficacy of praziquantel and albendazole in school children in Mwea Division, Central Province, Kenya," *Acta Tropica*, vol. 102, pp. 165–171, Jun 2007.

[213] E. Pisani and S. Botchway, "Sharing individual patient and parasite-level data through the WorldWide Antimalarial Resistance Network platform: A qualitative case study," *Wellcome Open Research*, vol. 2, p. 63, 2017.

[214] B. Pedrique, N. Strub-Wourgaft, C. Some, P. Olliaro, P. Trouiller, N. Ford, B. Pécoul, and J.-H. Bradol, "The drug and vaccine landscape for neglected diseases (2000–11): a systematic assessment," *The Lancet Global Health*, vol. 1, pp. e371–e379, Dec 2013.

[215] World Health Organization, "World malaria report 2017," Geneva: WHO, 2017. (Licence: CC BY-NC-SA 3.0 IGO).

[216] Centers for Disease Control and Prevention, "CDC - DPDx - Malaria." Available at: `https://www.cdc.gov/dpdx/malaria/index.html` [Accessed 30 Apr 2018].

[217] World Health Organization, "Guidelines for the treatment of malaria," Geneva: WHO Press, 2015.

[218] N. Fullman, R. Burstein, S. S. Lim, C. Medlin, and E. Gakidou, "Nets, spray or both? The effectiveness of insecticide-treated nets and indoor residual spraying in reducing malaria morbidity and child mortality in sub-Saharan Africa," *Malaria Journal*, vol. 12, p. 62, Feb 2013.

[219] WHO Global Malaria Programme, "Achieving and maintaining universal coverage with long-lasting insecticidal nets for malaria control," World Health Organization, 2017. Available at: http://www.who.int/malaria/publications/atoz/who_recommendation_coverage_llin/en/ [Accessed 24 Jun 2018].

[220] M. A. Phillips, J. N. Burrows, C. Manyando, R. H. van Huijsduijnen, W. C. Van Voorhis, and T. N. C. Wells, "Malaria," *Nature Reviews Disease Primers*, vol. 3, p. 17050, Aug 2017.

[221] Medicines for Malaria Venture, "MMV-supported projects," 2018. Available at: https://www.mmv.org/research-development/mmv-supported-projects [Accessed 24 Jun 2018].

[222] Uniting to Combat Neglected Tropical Diseases, "London Declaration on Neglected Tropical Diseases," 2012. Available at: http://unitingtocombatntds.org/london-declaration-neglected-tropical-diseases/ [Accessed 24 Jun 2018].

[223] Uniting to Combat Neglected Tropical Diseases, "Reaching a Billion – Ending Neglected Tropical Diseases: A gateway to Universal Health Coverage — Fifth progress report on the London Declaration on NTDs," 2017. Available at: https://unitingtocombatntds.org/reports/5th-report/ [Accessed 24 Jun 2018].

[224] M. S. Alilio, I. C. Bygbjerg, and J. G. Breman, "Are multilateral malaria research and control programs the most successful? Lessons from the past 100 years in Africa," *The American Journal of Tropical Medicine and Hygiene*, vol. 71, pp. 268–78, Aug 2004.

[225] IDDO, "Infectious Diseases Data Observatory — Home," 2018. https://www.iddo.org/ [Accessed 24 Jun 2018].

[226] J. B. Buckheit and D. L. Donoho, "WaveLab and Reproducible Research," 1995. Available at: https://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf [Accessed 4 Feb 2018].

[227] C. Tudur Smith, P. Williamson, A. Jones, A. Smyth, S. Hewer, and C. Gamble, "Risk-proportionate clinical trial monitoring: an example approach from a non-commercial trials unit," *Trials*, vol. 15, p. 127, Apr 2014.

[228] W. K. Michener, "Ten Simple Rules for Creating a Good Data Management Plan," *PLoS Computational Biology*, vol. 11, p. e1004525, Oct 2015.

[229] M. Williams, J. Bagwell, and M. Nahm Zozus, "Data management plans: the missing perspective," *Journal of Biomedical Informatics*, vol. 71, pp. 130–142, Jul 2017.

[230] The Global Health Network Editorial Team, "Data Management Systems – Your Views Needed," Global Health Data Management, 2017. Available at: https://globalhealthdatamanagement.tghn.org/articles/data-management-systems-your-views-needed/ [Accessed 4 Jul 2018].

[231] European Spreadsheet Risks Interest Group, "EuSpRIG Horror Stories," 2018. Available at: http://www.eusprig.org/horror-stories.htm [Accessed 30 Jun 2018].

[232] United Nations Statistics Division (UNSD), "Methodology: Standard country or area codes for statistical use (M49) — Geographic Regions," 2018. Available at: https://unstats.un.org/unsd/methodology/m49/ [Accessed 19 Jun 2018].

[233] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, "Data sharing by scientists: practices and perceptions," *PLOS ONE*, vol. 6, no. 6, p. e21101, 2011.

[234] C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett, "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLOS ONE*, vol. 10, no. 8, p. e0134826, 2015.

[235] "Everyone needs a data-management plan," *Nature*, vol. 555, pp. 286–286, Mar 2018.

[236] J. C. Wallis, E. Rolando, and C. L. Borgman, "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology," *PLOS ONE*, vol. 8, no. 7, p. e67332, 2013.

[237] A. Gonzalez and P. R. Peres-Neto, "Data curation: Act to staunch loss of research data," *Nature*, vol. 520, p. 436, Apr 2015.

[238] J. C. Greene, V. J. Caracelli, and W. F. Graham, "Toward a Conceptual Framework for Mixed-Method Evaluation Designs," *Educational Evaluation and Policy Analysis*, vol. 11, pp. 255–274, Sep 1989.

[239] A. Bryman, "Integrating quantitative and qualitative research: how is it done?," *Qualitative Research*, vol. 6, pp. 97–113, Feb 2006.

[240] The World Bank, "Republic of the Congo Overview," World Bank Group, 2018. Available at: http://www.worldbank.org/en/country/congo/overview [Accessed 19 Jun 2018].

[241] The World Bank, "Senegal Overview," World Bank Group, 2018. Available at: http://www.worldbank.org/en/country/senegal/overview [Accessed 19 Jun 2018].

[242] The World Bank, "World Bank Open Data — Data," World Bank Group, 2018. Available at: https://data.worldbank.org/ [Accessed 19 Jun 2018].

[243] The World Bank, "Congo, Rep. — Data," World Bank Group, 2018. Available at: https://data.worldbank.org/country/congo-rep [Accessed 19 Jun 2018].

[244] The World Bank, "Senegal — Data," World Bank Group, 2018. Available at: https://data.worldbank.org/country/senegal [Accessed 19 Jun 2018].

[245] The World Bank, "Data on Statistical Capacity," World Bank Group, 2018. Available at: http://datatopics.worldbank.org/statisticalcapacity/ [Accessed 19 Jun 2018].

[246] The World Bank, "Note on the Statistical Capacity Indicator," World Bank Group. Available at: http://datatopics.worldbank.org/statisticalcapacity/files/Note.pdf [Accessed 19 Jun 2018].

[247] J.-P. Bado, *Médecine coloniale et grandes endémies en Afrique 1900-1960: lèpre, trypanosomiase humaine et onchocercose*. Paris: Editions Karthala, 1996. (Book in French).

[248] Université Cheikh Anta Diop de Dakar, "Historique — 60 ans d'histoire," UCAD, 2012. Available at: https://www.ucad.sn/index.php?option=com_content&view=article&id=155&Itemid=320 [Accessed 20 Jun 2018] (Webpage in French).

[249] WorldWide Antimalarial Resistance Network, "WWARN's West Africa Regional Centre announces new management team — Worldwide Antimalarial Resistance Network," WWARN, 2016. Available at: http://www.wwarn.org/news/news-articles/wwarns-west-africa-regional-centre-announces-new-management-team [Accessed 20 Jun 2018].

[250] World Health Organization, "WHO policy recommendation: Seasonal malaria chemoprevention (SMC) for *Plasmodium falciparum* malaria control in highly seasonal transmission areas of the Sahel sub-region in Africa," WHO, 2012. Available at: http://www.who.int/malaria/publications/atoz/who_smc_policy_recommendation/en/ [Accessed 20 Jun 2018].

[251] FCRM, "Fondation Congolaise pour la recherche médicale — Qui sommes-nous," 2017. Available at: http://www.fcrm-congo.com/about [Accessed 20 Jun 2018] (Webpage in French).

[252] World Health Organization and Medicines for Malaria Venture, "Methods and techniques for clinical trials on antimalarial drug efficacy: genotyping to identify parasite populations," tech. rep., WHO, Geneva, 2008.

[253] J. Green and N. Thorogood, *Qualitative Methods for Health Research*. London: SAGE Publications, 4th ed., 2018.

[254] R. K. Yin, *Case study research and applications: design and methods*. Los Angeles, CA: SAGE Publications, 6th ed., 2018.

[255] O. C. Robinson, "Sampling in Interview-Based Qualitative Research: A Theoretical and Practical Guide," *Qualitative Research in Psychology*, vol. 11, pp. 25–41, Jan 2014.

[256] S. Reeves, A. Kuper, and B. D. Hodges, "Qualitative research methodologies: ethnography," *BMJ*, vol. 337, p. a1020, Aug 2008.

[257] P. Srivastava and N. Hopwood, "A Practical Iterative Framework for Qualitative Data Analysis," *International Journal of Qualitative Methods*, vol. 8, pp. 76–84, Mar 2009.

[258] C. Pope, S. Ziebland, and N. Mays, "Qualitative research in health care: Analysing qualitative data," *BMJ*, vol. 320, pp. 114–6, Jan 2000.

[259] QSR International Pty Ltd, "NVivo — What is NVivo?." Available at: `http://www.qsrinternational.com/nvivo/what-is-nvivo` [Accessed 20 Jun 2018].

[260] G. M. Miiro, O. Ouwe Missi Oukem-Boyer, O. Sarr, M. Rahmani, F. Ntoumi, K. Dheda, A. Pym, S. Mboup, P. Kaleebu, and N. programme, "EDCTP regional networks of excellence: initial merits for planned clinical trials in Africa," *BMC Public Health*, vol. 13, p. 258, Dec 2013.

[261] A. Raja, A. Tridane, A. Gaffar, T. Lindquist, and K. Prabadi, "Android and ODK based data collection framework to aid in epidemiological analysis," *Online Journal of Public Health Informatics*, vol. 5, p. 228, Feb 2014.

[262] O. Maduka, G. Akpan, and S. Maleghemi, "Using Android and Open Data Kit Technology in Data Management for Research in Resource-Limited Settings in the Niger Delta Region of Nigeria: Cross-Sectional Household Survey," *JMIR mHealth and uHealth*, vol. 5, p. e171, Nov 2017.

[263] F. C. Lunenburg, "Managing Change: The Role of the Change Agent," *International journal of Management, Business & Administration*, vol. 13, no. 1, pp. 1–6, 2010.

[264] J. H. Westover, "Managing Organizational Change: Change Agent Strategies and Techniques to Successfully Managing the Dynamics of Stability and Change in Organizations," *International Journal of Management & Innovation*, vol. 2, no. 1, pp. 45–50, 2010.

[265] J. Marques, "Understanding the Strength of Gentleness: Soft-Skilled Leadership on the Rise," *Journal of Business Ethics*, vol. 116, pp. 163–171, Aug 2013.

[266] C. J. Bonk and C. R. Graham, *The handbook of blended learning: global perspectives, local designs*. San Francisco, CA: Pfeiffer, 2006.

[267] A. O. Aragón and J. C. Giles Macedo, "A 'Systemic Theories of Change' Approach for Purposeful Capacity Development," *IDS Bulletin*, vol. 41, pp. 87–99, May 2010.

[268] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules, "Thematic Analysis: Striving to Meet the Trustworthiness Criteria What Is Already Known?," *International Journal of Qualitative Methods*, vol. 16, pp. 1–13, 2017.

[269] Collaboration for Health Policy and Systems Analysis in Africa (CHEPSAA), "Health Policy and Systems Teaching Resources — Introduction to Complex Health Systems." Available at: `http://hpsa-africa.org/index.php/introduction-to-complex-health-systems` [Accessed 20 Jun 2018].

[270] S. Elloker, P. Olckers, L. Gilson, and U. Lehmann, "Crises, Routines and Innovations: The complexities and possibilities of sub-district management," in *South African Health Review 2012/13* (A. Padarath and R. English, eds.), ch. 13, pp. 161–173, Durban: Health Systems Trust, 2013.

[271] J. van Dam, K. Omondi Onyango, B. Midamba, N. Groosman, N. Hooper, J. Spector, G. C. Pillai, and B. Ogutu, "Open-source mobile digital platform for clinical trial data collection in low-resource settings," *BMJ Innovations*, vol. 3, pp. 26–31, Feb 2017.

[272] D. W. Knipe, M. Pearson, R. Borgstrøm, R. Pieris, M. Weerasinghe, C. Priyadarshana, M. Eddleston, D. Gunnell, C. Metcalfe, and F. Konradsen, "Challenges and opportunities of a paperless baseline survey in Sri Lanka," *BMC Research Notes*, vol. 7, p. 452, Jul 2014.

[273] N. Manamley, S. Mallett, M. R. Sydes, S. Hollis, A. Scrimgeour, H. U. Burger, and H.-J. Urban, "Data sharing and the evolving role of statisticians," *BMC Medical Research Methodology*, vol. 16, p. 75, Jul 2016.

[274] WHO TDR, "Development of the TDR global competency framework for clinical research," Geneva: World Health Organization, 2016.

[275] WHO TDR, "Using the TDR global competency framework for clinical research: A set of tools to help develop clinical researchers," Geneva: World Health Organization, 2016.

[276] A. Julé, T. Furtado, L. Boggs, F. van Loggerenberg, V. Ewing, M. Vahedi, P. Launois, and T. Lang, "Developing a globally applicable evidence-informed competency framework to support capacity strengthening in clinical research," *BMJ Global Health*, vol. 2, p. e000229, Mar 2017.

[277] World Health Organization, "World Report on Knowledge for Better Health: Strengthening Health Systems," Geneva: WHO, 2004.

[278] S. R. Hanney and M. A. González-Block, "'Knowledge for better health' revisited – the increasing significance of health research systems: a review by departing Editors-in-Chief," *Health Research Policy and Systems*, vol. 15, p. 81, Dec 2017.

[279] United Nations, "The Millennium Development Goals Report 2015," New York: United Nations, 2015.

[280] B. R. Clark, ed., *The Research foundations of graduate education: Germany, Britain, France, United States, Japan.* Oxford: University of California Press, 1993.

[281] S. Marginson and M. Van Der Wende, "Globalisation and Higher Education," in *OECD Education Working Papers*, OECD Publishing, 2007. Available at: `http://dx.doi.org/10.1787/173831738240` [Accessed 30 Jun 2018].

[282] J. R. Andersen, I. Byrjalsen, A. Bihlet, F. Kalakou, H. C. Hoeck, G. Hansen, H. B. Hansen, M. A. Karsdal, and B. J. Riis, "Impact of source data verification on data quality in clinical trials: an empirical post hoc analysis of three phase 3 randomized clinical trials," *British Journal of Clinical Pharmacology*, vol. 79, pp. 660–668, Apr 2015.

[283] B. Krishnankutty, B. Naveen Kumar, L. Moodahadu, and S. Bellary, "Data management in clinical research: An overview," *Indian Journal of Pharmacology*, vol. 44, p. 168, Mar 2012.

[284] L.-A. McNutt, S. E. Furner, M. Moser, and E. M. Weist, "Applied Epidemiology Competencies for Governmental Public Health Agencies: Mapping Current Curriculum and the Development of New Curriculum," *Public Health Reports*, vol. 123, pp. 13–18, Jan 2008.

[285] S. Basu and C. Roberts, "Towards a Public Health Curriculum in Undergraduate Medicine," *Education for Health*, vol. 25, p. 98, Nov 2012.

[286] L. Allen, J. Scott, A. Brand, M. Hlava, and M. Altman, "Publishing: Credit where credit is due," *Nature*, vol. 508, pp. 312–313, Apr 2014.

[287] E. P. Laabes, R. Desai, S. M. Zawedde, and R. H. Glew, "How much longer will Africa have to depend on western nations for support of its capacity-building efforts for biomedical research?," *Tropical Medicine & International Health*, vol. 16, pp. 258–62, Mar 2011.

[288] D. O. Olaleye, G. N. Odaibo, P. Carney, O. Agbaji, A. S. Sagay, H. Muktar, O. O. Akinyinka, A. O. Omigbodun, A. Ogunniyi, W. Gashau, S. Akanmu, F. Ogunsola, C. Chukwuka, P. I. Okonkwo, S. T. Meloni, I. Adewole, P. J. Kanki, and R. L. Murphy, "Enhancement of Health Research Capacity in Nigeria Through North–South and In-Country Partnerships," *Academic Medicine*, vol. 89, pp. S93–S97, Aug 2014.

[289] J. Meyer, "Qualitative research in health care. Using qualitative methods in health related action research," *BMJ*, vol. 320, pp. 178–81, Jan 2000.

[290] P. Reason and H. Bradbury, eds., *The Sage handbook of action research: participative inquiry and practice*. SAGE Publications, 2nd ed., 2008.

[291] T. Chantler, P. Y. Cheah, G. Miiro, V. Hantrakum, A. Nanvubya, E. Ayuo, E. Kivaya, J. Kidola, P. Kaleebu, M. Parker, P. Njuguna, E. Ashley, P. J. Guerin, and T. Lang, "International health research monitoring: exploring a scientific and a cooperative approach using participatory action research," *BMJ Open*, vol. 4, p. e004104, Feb 2014.

[292] G. Ranmuthugala, J. J. Plumb, F. C. Cunningham, A. Georgiou, J. I. Westbrook, and J. Braithwaite, "How and why are communities of practice established in the healthcare sector? A systematic review of the literature," *BMC Health Services Research*, vol. 11, p. 273, Oct 2011.

[293] M. P. Bertone, B. Meessen, G. Clarysse, D. Hercot, A. Kelley, Y. Kafando, I. Lange, J. Pfaffmann, V. Ridde, I. Sieleunou, and S. Witter, "Assessing communities of practice in health policy: a conceptual framework as a first step towards empirical research," *Health Research Policy and Systems*, vol. 11, p. 39, Oct 2013.

[294] I. Pyrko, V. Dörfler, and C. Eden, "Thinking together: What makes Communities of Practice work?," *Human Relations*, vol. 70, pp. 389–409, Apr 2017.

[295] T. Furtado, S. Franzen, F. van Loggerenberg, G. Carn, S. Grahek, M. McBride, M. Power, J. O'Reilly, B. Savarese, M. A. Snowden, G. Stevens, A. Uys, and T. Lang, "Strengthening Neglected Tropical Disease Research through Enhancing Research-Site Capacity: An Evaluation of a Novel Web Application to Facilitate Research Collaborations," *PLoS Neglected Tropical Diseases*, vol. 8, p. e3225, Nov 2014.

[296] S. Van Tuyl and A. L. Whitmire, "Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia," *PLOS ONE*, vol. 11, p. e0147942, Feb 2016.

[297] A. J. Onwuegbuzie, R. M. Bustamante, and J. A. Nelson, "Mixed Research as a Tool for Developing Quantitative Instruments," *Journal of Mixed Methods Research*, vol. 4, pp. 56–78, Jan 2010.

[298] R. Luyt, "A Framework for Mixing Methods in Quantitative Measurement Development, Validation, and Revision," *Journal of Mixed Methods Research*, vol. 6, pp. 294–316, Oct 2012.

[299] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, pp. 144–151, Jan 2013.

[300] World Health Organization and ENVISION, "Data Quality Assessment (DQA) for NTDs," 2016. Available at: `https://www.ntdenvision.org/resource/publication/data_quality_assessment_dqa_for_ntds` [Accessed 19 Jun 2018].

[301] D. K. de Souza, E. Yirenkyi, J. Otchere, N.-K. Biritwum, D. K. Ameme, S. Sackey, C. Ahorlu, and M. D. Wilson, "Assessing Lymphatic Filariasis Data Quality in Endemic Communities in Ghana, Using the Neglected Tropical Diseases Data Quality Assessment Tool for Preventive Chemotherapy," *PLoS Neglected Tropical Diseases*, vol. 10, p. e0004590, Mar 2016.

[302] The Academy of Medical Sciences, BBSRC, MRC, and Wellcome Trust, "Reproducibility and reliability of biomedical research: improving research practice," tech. rep., The Academy of Medical Sciences, The Academy of Medical Sciences, 2015. Available at: `https://acmedsci.ac.uk/file-download/38189-56531416e2949.pdf` [Accessed 19 Jun 2018].

[303] National Academies of Sciences Engineering and Medicine, *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results*. Washington, D.C.: National Academies Press, Feb 2016.

[304] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, pp. 86–95, Nov 1996.

[305] LimeSurvey, "Professional online surveys with LimeSurvey." `https://www.limesurvey.org/` [Accessed 19 Jun 2018].

[306] Gephi.org, "Gephi: The Open Graph Viz Platform," 2017. `https://gephi.org/` [Accessed 19 Jun 2018].

[307] D. A. Luke and J. K. Harris, "Network Analysis in Public Health: History, Methods, and Applications," *Annual Review of Public Health*, vol. 28, pp. 69–93, Apr 2007.

End