Early development of decision support systems based on artificial intelligence: an application to postoperative complications and a cross-specialty reporting guideline for early-stage clinical evaluation

Baptiste Vasey

Lincoln College

Nuffield Department of Surgical Sciences

Supervised by

Prof. Peter McCulloch
Prof. David Clifton
Prof. Peter Watkinson
Dr. Lauren Morgan

Submitted: Hilary Term 2023

This thesis is submitted to the Nuffield Department of Surgical Sciences, University of Oxford, in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

To Cynthia,

whose love and unconditional support made this work possible.

ABSTRACT

Background: Complications after major surgery occur in a similar manner internationally but the success of response process in preventing death varies widely depending on speed and appropriateness. Artificial intelligence (AI) offers new opportunities to provide support to the decision making of clinicians in this stressful situation when uncertainty is high. However, few AI systems have been robustly and successfully tested in real-world clinical settings. Whilst preparing to develop an AI decision support algorithm and planning to evaluate it in real-world settings, a lack of appropriate guidance on reporting early clinical evaluation of such systems was identified. **Objectives:** The objectives of this work were twofold: i) to develop a prototype of AI system to improve the management of postoperative complications; and ii) to understand expert consensus on reporting standards for early-stage evaluation of AI systems in live clinical settings. **Methods:** I conducted and thematically analysed interviews with clinicians to identify their main challenges and support needs when managing postoperative complications. I then systematically reviewed the literature on the impact of AI-based decision support systems on clinicians' diagnostic performance. A model based on unsupervised clustering and providing prescription recommendations was developed, optimised, and tested on an internal hold out dataset. Finally, I conducted a Delphi process, to reach expert consensus on minimum reporting standards for the early-stage clinical evaluation of AI systems in live clinical settings. **Results:** 12 interviews were conducted with junior and senior clinicians identifying 54 themes about challenges, common errors, strategies, and support needs when managing postoperative complications. 37 studies were included in the systematic review, which found no robust evidence of a positive association between the use of AI decision support systems and improved clinician diagnostic performance. The developed algorithm showed no improvement in recall at position ten compared to a list of the most common prescriptions in the study population. When considering the prevalence of the individual prescriptions, the

algorithm showed a 12% relative increase in performance compared to the same baseline. 151 experts participated in the Delphi study, representing 18 countries and 20 stakeholder groups. The final DECIDE-AI checklist comprises 27 items, accompanied by Explanation & Elaboration sections for each. **Conclusion:** The proposed algorithm offers a proof of concept for an AI system to improve the management of postoperative complications. However, it needs further development and evaluation before claiming clinical utility. The DECIDE-AI guideline provides a practicable checklist for researchers reporting on the implementation of AI decision support systems in clinical settings, and merits future iterative evaluation-update cycles in practice.

RELATED PUBLICATIONS

The content of the present thesis was reported in the following publications. My original contribution to publication n° 6 (Youssef *et al*., 2021) consisted in a piece of code extracting the fraction of inspired oxygen (FiO2) received by patients based on proxy variables and was used in Chapter IV's code.

1. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, *et al*. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022.

2. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, *et al*. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022.

3. Vasey B, Clifton DA, Collins GS, Denniston AK, Faes L, Geerts BF, *et al*. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med.* 2021.

4. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, *et al*. Association of Clinician Diagnostic Performance With Machine Learning–Based Decision Support Systems: A Systematic Review. *JAMA Netw Open*. 2021.

5. Vasey B, Novak A, Ather S, Ibrahim M, McCulloch P. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. *Clin Radiol.* 2022.

6. Youssef A, Kouchaki S, Shamout F, Armstrong J, El-Bouri R, Taylor T, *et al*. Development and validation of early warning score systems for COVID-19 patients. *Healthc Technol Lett.* 2021.

ACKNOWLEDGEMENTS

I would like to thank the following people and organisations who, through their guidance, support and trust, played an important role in the completion of this thesis.

My fiancée Cynthia, for her resolute and unconditional support in good times and in bad, for her patience, for her faith in my ability to complete this work and for all the weekends and evenings sacrificed to the advancement of the thesis. My parents Sylvie and Philippe, for always supporting my choices, for believing in the value of education and for providing me the means to focus on my study. My sister Léa and brother Benoît, for being an indefectible home base and for keeping me on the right track mentally, emotionally and physically. My grandparents Anne-Marie, Jacqueline, Jean-Pierre and Harold, for the values they taught me and for giving me the material security to make ambitious intellectual choices.

The Berrow Foundation for making my dream of a DPhil at Oxford possible by funding my research through the Berrow Foundation Lord Florey scholarship and for offering me a lifelong tie to Oxford in Switzerland. My primary supervisor Prof. Peter McCulloch, for believing in the potential of my research project, for being supportive and available when needed while always providing me scope to explore and grow as an independent researcher, as well as for thought-provoking exchanges way beyond academia. My co-supervisors Dr Lauren Morgan, Prof. David Clifton, and Prof. Peter Watkinson, for welcoming me into their research groups, for offering me world-class research communities to share ideas with, and for guiding me through the most technical aspects of the thesis. The Rector, Fellows and Students of Lincoln College, for offering me a home away from home and amongst the best memories of my student life.

My Master thesis supervisor Dr. Cora Thiel, for teaching me scientific rigor, for guiding my first steps into research, and for inspiring me as scientific role model. The Swiss Study Foundation

and The Mercator Foundation, for broadening my horizons and for giving me the opportunity to develop some of the most important skills needed for this thesis. Ambassador Georges Martin, for showing me the power of diplomacy and for his defining support during the DPhil application process. Prof. Niklaus Labhardt and Prof. Valérie d'Acremont, for making introductions which would change the course of my studies and make this thesis possible. Guillaume Foutry, for his trust, for offering me my first mHealth experience, and for teaching me that a task is always better done than perfect. Dr. Guy Fones, for inspiring me with his ideal of a fair and universal access to healthcare, for introducing me to international health institutions and for teaching me how to operate in a multi stakeholder environment. Dr Irene Bosch and Dr Anuraj Shankar, for supervising my first publication and for preparing me to the work of a DPhil student. Dr Myura Nagendran, for his invaluable help in the development of DECIDE-AI. And last but not least, my friends and colleagues at Oxford, who gave a soul to these DPhil years and whose paths I wish to cross again as many times as possible in the future.

TABLE OF CONTENT

LIST OF FIGURES

List of Figures

LIST OF TABLES

List of Tables

LIST OF ABBREVIATIONS

ACS	American College of Surgeons
ACTA	Applied Cognitive Task Analysis
AE	Autoencoder
AI	Artificial Intelligence
AMLAS	Assurance of Machine Learning in Autonomous Systems
ANN	Artificial Neural Network
AUROC	Area Under the Receiver Operating Characteristic
BSI	British Standards Institute
CCI	Charlson Comorbidity Index
CCRCT	Cochrane Central Register of Controlled Trials
CCRG	Critical Care Research Group
CDSS	Clinical Decision Support System
CIEHF	Chartered Institute of Ergonomics and Human Factors
CNN	Convolutional Neural Network
CONSORT(-AI)	CONsolidated Standards Of Reporting Trials (extension to Artificial Intelligence)
COPD	Chronic Obstructive Pulmonary Disease
CORE-MD	Coordinating Research and Evidence for Medical Devices
COREQ	COnsolidated criteria for REporting Qualitative research
CTx	Core Trainee (in year x)
DECIDE-AI	Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence
E&E	Explanation and Elaboration
EHR	Electronic Health Record
EPR	Electronic Patient Record
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
EWS	Early Warning Score
FDA	(United States) Food and Drug Administration
HAVEN	Hospital Alerting Via Electronic Noticeboard
HF	Human factors
ICTRP	International Clinical Trials Registry Platform

ICU	Intensive Care Unit
IDEAL	Idea, Development, Exploration, Assessment, Long-term study
IEC	International Electrotechnical Commission
IMDRF	International Medical Device Regulators Forum
IQR	Interquartile Range
IRB	Institutional Review Board
ISO	International Standardization Organization
JAFROC	JAckknife Free-response Receiver Operating Characteristic
JTC	Joint Technical Committee
MDR	Medical Device Regulation
MFDS	(Korea) Ministry of Food and Drug Safety
MHRA	(United Kingdom) Medicines and Healthcare products Regulatory Agency
ML	Machine Learning
NEWS	National Early Warning Score
NHS	National Health Service
NICE	National Institute for health and Care Excellence
NIFDC	(China) National Institute for Food and Drug Control
NPV	Negative Predictive Value
NSQIP	National Surgical Quality Improvement Program
OSF	Open Science Framework
PC	Principal Component
PCA	Principal Component Analysis
PMDA	(Japan) Pharmaceuticals and Medical Devices Agency
POCT	Point Of Care Testing
PPI	Patient and public involvement
PPV	Positive Predictive Value
PROSPERO	Prospective Register of Systematic Reviews
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
ROBINS-I	Risk Of Bias In Non-randomized Studies - of Interventions
SaMD	Software as Medical Device
SEND	System for Electronic Notification and Documentation
SME	Subject Matter Experts
SPIRIT(-AI)	Standard Protocol Items: Recommendations for Interventional Trials (extension to Artificial Intelligence)

STARD(-AI)	STAndards for Reporting of Diagnostic accuracy studies (extension to Artificial Intelligence)
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology
STx	Specialty Trainee (in year x)
SVM	Support Vector Machine
T92	Toronto 1992 classification system
TIDieR	Template for Intervention Description and Replication
TRIPOD(-AI)	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (extension to Artificial Intelligence)
WHO	World Health Organization

CHAPTER I

Introduction

I.1 Thesis objectives and overview

The initial objective of this thesis was to develop a decision support system based on a patient similarity approach, to improve the management of postoperative complications on the ward. More specifically, it intended to provide a tool to junior clinicians, allowing them to map a patient currently presenting with complications to a group of clinically similar patients in a dataset. In doing so, the tool would support the junior clinicians in obtaining useful information about their target patient by looking into the clinical trajectories of similar patients in the past.

The first step of such a project was to evaluate the intended end users' support needs when dealing with the target medical condition, in this case surgeons facing postoperative complications on the ward. This was done in Chapter II through a series of 12 interviews using an Applied Cognitive Task Analysis approach. The second step was to review existing literature and proposed models. Building on the results of the interview analysis, a systematic review on the effect of machine learning based decision support systems on clinician diagnostic performance was conducted in Chapter III, with objectives oriented both toward this thesis' specific research questions and currently open questions in the field of clinical AI. The third step was to develop a model able to cluster patients in a clinically meaningful way. This was conducted in Chapter IV and the appropriateness of the clustering evaluated using drug prescriptions and procedure booking as a proxy for clinical similarity, reflecting a common definition of surgical complications according to escalation of care.

The fourth and fifth logical steps would have been to validate the model performance in an external dataset and to evaluate the potential clinical utility of an AI system's prototype in simulated clinical scenarios. However, the start of the COVID-19 pandemic in the spring of 2020, and the subsequent response from the University of Oxford and Oxford University Hospitals NHS Trust, rapidly made it clear that further research with medical staff as participants would not be possible within the timeframe of this thesis. It was therefore decided

to refocus the remaining work on a key methodological gap identified during the systematic review, namely the lack of guidance on how to evaluate AI systems when they are initially implemented at small scale in live clinical settings. Initial clinical evaluation being a crucial step in the development pathway of clinical AI, this refocusing was an anticipation of future problems which would have arisen in future work to move the model from in silico development to bedside application.

The objectives for the remaining chapters of the thesis following this adaptation were to produce, through a robust consensus process, a set of minimum reporting standards for the early-stage live clinical evaluation of decision support systems driven by AI. Because the focus of this guideline was AI systems used as adjunct rather than replacement to human intelligence, a special emphasis was given to human factors considerations at this stage of evaluation. Chapter V presents the design and conduct of an international Delphi process and analysis of the data collected. Chapter VI reports on the finalisation of the guideline through a consensus meeting and the following qualitative evaluation of the draft guideline.

Therefore, the present thesis follows a structure leading from preliminary investigations on a specific application of artificial intelligence in surgery, to the development of a general guideline applicable to most type of AI-based decision support systems in healthcare.

I.2 Background

I.2.1 The problem

Postoperative complications

Postoperative complications are a common problem in surgery and have been shown to be associated with a number of clinical outcomes including increased mortality, reoperation rates, length of stay, admission to critical care, needs of care after discharge, and patient psychosocial distress[1–5]. Reported rates of postoperative complications vary widely depending on speciality, procedure site and definition[2,6–10]. Indeed, the definitions of postoperative complication remain currently heterogenous and the field lacks a unified standardised system to characterise and report them[2], which probably explains some of the variation in complication rates between studies. Two systems are nonetheless more commonly accepted and used than others. First, the Accordion Severity Grading System[11], based on the T92 (Toronto 1992, a four grade classification system developpe by Clavien *et al*.[12]). The Accordion system offers a flexible grading system of four or six categories ranging from mild complication to death, with an expansion possible in the severe complication category, to allow for more granularity in larger studies or studies investigating procedures with higher rates of severe complications. Second, the Clavien-Dindo classification, also based on the T92 classification system, is a nine categories scale ranging from Grade 1 (any deviation from normal postoperative course) to Grad V (death of patient) and focusing on the therapy needed to address a complication. The Clavien-Dindo classification was first published in 2004 and re-evaluated in 2009, without prompting any changes to the scale[13,14]. Of note is also the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database[15], which defines specific complications to be reported by specially trained Surgical Clinical Reviewers in each of the participating centres.

Failure to rescue

Regardless of the definition and classification used, death represents the most extreme grade of complication. In 1992, Silber *et al.* showed for the first time that while complication rates are mainly defined by patient characteristics, the number of deaths amongst patients developing a complication is dependent on hospital characteristics[16]. They concluded that the factors associated with complication were different to these associated with preventing death in patients with complications and named the unsuccessful response to this second set of factors "failure to rescue". Failure to rescue has since become a field of research of its own, attracting growing interest over the years. Subsequent studies from Ghaferi, Birkmeyer and Dimick confirmed the initial observations, showing that whereas complication rates are similar or only slightly higher in the "worst" quintile centres in term of risk-adjusted mortality, the failure to rescue rates are markedly higher in these hospitals compared to the "best" quintile centres[17,18]. Similar results were observed when comparing low-volume to high-volume hospitals[19].

Several human and organisational factors have been shown to play a role in this failure to rescue patients who experience complications. Compliance to guidelines and adherence to surgical checklists were for example reported as decreasing mortality[7,20]. Endall-Gallagher *et al.* and Ludikhuiz *et al.* highlighted a positive association between nursing staff qualification and decrease failure to rescue rates[21,22]. Specific hospital staffing models, and more specifically those with a high number of physicians and nurses per bed, higher staffing level by hospitalists and residents, and more overnight coverage have also been associated with decreased failure to rescue rates[23–25]. Other hospital characteristics, such as the presence of a closed intensive care unit, teaching status, high level of technology and average daily census over 50% of the total capacity, proved to have similar effects[24,25]. The chain of events and factors leading to escalation of care, or lack thereof, also appears to play a dominant role in failure to rescue[26].

Early warning scores

Delay in escalation of care (including the activation of an emergency response team and ICU admission) is commonly cited as having a negative impact on outcomes (mainly mortality and length of stay)[27–31]. Even though such delays in escalation have multifactorial origins, the first step to appropriate escalation is the timely detection of a potential deterioration, which itself relies on regular monitoring[32]. The problem of late or missed detection of patient deterioration prompted the development of numerous Early Warning Scores (EWSs). A systematic review by Gerry *et al.* identified 34 unique EWS, of which 29 were based on statistical methods, three based on consensus, and two modifications of existing EWS[33]. The most common outcomes of interest are death, cardiac arrest and ICU admission, with prediction windows ranging from 12 hours to 30 days. Routinely collected vital signs such as respiratory rate, heart rate, systolic blood pressure, temperature and saturation, complemented by the level of consciousness, are the most commonly used predictors. Given the heterogeneity of outcomes and time horizons used, the review did not provide a quantitative summary of EWS performance. Out of 84 studies evaluating an EWS on an external dataset, 69 assessed model discrimination using the C index (equivalent to the area under the Receiver Operating Characteristic curve). Performance ranged from 0.55 to 0.96.

The UK National Health Service (NHS) is currently using the second version of the National Early warning Score (NEWS 2) to detect inpatient deterioration on its wards. NEWS 2 is a score based on the six predictors already listed above, using death, cardiac arrest and unplanned ICU admission within 24 hours as outcome, and recommended by the Royal College of Physicians[34,35]. The College recommends that every patient with a NEWS 2 ≥ 5 (or ≥ 3 in any single score component) should be reviewed by a physician with competencies in the assessment of acute illness and those with a score ≥ 7 by medical staff with critical care competencies. At the Oxford University Hospitals NHS Foundation Trust, an observation set should be collected from inpatients every six hours and a score generated for each observation set. Although more elaborate approaches have demonstrated better

performance[36] or been developed for specific patient subgroups[37], these have not been deployed at scale yet. An important limitation of most EWS remains the trade-off that needs to be made between precision and recall, which often imply high rates of false positive alerts at level of sensitivity acceptable for clinical use. In a study by Pimentel *et al.*, the positive predictive value of NEWS 2 varied for example between 2.7 and 3.0 for a threshold of 5 in a cohort of hospitalised patients[38]. This low positive predictive value leads in turn to alarm fatigue that undermine these systems' acceptance in clinical settings and can confound the evaluation of their real impact on patient outcomes[39]. Given the current high false alarm rate, a NEWS 2 alarm is thus far from being a clear indication for escalation, and most of the decision-making stays in the hands of the attending clinician.

Extensive research has been done in improving the sensitivity of EWS while reducing the number of false positive alarms reviewed by ward doctors, but much less is known about how to best to ensure optimal response to alarm triggers. Detection of potential deteriorations alone is not sufficient; actual deteriorations also have to be correctly identified, communicated effectively and an appropriate response initiated. Some of the major barriers to correct identification and further escalation were found to be issues with documentation, lack of clinical experience/confidence, lack of standardisation, lack of accountability, overconfidence, fear of the hierarchy, relationship within and between teams, communication problems, high workload, staff organisation, reduced staff availability, lack of essential equipment, organisational processes, and patient variability[40–45].

The introduction of rapid response teams has been shown to have a positive impact on escalation of care but their effect on critical outcomes remains debated[46,47]. Rapid response teams are costly to introduce and only have capacity to review a limited number of patients each day[35]. Furthermore, these interventions stand at the end of the escalation cascade and rely on prior correct identification of a deterioration, appropriate decision to escalate and effective communication, for which little support is currently available. Multiple interventions to improve communication during handover and teamwork more generally have been proposed,

but two systematic reviews on the subject found no robust evidence of efficacy[48,49]. Avenues to assist the initial decision to escalate and selection of immediate diagnostic or therapeutic actions remains largely unexplored. The work of this thesis aims to propose one possible solution to fill this gap, taking a NEWS alarm as starting point for decision support and providing less experienced junior clinicians with recommendations to orient their subsequent actions.

I.2.2 Potential value and applications of clinical Artificial Intelligence

Computerised clinical decision support

Early attempts to automate or "mechanise" differential diagnoses were made before the dawn of medical computing. In 1954, F. A. Nash proposed an "apparatus" made of interchangeable wooden pieces, one for each symptom or sign, linking to different diagnoses. The wooden pieces were chosen according to the patient's presentation and the final diagnostic could be read from the aligned marks[50]. Lipkin and Hardy published in 1958 the description of a system using marginal punched cards to support the diagnosis of haematological disorders. 26 cards punched according to a determined coding system and each representing a known disease were compared at once with a patient case encoded using a system of rods[51].

One of the first article mentioning the potential application of computers to improve medical diagnosis was published in Science in 1959, advocating the suitability of computers to provide advanced reminders of possible diagnoses.[52]. A multitude of computerised decision support systems (CDSS) have since then been developed[53], mostly using a rule-based approach, where the CDSS behaviour is hard coded by its developers. Common areas of applications are patient monitoring, clinical studies ordering, therapy prescription, drug interaction or allergy warnings, preventive care reminders and diagnosis support (including symptom checkers for patient use only). Many of these systems have been integrated into Electronic

Health Record (EHR) systems and are commonly used in practice. Despite this integration and some evidence of effectiveness, the quality of evaluation is often inconsistent, the study results display large variations and traditional CDSS use remains debated, both in term of actual clinical value and efficiency in clinical workflows[53–57]. However, over the past two decades, renewed interest has arisen in computerised decision support due to developments in a related field of computer science.

Artificial intelligence in healthcare

Artificial intelligence (AI) is a field of computer science interested in synthetising intelligent agents. The exact definition of AI varies between authors, but AI is in general described as the property of agents capable of executing tasks otherwise requiring human intelligence. Poole *et al.* give a more precise definition of an intelligent agent as "a system that acts intelligently: what it does is appropriate for its circumstances and its goal, it is flexible to changing environments and changing goals, it learns from experience, and it makes appropriate choices given perceptual limitations and finite computation[58]."

Machine learning (ML) is the subset of AI which enables computers to interpret available data and adapt to a changing environment. It is defined as the ability for a computer to learn from data without being explicitly programmed. This definition derives from A. L. Samuel first mention of machine learning in 1959, describing it as "the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program."[59]

ML has been applied to medical problems since the 1970s, but only recently, and with the development of new mathematical approaches and increased computational power, have these methods claimed to match or surpass human clinician performance in some specific areas[60]. Targeted applications range from screening, triage and diagnosis to quality control and resource use anticipation, through personalised risk prediction and treatment recommendation. Being heavily data-driven, radiology, pathology, oncology and dermatology

are the specialties where most of the ML applications for diagnostic support in medicine have been developed so far and are expected to be the specialties where common use of this technology in clinical settings will be first observed[61,62]. According to Benjamens *et al.*, Arterys Cardio DL became in 2016 the first medical device approved by the U.S. Food and Drug Administration (FDA) to be specifically considered an AI/ML-based technology in the official FDA announcement[63]. IDx-DR, an AI-based diagnostic system aimed at detecting diabetic retinopathy in diabetic patients, followed in April 2018, and was the first fully autonomous "Software as Medical Device" (SaMD) to get FDA approval[64]. The list of FDA approved AI system has been growing ever since, with most instances in the field of radiology, oncology and cardiology[63], even though the development and evaluation of these new products remain mostly unpublished, hence escaping transparent and rigorous peer-review process[65].

The promises of artificial intelligence for decision making in medicine

The promises of AI in medicine are plentiful[62,66,67] and can overall be grouped in four predicates.

1) AI will increase the quantity of information which can be processed to inform medical decision making. This should be understood both in term of the number of patient records with commonly used clinical variables and the breadth of processable variables. It is not uncommon for ML models to learn from many thousands of cases in a matter of hours, whereas most human clinicians will only see this number of patients in their lifetime. Publicly available datasets, such as MIMIC-III or eICU for example, contain 38,597 and 139,367 patient records respectively[68,69]. Such amounts of information simply cannot be processed by a single human brain. With the increasing availability of so-called "omics data" (genomics, radiomics, surgomics, etc) relevant to diagnosis and treatment planning, computer-aided decision support will make it possible, and perhaps even become unavoidable, to integrate and process the whole diversity of data generated in modern healthcare settings. In radiology for example, the use of raw data from scanners and MRIs opens new realms of information processing simply not accessible to the human eyes[70,71]. In the operating theatre, the integration of

patient, surgeon, environment, and process data will allow more granularity in the analysis of surgical outcome[72]. Where human clinicians resort to subspecialisation, AI systems could potentially integrate data from all medical specialties at once.

2) AI will improve the quality of medical decision making. Given the importance of pattern recognition in the medical practice, it is hoped that the integration and analyses of a large amount of data will lead to more accurate predictions or recommendations, or at least outputs as good as the ones of the best human clinicians. Esteva *et al*. were amongst the first to demonstrate expert level classification of skin lesions using a deep neural network[73]. Diabetic retinopathy screening also benefited from early development of ML algorithms with, for example, Ting *et al*. presenting a deep learning model for diabetic retinopathy and related eye diseases recognition based on retinal images and trained on more than 70,000 images from 14,880 patients[74]. A systematic review by Liu *et al*. recently concluded that AI systems match human expert performance on specific tasks, in this case diagnosis from medical imaging[60]. Even if not exceeding the performance of experts, bringing decision-making to their level would already be a considerable gain for patients, who are often attended by more junior clinicians.

3) AI will allow more personalised medicine. Modern medicine is based on clinical evidence. However, this evidence is generated by a subset of the population, not always representative of the whole, with for example women often underrepresented in clinical trials[75–78]. Moreover, even amongst the most studied subgroups, genetic differences and other patient characteristics can influence the risk of developing a condition or a response to a treatment. For example, it has been shown that genetic mutation can influence the pharmacological response to some treatments[79–81]. If considered during the AI training process, these genetic variations could be identified and a potential treatment adapted. One of the most advertised promises of clinical AI is therefore to provide, through the integration of granular details and large numbers of instances, recommendations tailored to patient specific characteristics, moving from population-level to patient-level evidence.

4) AI will optimise processes. In healthcare systems under constant resource strain, enabling smarter workflow and more timely decision making could also make a consequential difference to outcomes. Early detection of deterioration, bed occupancy anticipation, case review prioritisation, or double reader replacement are amongst many examples of such process optimisation. The sheer speed of data analysis would also be impacted, as demonstrated for example by a 3D convolutional neural network interpreting certain head-CTs 150 time faster than radiologists (even if yielding poorer accuracy in this case).[82]

Current limitations of artificial intelligence for clinical decision making

However, the field of AI is at present still far from displaying its full potential and many hurdles remain to be overcome, challenging the optimism of AI enthusiasts and suggesting more caution as to what the real impact of AI technology in medical decision making will be. Even if some successful examples have attracted attention, the actual added value of complex ML models compared to traditional statistical approaches[83] and the quality of the evaluation supporting these claims are still contested[84–86]. Here is a non-exhaustive list of current issues regarding the development and implementation of AI systems in healthcare.

First of all, every AI system requires data to train on and perform its task. To train efficient models, researchers need high-quality and often large datasets and issues around data quality and privacy are central in AI system development. Data quality can vary greatly between datasets and depends on various factors such as the devices used to generate the data, the way they are entered in the dataset, the population they are collected from, the setting in which they are extracted, the practitioner collecting them or the completeness of the data collection. For example, hospital acquired data are more likely to over-represent the sickest patients, billing codes pathologies that are well compensated, and information gathered through wearable sensors the rich and healthy.[87] Data linkage is also often a problem as pseudonymised databases focusing on the primary research questions they were developed for may be impossible to re-associate at patient level for broader application. Database linkage

across centres is yet another challenge due to differences in acquisition settings, formatting, storing, or legal requirements to access them. Mathematical solutions, such as federate learning, are being developed to overcome some of these problems[88], but they remain a major barrier to the constitution of very large and diverse datasets.

Second come the problems of transparency, explainability and interpretability, also often referred to as the "black box" problem of AI. Although often used interchangeably, these three are distinct concepts. Transparency relates to the ability for the public to openly access all the information needed to scrutinise and appraise the model, including the code of the algorithm, the data it was trained on, and how it was evaluated[89]. Explainability is the ability to communicate in a comprehensible way how a model deals with the input data and produces its output. Interpretability is probably the most sensitive aspect and define "the degree to which a human can understand the cause of a decision"[90]. A thought-provoking alternative is to define interpretability as "the degree to which [an expert] human can consistently predict the model's result"[91]. The AI black box poses problems in terms of integration of an output within the broader clinical context, of attribution of liability, of bias propagation, or of scientific understanding in the prospect of generating new knowledge. The question of the interpretability of AI models remains quite divisive, though, with some authors stating that transparency, explainability and interpretability are conditions *sine qua non* for the implementation of AI systems in healthcare settings[67,92], and other arguing that fully interpretable AI systems are a myth, can reduce performance and that after all, if a system robustly demonstrates an improvement in patient care, it is only of lesser importance to understand exactly how it does it[93,94].

Third, there are the intrinsic limitations related to the training data. Indeed, ML models can only know as much as is encoded in the data they learn from, conditioned to the task they have been trained for. Etymologically speaking, from its Latin roots *inter* and *legere*, "intelligence" means to read between or to choose between. Often the solution to a clinical situation is not clearly defined by existing evidence or past experience and necessitates the

navigation of a grey area by reading and connecting between the closest available pieces of knowledge, a feat that human intelligence manages quite well, but that artificial intelligence still often struggles with. This limitation of ML models comes not only from the granularity and richness of the data they are trained on, but also from the way ground truth is attributed to them. Gold standards are not always available, sometimes only partially accurate and mainly represent the current state of knowledge, rather than an absolute truth. Solutions to this problem have been proposed, such as detecting deviations from the norm, rather than identifying specific pathological findings, but these remain restricted to specific applications[95].

Fourth, clinical AI raises a series of ethical consideration. Through their amplifying effect, AI systems have the potential to magnify biases already imbedded in healthcare systems and therefore in historical data. For example, it has been shown that an algorithm widely used to identify patients eligible for high-risk care management based on future occurring costs systematically under-evaluated the risk for Black patients because these patients usually had more restricted access to care and thus generated lower healthcare costs[96]. Biases are also present in the distribution of data in training sets, where minorities are often underrepresented. If trained on non-representative datasets, AI systems are likely to underperform when applied to patients from the underrepresented groups, sometimes further amplifying already existing disparities. This is for example the case of some dermatology databases used to train models detecting skin cancer. The training sets being skewed towards lighter skin tone on the Fitzpatrick scale, the performance of the models trained on these datasets are often better on similar skin tones than on darker ones[97,98].

Fifth, there is the broad topic of AI systems integration within the existing clinical settings and workflows. This includes amongst others: technical issues such as connectivity to the local data infrastructure; adaptation to clinicians' real-life support needs, issues relating to the differences between the training condition and use conditions; user related issues such as training, usability, trust and acceptance; patient related issues such as their attitude toward AI-driven decision making impacting their health. Although often thematised[62,67,87,99], relatively

little research on these aspects (in an AI-specific context) has been published in the medical literature so far, with some examples nonetheless yielding interesting results and directions for further studies[100–103].

Last but not least comes the issue of evaluation. Most of the algorithms developed so far have only being tested on retrospective datasets or datasets prospectively acquired in strictly controlled research settings. To become part of common care practice, they will have to be proven effective in real-world clinical settings[66,104–106], ideally in situations where the AI system outputs have an actual influence on patient care, and through randomised controlled trials where appropriate[99,107]. Pre-clinical algorithmic performance alone doesn't necessarily mean impact on patient outcomes, as exemplified by several trials showing little or no improvement in clinical outcomes for AI systems which looked promising during preclinical studies[108–110]. It also cannot properly assess the risks for patients created by live use of the AI system. However, the number of robust prospective trials remains small compared to the total number of proposed AI systems and the claims made by their authors[84,111]. This phenomenon is only exacerbated by the discrepancy between regulatory requirements and scientific expectations, allowing many AI systems to enter the market without any robust clinical studies[63,112,113].

For these reasons, and despite all the current and predicted progress of clinical AI, it is unlikely that human physicians will disappear from the medical decision making process in the near future[66]. As long as the responsibility and liability for patient care remains with clinicians, the human perception of a problem, and their decision regarding the most appropriate solution will probably remain the main factors influencing patient outcomes, not the stand-alone recommendations of AI systems[85,105,114,115]. This is why a combination of AI and human physicians, also referred to as augmented intelligence, might be the most promising way to improve patients care in the short- to mid-term[87]. In a field where most projects still aim mainly at outperforming human physicians, a few successful examples have shown that supporting human users' decision-making might actually yield better results than replacing it[116,117]. Steiner *et al*. argue for example that human-computer interaction in lymph node metastasis detection

is an efficient way to combine the algorithm's high sensitivity with clinicians' good specificity[116]. Such human-AI collaboration could also offer practical solutions to some of the limitations noted above, such as the risk of unexpected algorithm outputs or difficult integration in existing pathways, as well as providing a more broadly accepted way to introduce this technology into clinical settings at the beginning. However, little data exist on the actual impact of AI recommendations on human performance and, inversely, on the impact of a human second read on AI performance.

The work of this thesis therefore focuses on AI systems developed to augment rather than replace clinician intelligence. The addition by design of a human in the loop offers many opportunities to create a system better suited to later implementation but also introduces several challenges, which need to be considered carefully. A field of research, human factors, focuses on investigating these challenges and finding solutions to the issues they raise.

Human factors in healthcare and clinical AI

Human factors (HF), also called ergonomics, is described by the International Ergonomics Association as "the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimise human well-being and overall system performance"[118]. In an explanation adopted by the World Health Organization in one of its technical documents[119], three domains of system design are identified: physical, cognitive and organisational. Physical ergonomics relates to the "human anatomical, anthropometric, physiological and biomechanical characteristics" of a user, cognitive ergonomics to their "mental processes, such as perception, memory, reasoning, and motor response", and organisational ergonomics to the "organisational structures, policies, and processes" in which a task takes place[118]. In summary, HF comprises a set of bidirectional physical, cognitive and environmental interactions impacting human performance within a system.

The study of the discipline has long been applied to safety-critical industries, such as the military, aviation and energy sectors. One of the seminal papers in the field is Bainbridge's 1983 *Ironies of Automation*[120]. It describes the ironies of an automated industry, in which human inputs, and hence the opportunity for learning and training, are less and less frequent, yet much more specialised and impactful when needed. Since then, human factors specialists have developed a wide range of methods to analyse the interactions of users with and within a system, ranging from cognitive task analysis to team assessment and interface analysis methods[121]. Human factors as a discipline has also grown an intricate relationship to safety, with a strong emphasis on developing safer, as well as more efficient, systems.

Healthcare is also a safety-critical sector and unsurprisingly human factors research is being conducted in this field, although belatedly and at smaller scale. Current research includes studies of the usability of medical devices or information systems[122], HF for patient safety[123], and a focus on teamwork and non-technical skills[124,125]. Both the International Ergonomics Association and its British counterpart the Chartered Institute of Ergonomics and Human Factors (CIEHF) have dedicated sections or committees for HF in healthcare[126,127]. In the regulatory sphere, HF evaluation of new medical devices is mandated both in Europe and North America[128–130] and several specific guidelines have been developed by the International Standardization Organization (ISO)[131–133].

Over the past years, experts have started to subspecialise in HF for clinical AI. Asan *et al*. retrieved 48 studies investigating HF in the context of AI in healthcare. User perception, device usability, cognitive workload, and user's trust in AI were the most reported subjects[134]. Sujan *et al.* published a list of HF challenges likely to emerge as more AI systems are incorporated into clinical pathway, such as automation bias, inappropriate handover procedures, limited situational awareness and modified patient-practitioner interactions[135]. A CIEHF special interest group recently published a white paper describing current issues in the HF evaluation of clinical AI, such as the tendency of developers to "design out" human errors rather than understanding them and the limited consideration of cultural and social aspects in the human-

AI interaction[136]. The white paper also presents important applications of HF methods specific to healthcare AI, highlighting avenues of research for the near future. Output explanation and training needs are listed in addition to other concepts already mentioned in this paragraph.

I.2.3 Methodological gap in the evaluation of clinical Artificial Intelligence and opportunities for new guidelines

AI pre-clinical and clinical evaluation

Once developed, any medical devices, AI-based decision support systems being no exception, need to be evaluated. The evaluation constraints faced by developers of medical devices are manyfold. First, they have to abide by regulatory requirements if they one day wish to commercialise their products. Then, they have to obtain authorisation from ethics committees or Institutional Review Boards (IRBs) to access clinical data or clinical settings. Economic limitations and competitive pressure also have to be considered in the design of the evaluation plan. Last but not least, they need to follow the rules of scientific evaluation if they want their product to be accepted by healthcare professionals used to evidence-based medicine. These constraints are not always aligned and developers have to find a fine balance between competing sets of interests.

In the regulatory context, AI systems belong to the broader category of software as medical devices (SaMDs). In Europe, regulatory requirements are defined by the Medical Device Regulation (MDR), or more accurately the Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices, dated 5 April 2017[128]. European guidance specific to clinical AI evaluation has not yet been published, but several projects are underway, including the Coordinating Research and Evidence for Medical Devices (CORE-MD) initiative[137]. Furthermore, any future European regulation will probably be based on the High-Level Expert Group on Artificial Intelligence's Ethics Guideline for Trustworthy AI[92]. In the United States, the Food and Drugs Administration FDA regulates medical devices and has

started to reflect on dedicated evaluation models more adapted to clinical AI[138,139]. In the United Kingdom, the Medicines and Healthcare products Regulatory Agency (MHRA) has launched a large change programme for the regulation of software and AI as medical devices[140] and just concluded a consultation on the future of medical devices regulation, with a section dedicated to SAMDs[141]. The Chinese National Institute for Food and Drug Control (NIFDC), the Japanese Pharmaceuticals and Medical Devices Agency (PMDA) and the Korean Ministry of Food and Drug Safety (MFDS) are also playing an increasing role in the regulation of clinical AI technology and deserve to be listed here as well. At the international level, the International Medical Device Regulators Forum (IMDRF) sets the general standards, which are translated into regulations by national agencies. Over the past decade, the IMDRF has published a series of guidance documents on SaMD risk classification[142], SAMD clinical evaluation[143], and more general updates on clinical evaluation[144] and clinical investigation[145]. The IMDRF has an ongoing working group tasked to achieve an aligned approach to the management of AI- based medical devices, which has already produced a document on terminology, currently open to consultation[146]. At multilateral policy level, the World Health Organization (WHO) has also recently published its *Ethics and governance of artificial intelligence for health* guidance, which endorses a set of six key ethical principles for the use of AI for health. This document is primarily aimed at health ministries but provides as well practical advice for technology developers and healthcare providers[147].

Although regulatory requirements should ideally align with robust scientific evidence generation, this is unfortunately not always the case, as demonstrated by the considerable number of FDA-approved AI systems with no available clinical studies, let alone peer-reviewed publication of their resuts[63,112,113]. To fill this gap in guidance, numerous academic initiatives have emerged (see section “Existing guidelines and guidelines under development”), including many opinion pieces and a few large consensus processes, but no unified development pathway has been widely accepted to date.

Nonetheless, an overall trend of stage-based evaluation has started to crystallise itself from the literature over the past years[148–152]. Like the now accepted development pathway for drugs, it is broadly divided into a pre-clinical stage, a clinical stage (that some authors further divide into an offline or *shadow mode* stage, a small-scale or formative stage and a large-scale or summative stage), and long-term follow-up, also named algorithmovigilance by analogy[153]. Although no harmonised framework exists yet, key components of AI evaluation, such as performance testing on external datasets, usability testing, clinical integration assessment, comparative outcome analysis or error analysis, feature in most proposed pathways. Divergences of opinion remain mainly about the timing and depth to which these components should be evaluated.

Existing guidelines and guidelines under development

To help researchers and developers navigate the complexity of AI evaluation throughout theses different stages and across requirement sets, several guidelines have been proposed, both to facilitate regulatory compliance and to improve the quality of scientific evidence generation, sharing, and appraisal.

These initiatives can be grouped into four main categories: methodological guidance, quality standards, reporting guidelines and risk of bias assessment/quality control tools. Methodological guidance has so far come from small expert groups, if not single research laboratories. McCradden *et al*. proposed a 3-phase approach to strengthen the ethical integration of AI through clinical research[148]. Park *et al*. proposed a 5-stage approach (including a preclinical stage 0) echoing the current development pathway for drugs[151]. Sendak *et al.* published a 4-phases translational pathway with an emphasis on integration and scaling up of the AI intervention[149]. Higgins *et al.* described a 3-step framework to take AI systems from “bit to bedside”, with a focus on clinical needs and risks[150]. Further methodological guidance was produced for specialty-specific application, such as FUTURE-AI in medical imaging[154], PRIME in cardiology[155], or recommendations by Schwendicke *et al*. in dentistry[156].

Of a broader scope but nonetheless applicable to clinical AI, Hawkins *et al.* have published the Assurance of Machine Learning in Autonomous Systems (AMLAS) guidance with a focus on safety and risk management[157].

Quality standards are developed by recognised standardisation organisations, such as the British Standards Institute (BSI), the International Organization for Standardization (ISO), or the International Electrotechnical Commission (IEC). To date, standards relevant to clinical AI development and evaluation include: Information technology – artificial intelligence - Artificial intelligence concepts and terminology (ISO/IEC 22989:2022)[158]; Framework for artificial intelligence (AI) systems using machine learning (ML) (ISO/IEC 23053)[159]; Medical devices - Application of risk management to medical devices (ISO 14971:2019 and ISO/TR 14971:2020)[160,161]; Medical device software - Software life cycle processes (IEC 62304:2006)[162]; Ergonomics of human-system interaction - Usability methods supporting human-centred design (ISO/TR 16982:2002)[132]; Medical devices - Part 1: Application of usability engineering to medical devices (IEC 62366-1:2015)[163]; Medical devices - Part 2: Guidance on the application of usability engineering to medical devices (IEC/TR 62366-2:2016)[164]; Medical devices. Application of usability engineering to medical devices (BS EN 62366-1:2015+A1:2020)[165]; Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (ISO 9241-11:2018)[131]; Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems (ISO 9241- 210:2019)[133]; Clinical investigation of medical devices for human subjects - Good clinical practice (ISO 14155:2020)[166]. The Joint Technical Committee (JTC) ISO/IEC JTC 1/SC 42 and the BSI have published other standards on AI whose full lists can be found on their webpages (www.iso.org/committee/6794475.html; www.bsigroup.com/en-GB/standards/). Further AI-specific standards are also under development such as the Validation framework for the use of AI within healthcare (BS 30440) and Artificial intelligence – Quality evaluation guidelines for AI systems (ISO/IEC AWI TS 5471). A comprehensive list of upcoming ISO/IEC and BSI standards on AI can also be found on the ISO/IEC JTC 1/SC 42 and BSI webpages, respectively.

AI-specific reporting guidelines are still few in numbers but are often derived from large consensus statements based on Delphi studies. The pioneers in the field were CONSORT-AI and SPIRIT-AI, focusing on the reporting of randomised controlled trials involving an AI-based intervention, and their protocols, respectively[167,168]. Further announcements have been made, although the final guidelines have not been published yet. This is the case for TRIPOD-AI[169], an extension of TRIPOD[170] for the reporting of the development and validation of prediction models based on AI, STARD-AI[171], an extension of STARD[172] for the reporting of diagnostic accuracy study of AI systems, and the Computer Vision in Surgery International Collaborative[173] for AI-assisted surgical computer vision. In the group of reporting guidelines not based on large consensus statement, MI-CLAIM[174], MINIMAR[89] and CAIR[175], three cross-specialty guidelines focusing on model design, study population and generalisability of the output, can be mentioned, as well as CLAIM[176], for the reporting of AI models in medical imaging. Also worth notice is a proposal by Sendak *et al.* to develop standardised ML model fact labels for the quick description of AI systems to clinical end users[177]. The EQUATOR network (www.equator-network.org) remains the source of choice for up-to-date information about existing and upcoming reporting guidelines, whether AI-specific or more general.

Finally, risk of bias and quality assessment tools have had mixed origins. The upcoming PROBAST-AI[169], an extension of PROBAST[178], will be based on a large expert consensus and focus on the risk of bias evaluation for machine learning based prediction model studies. CLEAR Derm is a checklist for the evaluation of image-based AI studies in dermatology and was also based on a formalised expert consensus approach, although with smaller number of participants[179]. Petzold *et al*. gathered extensive feedback from different stakeholder groups to develop an AI-specific extension to the OSCAR-IB consensus-based quality control criteria for optical coherence tomography[180]. A few other groups have published recommendations for the quality assessment of AI studies based on expert opinions, such as Vollmer *et al*. and their 20 critical questions on transparency, replicability, ethics and effectiveness[181], or Scott *et al.* with their checklist for assessing the suitability of ML in healthcare[182].

The number of guidelines and recommendations is constantly expanding and the present listing is probably not an exhaustive depiction of the currently published guidance landscape. However, this overview allows some useful considerations. When synthesising existing or upcoming guidelines, comparing them to the established drugs development pathway, and considering the current blockade at the translation from pre-clinical to clinical studies[111], one observation stands out. There is hardly any guidance for the early stage of clinical evaluation, when AI systems are deployed for the first time in the conditions of their intended use, with a direct influence on patient care. Inadequate evaluation at the early stage of clinical implementation is a serious problem, as it will very likely complicate the conduct of robust evaluation further downstream in the translation pathway and jeopardise attempts to acquire definitive evidence of clinical effectiveness.

The IDEAL Framework and related complex intervention evaluation framework

To understand the relevance of this early stage of clinical evaluation, a comparison with innovations sharing similar characteristics can be helpful. When considering the specificities of clinical AI evaluation, it appears that AI systems display many features of so-called complex interventions. Complex interventions are described as “interventions that contain several interacting components”, with or without a variation in the number and difficulty of behaviours required by those delivering or receiving the intervention, in the number of groups or organisational levels targeted by the intervention, in the number and nature of outcomes, and in the degree of flexibility or tailoring of the intervention permitted[183]. Their evaluation does not fit into the classic research methodology paradigms. The Medical Research Council therefore produced a dedicated guidance for the development and evaluation of complex intervention. The original guidance from 2000[184] has been amended twice already, with the latest version dating from 2021[183,185]. Briefly summarised, the guidance suggests broadening the scope of evaluation beyond outcomes measurement, integrating consideration about the intervention’s integration in the implementation context, its value compared to the resources needed, and its

impact on real-world decision-making. It highlights the trade-off existing between "precise unbiased answers to narrow questions and more uncertain answers to broader, more complex questions" and encourages authors to choose the questions asked based on their usefulness to decision-makers, rather than on the prospect of giving precise answers. It also recommends a phased approach with opportunities for iterative improvement between each stage.

Complex interventions vary in form and nature, but one type, the development of new surgical procedures, shares many parallels with AI systems and has methodological guideline readily available. For example, differences amongst users and contexts, in both surgery and clinical AI, create heterogeneity in outcomes but also in opinions on the applicability of the intervention to different patient groups and situations. These differences need to be resolved to arrive at an agreed definition of the intervention and its indications. Surgical complex interventions also inevitably require a stage of iterative modification to adapt them to the realities of clinical practice. Likewise, AI systems will be modified during their life cycle, either through conscious design upgrade or automated self-learning updates. If these changes are not properly documented information will be lost which may frustrate later attempts at evaluation. When learning a new procedure or implementing a new AI system, user learning curves are inevitable too. If not dealt with properly, they are a powerful confounder which usually reduce the apparent efficacy of the intervention in RCTs.

The IDEAL framework was originally developed by the Balliol Collaboration in 2009, before being updated ten years later in 2019[186,187]. The framework describes a five-stage approach to the development and evaluation of new surgical procedures, accounting for the specificities mentioned above. In stage I (first-in-human or Idea), the procedure is tested on one highly selected individual and can be regarded as the proof of concept. During stage IIa (Development), the procedure is iteratively refined by the same group of surgeons who developed it, once again on a strictly selected group of patients. Each modification is recorded and their impact on the outcomes carefully analysed until the procedure reaches stability. Stage IIb (Exploration) is the preparation to larger, summative trials. The indications for the

new procedure are broadened and more surgeons are invited to use it. Important outcomes of this stage are multi-centre agreement on the procedure execution, definition of quality standards and description of the operative indications. In stage III (Assessment), large comparative studies, ideally randomised-controlled, are conducted to demonstrate the procedure's effectiveness. The final stage IV (Long-term follow up) is the surgical equivalence to pharmacovigilance, at which registries should be created to record adverse events and detect more subtle signal during wide-scale use of the procedure.

Since the original publication, several extensions of the IDEAL framework have been developed and published. IDEAL-D proposes an adaptation to medical devices, introducing a pre-clinical stage (stage 0) and suggesting a merger of stage IIa and IIb given the difference in the refinement process between surgical procedure and medical devices[188]. A specific consensus statement was later produced to define the key components of the newly created stage 0[189]. Given its focus on medical devices, IDEAL-D is of particular relevance to the evaluation of AI systems in clinical settings. Specialty-specific extensions have also been produced such as IDEAL-Physio for physiotherapy intervention evaluation[190] and R-IDEAL for radiation oncology[191]. More recently the IDEAL reporting checklist[192] was developed to provide additional guidance to authors at the time of submission and addressing all IDEAL stages except stage III, already covered by the CONSORT checklist as this stage would ideally be a randomised controlled trial[193,194]. The parallels between IDEAL Stages 1 to 2b and the needs of early-stage evaluation of AI systems make the IDEAL reporting checklist a valuable model which could guide efforts to develop a dedicated guideline for the reporting of early clinical evaluation of AI systems.

This thesis hence takes place in a context where surgical complications and failure to rescue remain major drivers of morbidity and mortality, AI has shown promising results in data analytics, opening the way to broader application in healthcare, but a barrier still exists at the stage of translation from pre-clinical development studies to clinical impact demonstration, probably due to incomplete evaluation methodology and guidance.

CHAPTER II

Clinician cognitive process and desired computerised support needs when facing postoperative complications on the ward – a qualitative study

II.1 Introduction

Progress in healthcare digitalisation, computing power and machine learning models has opened a new era of advanced information processing, offering clinicians extended opportunities to acquire and analyse patient data. Clinical decision support systems (CDSS) now have the potential to go beyond the simple display of routinely collected information, hard coded recommendations or set guidelines. In surgery, as in most other specialties, adequately designed CDSS could enhance physicians' understanding of a case and improve their decision-making by providing complex, tailored support. However, the exact support needs of surgeons facing postoperative complications on the wards have not been studied in detail so far, and neither have their expectations regarding possible CDSS outputs. This is despite numerous authors highlighting synergetic human-computer interaction and a good integration in practitioners' workflow as essential characteristics of any future efficient CDSS[62,66,67,106].

Several studies have investigated physicians' information display needs in order to appropriately design software in settings such as general wards[195,196], the operating room and post-anaesthesia care unit[197], and the neonatal ICU[198]. However, these studies focused on single variables (e.g. heart rate, urine output, admission status) and did not investigate the complex information needs linked to a specific cognitive process, like for example the management of a postoperative complication. In surgery, attempts have been made to better understand the decision-making process, mostly in the domain of surgical education[199,200] or for specific procedures[201,202]. To the best of my knowledge, no such efforts have been carried out regarding postoperative complications.

A recognised approach to investigate cognitive processes, such as decision-making, of specific professional groups is cognitive task analysis. Cognitive task analysis is composed of a variety of methods aimed at "describing and representing the cognitive elements that underlie goal generation, decision making, judgment, etc."[203]. By deconstructing cognitive processes into lower-level elements and strategies (e.g. situation assessment, perceptual

distinction or critical cues identification), it allows participants to express knowledge which can otherwise be difficult to verbalise, and is often used to develop training scenarios or improve interfaces. One of these methods is the applied cognitive task analysis (ACTA) methodology, developed by Militello *et al.* in 1998 as part of a project with the Navy Research and Development Center and described by its authors as "a set of streamlined cognitive task analysis tools that have been developed specifically for use by professionals who have not been trained in cognitive psychology, but who do develop applications that can benefit from the use of cognitive task analysis"[203]. It aims at understanding the cognitive elements underlying decision-making and goals generation in a specific situation. To this end, it uses a four-step approach, namely (i) the creation of a task diagram, (ii) knowledge audits, (iii) simulation interviews, and (iv) the creation of a cognitive demands table. Interviews with subject matter experts (SMEs) are the primary mean of data collection.

Alternative cognitive task analysis methods such as the critical decision method[204] or cognitive walkthrough technique[205] exist to analyse professionals' cognitive processes. In the context of this study, ACTA was preferred for two reasons. First, it was validated for use by system designers without training in cognitive psychology or extensive human factors research experience. Second, the method is design to identify how experience and expertise are used to perform a task, which is particularly well suited to the secondary objective of the study, namely to highlight any differences in clinician's approach to the management of surgical complications based on their level of experience.

The following study on clinician cognitive process and desired computerised support needs when facing postoperative complications on the ward was designed to qualitatively: (i) describe the perceived cognitive challenges faced when attending patients presenting with postoperative complications on the ward; (ii) highlight the differences between junior and senior clinicians in terms of cognitive challenges and management strategies; and (iii) produce a list of support modalities desired by junior surgeons to help bridging the gap between junior and senior management of surgical complication on the ward.

II.2 Methods

This study was a series of single time point semi-structured interviews based on the ACTA methodology[203]. An inductive thematic analysis was used to analyse the interview transcripts and produce the cognitive demands table. The study is reported according to the recommendation of the consolidated criteria for reporting qualitative research (COREQ)[206]. The University of Oxford Clinical Trials and Research Governance study classification group reviewed the study protocol (**Annex II-1**) on October 22nd 2020, and considered the proposed study as service development/service improvement and waive the requirements for ethics review and approval.

II.2.1 Participants

Inclusion and exclusion criteria

Participants were recruited according to two inclusion criteria (both had to be fulfilled): 1) Participants were employee of an NHS Trust; and 2) Participants were either: core surgical trainees (CT1-2), surgical specialty trainees in year 3, 4, 6, 7 or 8 (ST3-4, ST6-8), consultants in general surgery or a general surgery subspecialty, specialty trainees in year 6, 7 or 8 (ST6-8) in perioperative medicine/intensive medicine, or consultants in perioperative medicine/intensive medicine. **Figure II-1** shows an overview of the current training pathway in the NHS for illustration. There were no exclusion criteria. However, the investigator retained the right to discontinue a participant from the study at any time for significant non-compliance with the study requirements or an obvious lack of consideration in their answers. Participants were allocated to one of two groups, based on their level of clinical experience:

1. “junior”: core surgical trainees (CT1-2) or specialty surgical trainees (ST3-4);
2. “senior”: specialty trainees (ST6-8), fellows or consultants.

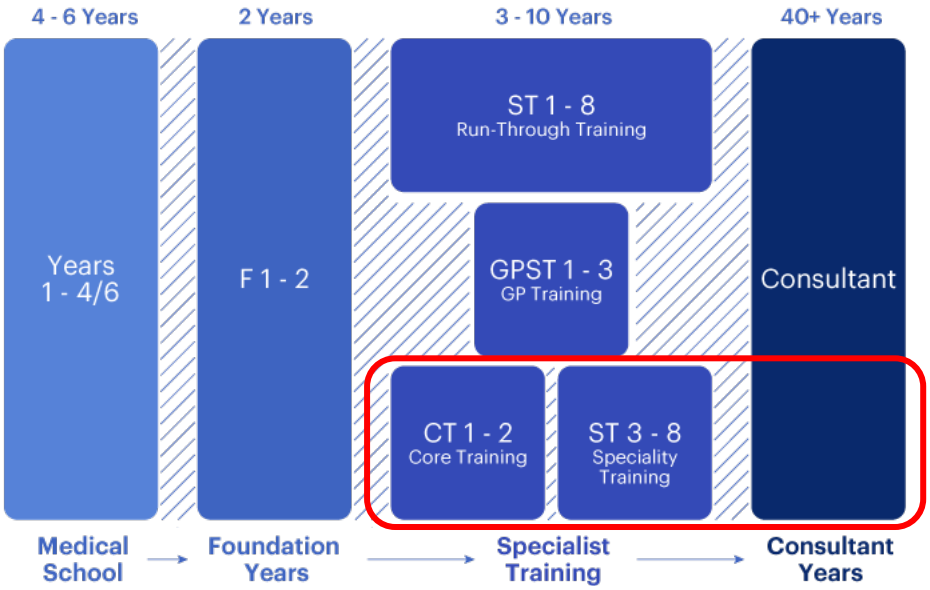


Figure II-1 : NHS clinical training pathway. The red quadrangle highlights the training level included in the study. F = foundation; CT = core training; ST = specialty training; GP = general practitioner.
Figure adapted from https://medmentor.co.uk/blog/what-happens-after-medical-school-in-the-uk (last opened on January 4th 2023).

Recruitment

Participants were recruited from 3 different hospitals of the Health Education England South East region (Oxford Deanery), of different sizes and structures. Participants were recruited based on a convenience sample, constrained to balance the two experience groups and main hospital affiliations. “Junior” surgeons were first informed about the study through their Training Program Director and a general invitation to participate was circulated. Subsequently, they were contacted through direct invitations or personal introduction from colleagues. “Senior” clinicians were contacted directly or through personal introduction from colleagues. The participants did not receive any form of compensation for taking part in the study.

All participants received a participant information sheet (**Annex II-2**) prior to the interview and were consented using an electronic consent form (**Annex II-3**). Consent could be withdrawn at any time. Interviews were conducted in parallel for the two study groups and without pre-defined order.

Sample size

The initial sample size was set at 10 participants in each experience group, with the possibility to recruit additional participants if thematic saturation was not reach after the initial recruitment round. In order to reduce the burden on clinicians in the context of the second and third wave of COVID-19 in the United Kingdom, it was subsequently decided to reduce this number to six participants in each group. Indeed, previous research have shown that, when investigating a well-defined question in a relatively homogenous population, the majority of topics were usually found within the six to 12 first interviews[207,208]. This corrected sample size would remain subject to confirmation by a thematic saturation analysis (see section II.2.3).

II.2.2 Applied cognitive task analysis

Design of the Applied Cognitive Task Analysis

The ACTA methodology is articulated around four components. The first step was to create a Task Diagram for ACTA analysis, highlighting the cognitively challenging aspects of a task (in this case, the management of a postoperative complication). SMEs were presented with a general scenario of postoperative complication on the ward and asked to describe, in four or five steps, their decision-making process when attending this patient. Subsequently, they were asked to identify the steps most cognitively challenging for them.

Second, a knowledge audit was conducted, articulated around key knowledge categories characterising expertise and selected from literature on expert-novice differences[203]. In the complete ACTA knowledge audit, these categories include: “past & future” (experts can anticipate future developments based on how a situation evolves), “big picture” (experts quickly develop an understanding of the situation as a whole), “noticing” (experts are better at detecting cues), “job smarts” (experts know how to apply the tricks of the trade to work more efficiently), “opportunities/improvising” (experts are more comfortable improvising), “self-monitoring” (experts are aware of their performance and know when to make adjustments, “anomalies” (experts have seen more cases and can better detect deviation from the norm)

and “equipment difficulties” (experts have more confidence to challenge misleading equipment). For the present studies, only five of these categories were selected after the first interview piloting. This was because it became clear that the full audit would last well over the 30 minutes target and might have decrease the willingness of SMEs to participate. Expert opinion (Dr Lauren Morgan) was sought to select the most relevant categories. “Past & future” and “Noticing” were merged into a single question. “Opportunities/Improvising” was dropped based on previous experience of Dr Morgan. “Equipment difficulties” was also dropped based on previous experience and to avoid diverting the interview toward a list of complaints about electronic health record functionalities.

Third, simulation interviews were conducted, which placed SMEs in simulated operational settings and echo incident-based interviews. Because of time constraints, it was originally planned not to perform this step. However, given its usefulness to elucidate differences in perspective between experts and novices, it was decided to combined step one and step three, by presenting a clinical scenario while interrogating the SMEs on the task diagram.

Fourth, a tailored cognitive demands table was created to organise and analyse the data. The table was based on the recommended template and modified to reflect the objectives of the study as well as facilitate the integration of the results within the broader thesis project. For each of the challenges identified from the task diagram, its main categories comprised: reasons for being difficult, common errors, strategies used to face challenges and difficulties, and desired computerised support. An inductive thematic analysis (see section below) was first performed on the interview transcript to formalise the qualitative data analysis process[209], and the identified themes were subsequently organised into the table.

Interviews

As a result of the COVID-19 pandemic, interviews were conducted online using the Microsoft Teams software (www.microsoft.com/en/microsoft-teams/log-in) and by telephone in one instance. All interviews were conducted by myself, a qualified medical doctor writing a DPhil thesis under the supervision of two academics with extensive experience in human factors

research (Dr Lauren Morgan and Prof Peter McCulloch). I had at the time no previous experience with qualitative research. Interviews were audio recorded using a Homder TF-85 digital voice recorder device and professionally transcribed through the TP Transcription Limited (reference TP-4199) transcription service. Transcription were made at "intelligent verbatim" level, described as verbatim where "certain elements are omitted if they add no meaning to the script. This is mainly ums, errs and repetitions and false starts" by the service provider[210]. Brief additional field notes about any relevant environmental factors (e.g. disruption during the interview) and a personal assessment of the interaction quality with the participants were also recorded. Participants had no opportunity to review their interview transcript.

The interview protocol comprised: an introduction, identically read to all study participants and presenting the main objectives of the study, the role of the study within the thesis' larger research interest, as well as the methodology used; a personal information section, to collect information about the clinical experience and responsibilities of the participants as well as about their current work environment; an ACTA-specific section with the clinical scenario and five knowledge category probes, complemented by the same follow up question every time; and an opportunity to share any additional thoughts on the management of postoperative complications and desired support modalities. The protocol was designed for the interview to last around 30 minutes.

Two versions of the interview protocol were developed, one for junior and one for senior participants. The main differences were in the scenario proposed and the follow up question. The scenario was adapted to the role of senior staff in hospital. The follow up question for junior participants was "How could a computer or intelligent algorithm support you along this decision-making process?", whereas the one for senior participants was "What do you think novices would do differently in such situation?". The interview protocols were piloted with one consultant surgeon, one junior surgeon and one qualitative research specialist. They can be found at the end of **Annex II-1**.

Thematic analysis

Thematic analysis is a versatile approach to qualitative data analysis, whose main objectives are to identify, analyse and report recurrent themes (i.e. patterns of meaning) in the examined sources, based on an active intermediary coding process[211]. It is one of the basic and most accessible qualitative analysis methods, and can support the conduct of other, more complex, analysis methods or theory. It usually consists of six steps: a familiarisation stage, the generation of initial codes, the search for themes, the review of themes, the definition and naming of themes, the report production. Thematic analysis is not a linear, but a recursive process, during which authors go back and forth through the stages and writing should begin from stage one rather than only at the end of the analysis[211].

In this study, thematic analysis was preferred to other qualitative analytic methods, such as narrative analysis, or grounded theory, for its balance between simplicity of use, flexibility (it can be use across a range of theoretical frameworks), and rigor of application. An inductive approach (i.e. codes and themes are derived from the collected data and strongly linked to them[212]; as opposed to a deductive approach where themes and codes are pre-defined or informed based on existing theory) was chosen to not restrict the range of support modalities identified by the study and because literature on the subject is still sparse. No specific prevalence criteria were set for a theme to be considered as such (i.e. a theme could be mentioned in as little as one single interview). The analysis focused on a rich description of the data set, rather than a detailed account of a subset of themes. Despite the clear emphasis on AI-based decision support, the objective was indeed also to contextualise the expectations for this technology within the broader landscape of decision support in hospitals (e.g. EHR, on-call colleagues, etc.). Themes were constructed at a semantic level to limit the influence of the author's personal biases on this topic. This means that themes were constructed to convey the explicit meaning of the data, as opposed to latent/interpretative themes which would reflect underlying ideas or assumptions on the data. High level of granularity was preferred to broad, overarching concepts when selecting themes.

Familiarisation happened through data collection and a transcript check performed by the main author. Initial ideas for codes were written down during the transcript check. The NVivo software (QSR International Pty Ltd., v1.0) was used for the coding process and reviewing of themes. The attributed codes were arranged into thematic maps according to the three categories of the designed cognitive demands table ("reasons for being challenging", "common errors" and "coping strategies") and the associated desired computerised support needs, using the draw.io software (JGraph Ltd., v15.8.7). Common challenges were not analysed thematically, as they were directly identified by the participants in reference to the developed task diagram. Themes were identified from the thematic map and their content reviewed for "internal homogeneity" and "external heterogeneity"[212]. Adjustment to coding and thematic grouping were performed where necessary. Final themes were named and charted into the cognitive demands table. Beside its name and location within the cognitive demands table, a brief narrative description was also produced for each theme.

II.2.3 Thematic saturation analysis

Analysis of thematic saturation was conducted according to the method proposed by Guest *et al* [213], described in the next paragraph. The main advantages of this method are that it can be used both prospectively and retrospectively, does not require prior knowledge of theme prevalence, and conceptualises saturation as a relative measure (therefore, the influence of researcher's coding granularity will be neutralised as it influences both the numerator and the denominator). Analysis of saturation was made at code, rather than theme, level, reflecting the definition of saturation originally published by Glaser *et al.* that saturation not only pertains to categories (themes) but also to data from which researcher can "develop properties of the categories"[214].

Only codes pertaining to categories relevant to the cognitive demands table were considered in the analysis of saturation. A base size (i.e. number of interviews whose codes form the denominator) of six, a run length (i.e. the number of interviews whose codes form the running

numerator) of two and a new information threshold of ≤ 5% were selected, representing an intermediate level of confidence than saturation has been reached[213]. Interviews were chronologically ordered. The number of new codes for each interview was calculated. The numbers of new codes within the six first interviews (the basis) were summed to form the number of base codes (the denominator). The number of new codes in the subsequent 2 interviews (interviews n°7 and 8, the run window) were summed to form the number of new run codes (the numerator). The number of run codes was divided by the number of base codes. If the result was smaller than or equal to the specified 5% threshold, it was concluded that saturation had been reached, otherwise the run window was advanced one position (to interview n°8 and 9), and the same process was repeated. The analyses were performed using the NVivo software (QSR International Pty Ltd., v1.0) and Python (Python Software Foundation, v.3.8.5, including standard packages). To address the limitation that this method is dependent on the order in which the interviews are analysed and that hence saturation might have been reached by chance only, the same analysis was performed again for all relevant interview order permutations.

II.3 Results

II.3.1 Participants

A total of twelve participants (six in each experience group) participated in the study. No surgical trainees answered the first, general invitation circulated through the Program Director. Subsequently, all the clinicians were contacted directly or through personal introductions and accepted to participate in the study, although one consultant contacted through personal recommendation never responded to follow up after initially accepting the invitation. No participant withdrew consent after being interviewed. I had at the time previous professional relationships with four of the participants, including professional as well as personal relationship with one of them. Interviews took place between November 2020 and August 2021. **Table II-1** gives an overview of the participants characteristics and length of interviews. All interview transcripts were included in the analysis.

Interview #	Participant ID	Seniority level	Specialty	Length of interview	Length of transcript[a]
1	13115	CT2	General Surgery	33:33	4,286 words
2	11468	ST6	Intensive medicine and Anaesthetics	34:54	3,911 words
3	11385	ST3	Paediatric Surgery	40:15	4,486 words
4	12218	ST6	Intensive Medicine	32:49	3,918 words
5	11308	ST4 + 2 years[b]	Transplant Surgery	32:54	5,030 words
6	11315	CT2	Orthopaedic	32:13	4,389 words
7	11398	Consultant	HPB Surgery	31:13	3,426 words
8	10355	CT1	General Surgery	30:38	4,548 words
9	10025	CT2	Orthopaedic and Trauma	25:37	4,114 words
10	11485	CT1	General Surgery	23:04	2,545 words
11	10098	Consultant	Vascular	28:15	4,000 words
12	20458	Consultant	Plastic and Hand Surgery	30:02	3,559 words

Table II-1: participants and interviews characteristics. Details about hospital workplaces are not individually reported to avoid participant identification. The length of interviews is reported in minutes. [a] excluding introduction and personal information sections; [b] the participant was considered as senior; HPB = Hepato-Pancreatico-Biliary.

Participants were either at home or in their hospital office during the interview. They were alone and undistracted, although some instances of brief interruption (colleagues walking in, dog barking, phone ringing, charger needed) occurred during five interviews.

II.3.2 Study sites

Participants worked in the following hospitals: Churchill Hospital (specialist cancer centre), John Radcliffe Hospital (University teaching hospital), Royal Berkshire Hospital (district general hospital), Stoke Mandeville Hospital (district general hospital), Wexham Park Hospital (district general hospital). All hospitals had ICU facilities available. All participants in the junior group reported that there were no consultant surgeons on site during the night.

II.3.3 Task diagram

A task diagram was constructed from the interviews leading from the initial alarm to the definitive treatment plan. Six steps were identified as cognitively or logistically challenging (initial task prioritisation, producing a differential diagnosis, prescribing additional investigations, generating a treatment plan, severity assessment, escalation) and are discussed in more details in the cognitive demand table (see section II.3.7). **Figure II-2** presents the ACTA task diagram.

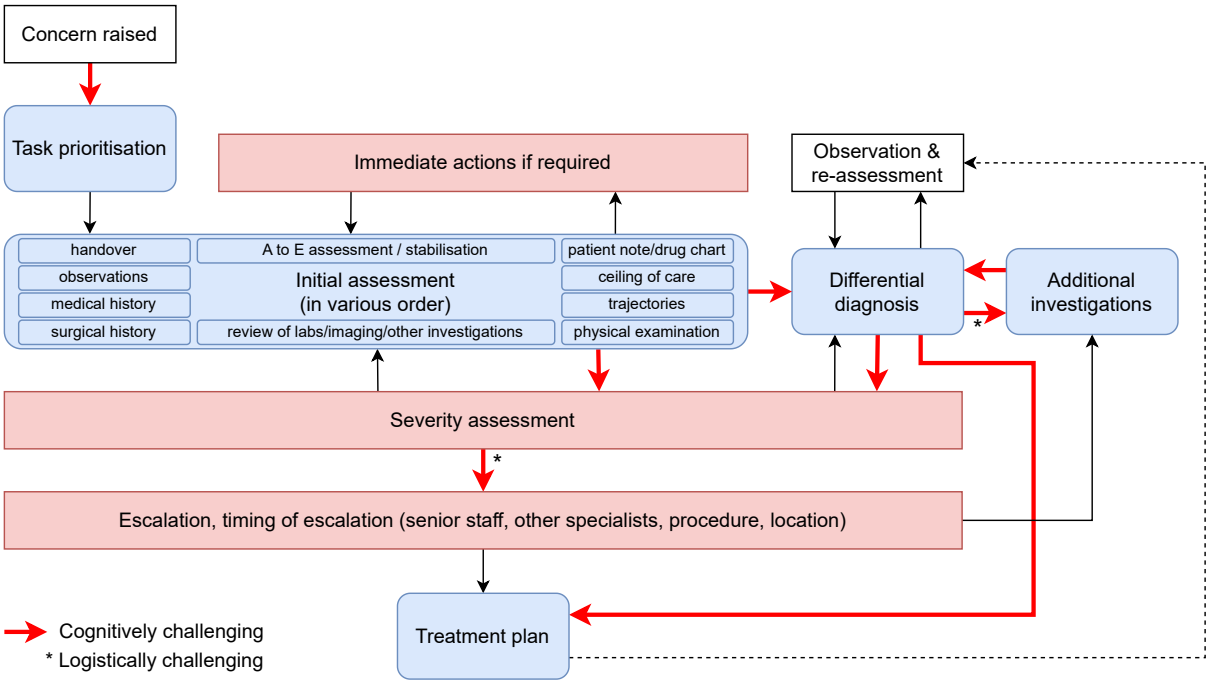


Figure II-2: ACTA task diagram. The components of the initial assessment step were consistent across interviews but not their order. Some clinicians also produce a list of problems, as intermediary step between the initial assessment and differential diagnosis (not shown). Re-assessment after having started the treatment plan is common and might lead to adapt the differential diagnosis.

II.3.4 Thematic analysis

A total of 134 codes were identified, of which seven pertained to the cognitively challenging decision-making steps, 111 to categories of the cognitive demands table (34 about the reasons why these steps are difficult, 12 about commonly made errors, 39 about the strategies adopted to face challenges and difficulties, and 26 about desired computerised support). 14 of the remaining codes were about problems encountered when working with computers, one was about the nature of information coded in EHR and one about the perception of being a junior or a senior clinician. **Figure II-3** to **Figure II-6** display the thematic maps for the four specified categories, including codes and identified themes. The total number of unique codes attributed to each interview can be found in **Table II-2**. The detailed prevalence of all codes in each interview can be found in **Suppl. Table II.1**. Themes are described in more details in **Table II-4** to **Table II-7**.

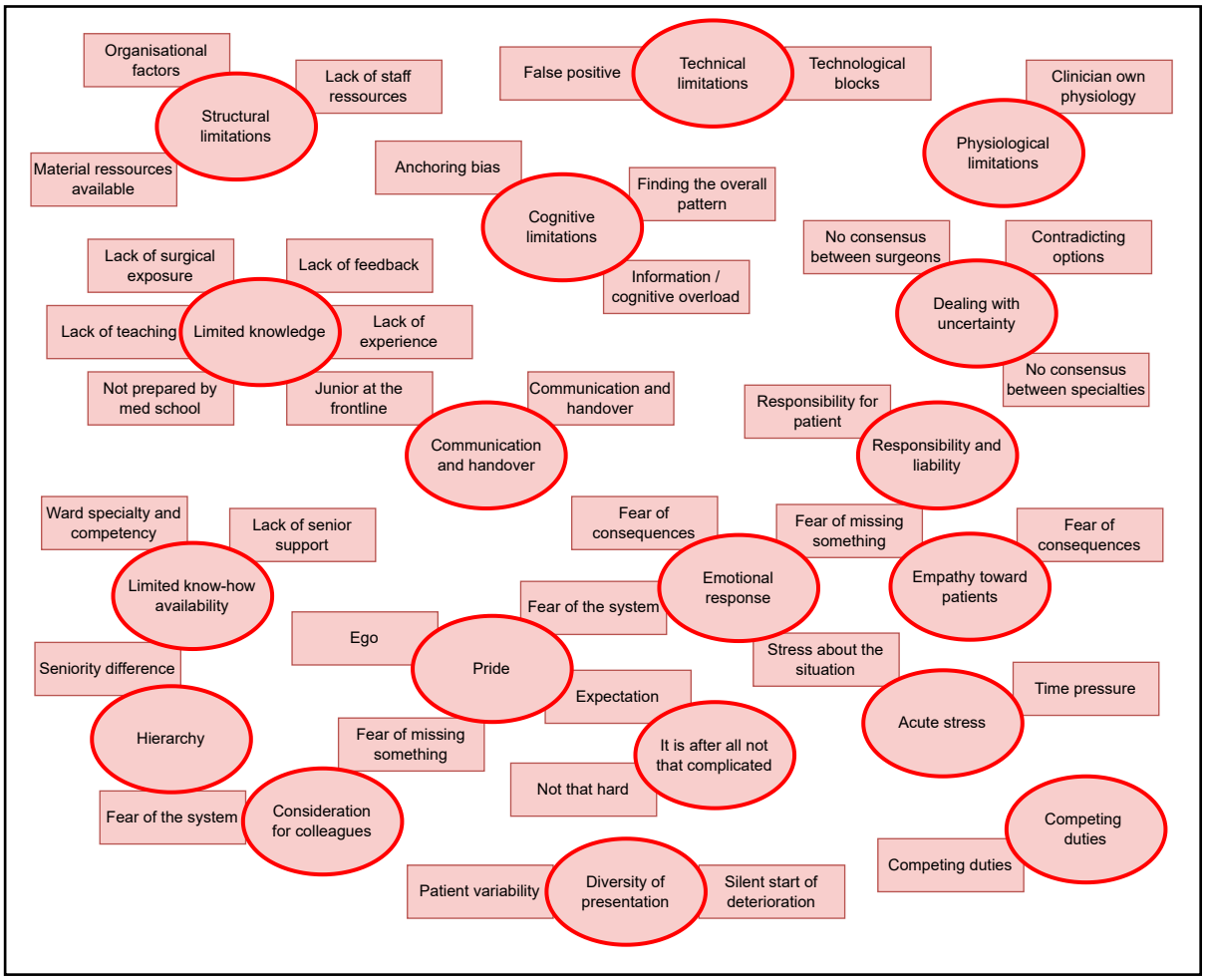


Figure II-3: reasons for being difficult. The themes' names refer to reasons given to explain why the reported steps are challenging. Themes are oval, code rectangular.

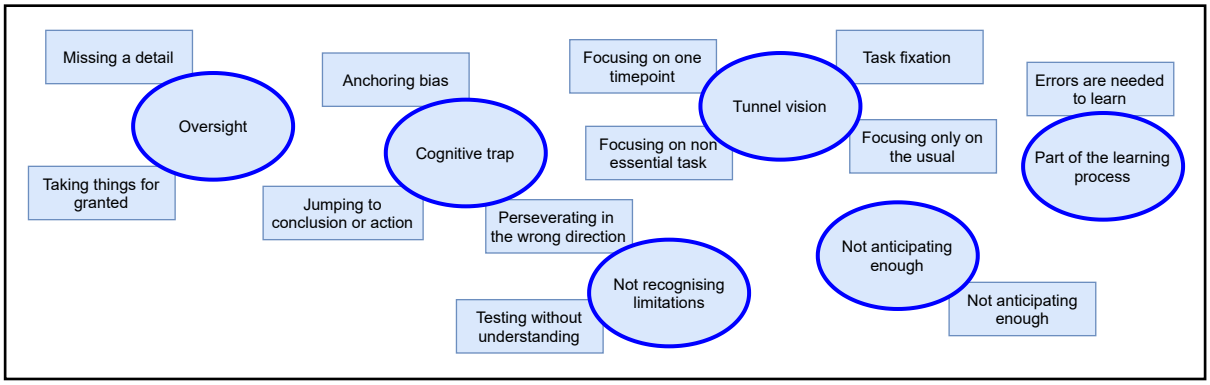


Figure II-4: common errors. The themes' names refer to common errors made by (often junior) clinicians. Themes are oval, code rectangular.

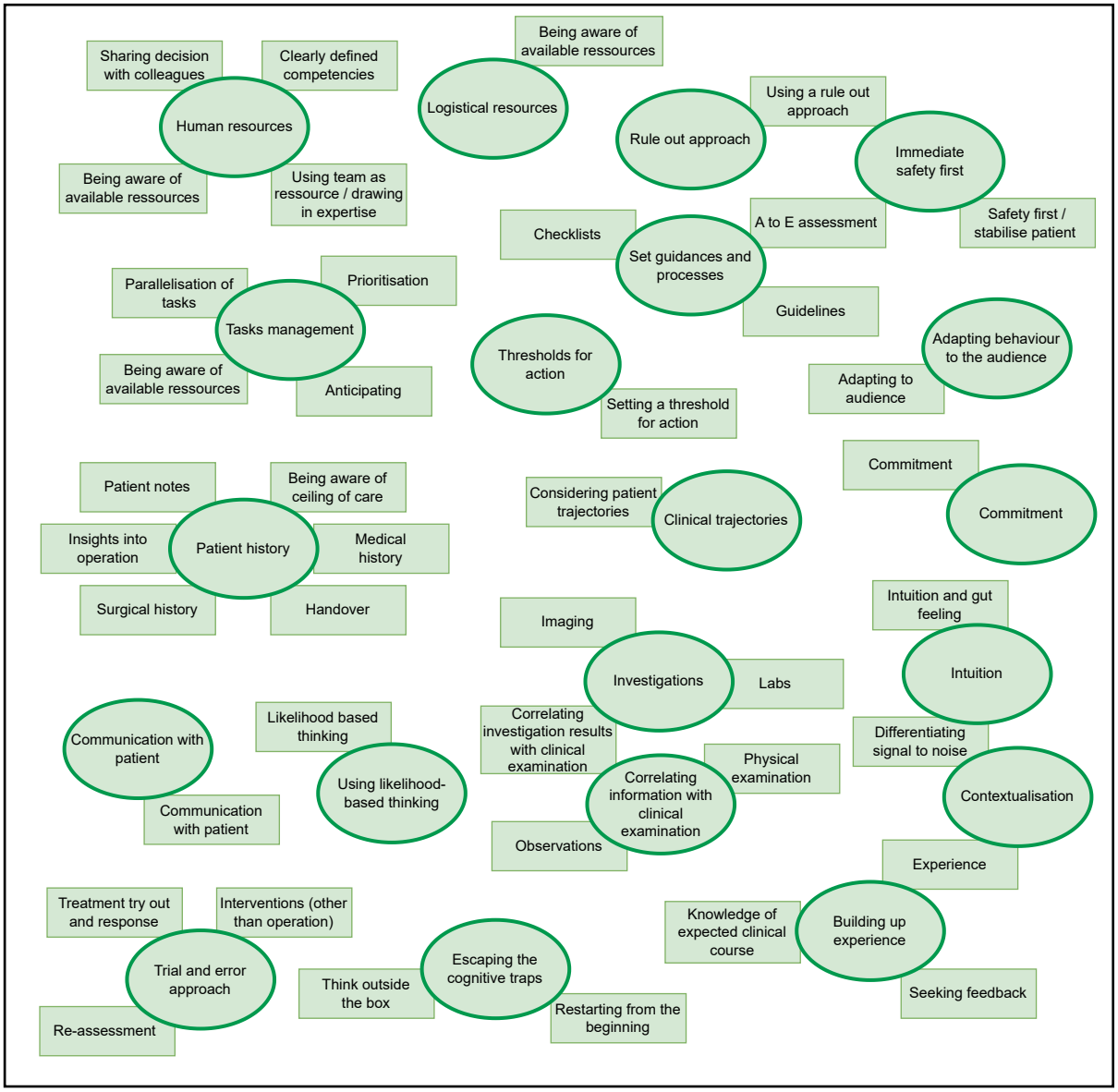


Figure II-5: coping strategies. The themes' names refer to strategies used by clinicians to overcome the reported challenges and difficulties. Themes are oval, codes rectangular.

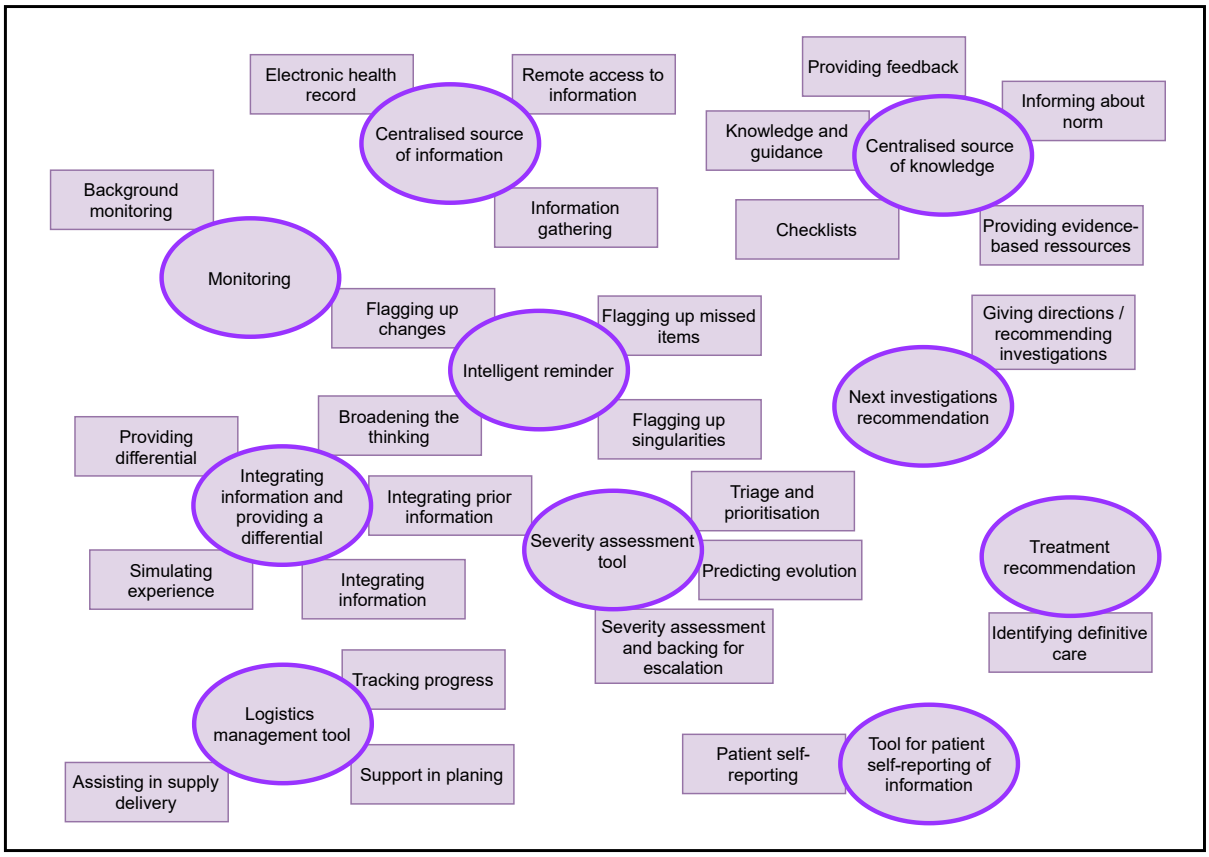


Figure II-6: desired computerised support. The themes' names refer to what the computerised support should provide. Themes are oval, codes are rectangular.

II.3.5 Thematic saturation analysis

Thematic saturation was reached at 8^{+2} (8 interviews + a run window of 2 interviews). **Table II-2** present a summary of the thematic saturation analysis.

	Interview1	Interview 2	Interview 3	Interview 4	Interview 5	Interview 6	Interview 7	Interview 8	Interview 9	Interview 10	Interview 11	Interview 12
Number of codes	33	22	34	32	40	33	28	33	24	36	45	44
Number of new codes	33	17	20	13	13	4	5	4	2	1	2	3
Number of new codes in basis/run window	100							9	6	3	3	5
% new code	N/A							9 %	6 %	**3 %**	3 %	5 %

Table II-2: thematic saturation analysis results. The interviews are chronologically ordered. The number of new codes is based on the difference with all the precedent interviews (i.e. not just the first six basis).

To investigate the impact of interview permutations on the interview position at which saturation was reached, the same process was repeated 665,280 times (see equation 1, with total number of interviews n = 12 and basis size k = 6), representing all unique permutations with analytical relevance (i.e. the interview order in the first six basis is irrelevant).

$$C_k(n) * (n-k)! = \frac{n!}{k!\,(n-k)!} * (n-k)! = \frac{n!}{k!} = \frac{12!}{6!} = 665{,}280 \quad (1)$$

Saturation was reached at or before the 12th interview in 627,254 cases (94.3%). **Figure II-7** shows the distribution of the interview positions at which saturation was reached.

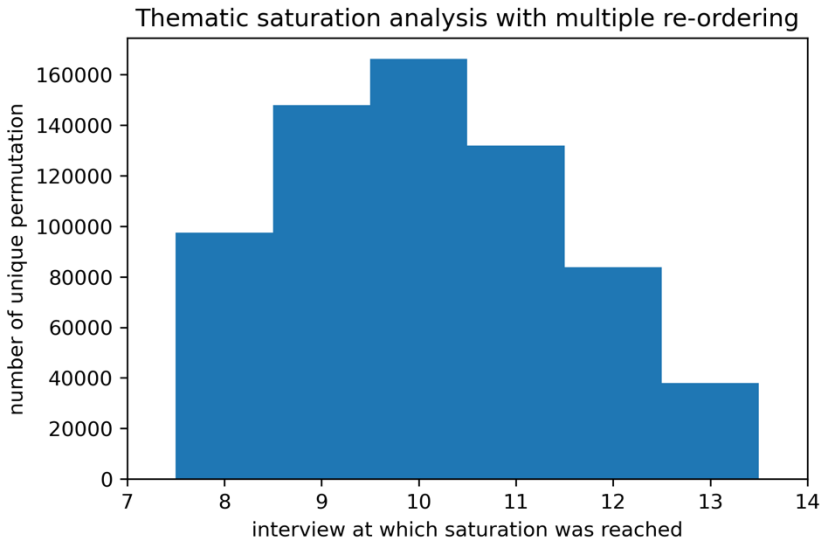


Figure II-7: thematic saturation analysis with multiple re-ordering. Distribution of the interview at which thematic saturation was reached for all relevant interview order permutations. 13 represent permutations for which saturation was not reached after 12 interviews, but not necessarily that saturation would have been reached with a 13th interview.

II.3.6 Difference between junior and senior clinicians

Throughout the interviews, it could be observed that all participants, including these considered as juniors according to the study's protocol, saw themselves as someone else's senior. Even when not directly prompted to do so, junior participants described what their own juniors would do differently and potentially wrong. From this perspective, the difference between junior and senior was less marked than anticipated. It additionally appeared that even the more junior participants had a range of coping strategies at their disposal and that the differences between juniors and seniors in terms of strategies used was also less clear cut that expected. Therefore, and with the exception of the task diagram, the results are presented for the whole cohort, as opposed to segregated by seniority level.

Regarding the task diagram, the challenges reported in section II.3.3 reflects mainly the junior perspective. Seniors principally identified challenges around arriving at a correct differential diagnosis, therapeutic escalation and analysing the nuances of treatment plans. These differences between senior and junior perspectives are partially explained by organisational factors. Seniors only become involved in second or third line after a junior has already assessed the situation and have the long-term responsibility about the definitive management. Logistic challenges appeared to impact more on juniors than seniors for reasons explained under the theme "hierarchy" in **Table II-4**.

II.3.7 Cognitive demands table

Table II-3 displays the cognitive demands table. The identified themes are attributed to the relevant challenges (initial task prioritisation, producing a differential diagnosis, prescribing additional investigations, generating a treatment plan, severity assessment, escalation) and categories (reasons for being difficult, common errors, coping strategies, desired computerised support). The same theme can be relevant to several challenges. Theme content and explanatory quotations are detailed thereafter.

Cognitive challenges identified

Initial task prioritisation: after noticing an alarm, a NEWS score trigger or receiving a call from the ward staff, clinicians have to make a first decision, whether to attend right away or after some delay, considering the other ongoing tasks.

Producing a differential diagnosis: after their initial assessment of the patient and review of the available information, clinicians have to produce a list of potential diagnosis, which explain the current clinical presentation. This differential diagnosis is the basis which informs the additional investigations and treatment plan. Some clinicians also produce a list of problems, as intermediary step between the initial assessment and differential diagnosis.

Prescribing additional investigations: clinicians have to select additional investigations to refine their differential diagnosis. Compromises must be made between adequate use of resources and information gain.

Generating a treatment plan: based on the most likely and potentially most serious diagnosis, clinicians have to prescribe a list of therapeutic measures. This includes observation and conservative treatment.

Severity assessment: all along their decision-making process, and with frequent re-assessment, clinicians have to evaluate the severity of the patient's condition. This dictates the level of measures used and their urgency.

Escalation: based on the severity assessment and clinical presentation, clinicians have to decide whether to escalate care. Escalation can happen in terms of staff seniority, level of intervention (e.g. conservative *vs* invasive), or care settings (e.g. general ward *vs* ICU).

The step of initial assessment (comprising handover, patient note review, observations, history taking, physical examination, as well as existing laboratory and imaging review) and the decision-making element about taking immediate actions if necessary (immediate measures to stabilise the patient or alleviate their suffering) were mentioned by several participants, but none described them as particularly cognitively challenging.

(legend on page 47)

Challenges	Reasons for being difficult	Common errors	Coping strategies	Desired computerised support
Initial task prioritisation	• Acute stress • Cognitive limitations • Communication and handover • Competing duties • Diversity of presentation • Emotional response • Hierarchy • Limited knowledge • Physiological limitations • Structural limitations • Technical limitations	• Cognitive traps • Not anticipating enough • Oversight • Tunnel vision	• Building up experience • Clinical trajectories • Commitment • Contextualisation • Heuristic • Immediate safety first • Investigations • Patient history • Set guidance and processes • Tasks management • Using likelihood-based thinking	• Centralised source of information • Centralised source of knowledge • Integrating information and providing differential • Intelligent reminder • Monitoring • Severity assessment tool • Tool for patient self-reporting of information
Producing a differential diagnosis	• Acute stress • Cognitive limitations • Communication and handover • Competing duties • Dealing with uncertainty • Diversity of presentation • Emotional response • Hierarchy • Limited know-how availability • Limited knowledge • Consideration for colleagues • Physiological limitations • Pride • Structural limitations • Technical limitations	• Cognitive traps • Not recognising limitations • Oversight • Tunnel vision	• Adapting behaviour to the audience • Building up experience • Clinical examination at the centre • Clinical trajectories • Commitment • Communication with patient • Contextualisation • Escaping the cognitive traps • Heuristic • Human resources • Immediate safety first • Investigations • Patient history • Rule out approach • Set guidance and processes • Trial and error approach • Using likelihood-based thinking	• Centralised source of information • Centralised source of knowledge • Integrating information and providing differential • Intelligent reminder • Severity assessment tool • Tool for patient self-reporting of information

(legend on page 47)

Challenges	Reasons for being difficult	Common errors	Coping strategies	Desired computerised support
Prescribing additional investigations	• Acute stress • Cognitive limitations • Competing duties • Dealing with uncertainty • Empathy toward patients • Emotional response • Hierarchy • Limited know-how availability • Limited knowledge • Consideration for colleagues • Physiological limitations • Pride • Responsibility and liability • Structural limitations • Technical limitations	• Cognitive traps • Not anticipating enough • Not recognising limitations • Tunnel vision	• Adapting behaviour to the audience • Building up experience • Commitment • Communication with patient • Escaping the cognitive traps • Heuristic • Human resources • Immediate safety first • Investigations • Logistical resources • Patient history • Rule out approach • Set guidance and processes • Tasks management • Using likelihood-based thinking	• Centralised source of information • Centralised source of knowledge • Intelligent reminder • Logistics management tool • Next investigations recommendation
Generating a treatment plan	• Cognitive limitations • Competing duties • Dealing with uncertainty • Empathy toward patients • Emotional response • Hierarchy • Limited know-how availability • Limited knowledge • Consideration for colleagues • Physiological limitations • Pride • Responsibility and liability • Structural limitations • Technical limitations	• Cognitive traps • Not anticipating enough • Not recognising limitations • Tunnel vision	• Adapting behaviour to the audience • Building up experience • Commitment • Communication with patient • Human resources • Investigations • Logistical resources • Patient history • Set guidance and processes • Tasks management • Threshold for action • Trial and error approach • Using likelihood-based thinking	• Centralised source of information • Centralised source of knowledge • Intelligent reminder • Logistics management tool • Treatment recommendation • Tool for patient self-reporting of information

Challenges	Reasons for being difficult	Common errors	Coping strategies	Desired computerised support
Severity assessment	• Acute stress • Cognitive limitations • Communication and handover • Dealing with uncertainty • Diversity of presentation • Empathy toward patients • Emotional response • Hierarchy • Limited know-how availability • Limited knowledge • Consideration for colleagues • Physiological limitations • Pride • Responsibility and liability • Structural limitations • Technical limitations	• Cognitive traps • Not anticipating enough • Not recognising limitations • Oversight • Tunnel vision	• Building up experience • Clinical examination at the centre • Clinical trajectories • Commitment • Communication with patient • Contextualisation • Escaping the cognitive traps • Heuristic • Human resources • Immediate safety first • Investigations • Logistical resources • Patient history • Rule out approach • Set guidance and processes • Tasks management • Using likelihood-based thinking	• Centralised source of information • Centralised source of knowledge • Integrating information and providing differential • Intelligent reminder • Logistics management tool • Severity assessment tool • Tool for patient self-reporting of information
Escalation	• Acute stress • Cognitive limitations • Communication and handover • Dealing with uncertainty • Empathy toward patients • Emotional response • Hierarchy • Limited know-how availability • Limited knowledge • Consideration for colleagues • Physiological limitations • Pride • Responsibility and liability • Structural limitations	• Cognitive traps • Not anticipating enough • Not recognising limitations • Oversight • Tunnel vision	• Adapting behaviour to the audience • Building up experience • Clinical examination at the centre • Clinical trajectories • Contextualisation • Heuristic • Human resources • Immediate safety first • Investigations • Logistical resources • Patient history • Set guidance and processes • Tasks management • Threshold for action • Using likelihood-based thinking	• Centralised source of information • Centralised source of knowledge • Integrating information and providing differential • Intelligent reminder • Logistics management tool • Severity assessment tool • Treatment recommendation

Table II-3: cognitive demands table. This table presents, for each main cognitive challenge, the themes identified as reasons for being difficult, common errors, strategies used and desired computerised support. Themes are displayed in alphabetical order. A detailed explanation of each theme can be found in the following section.

Reasons for being difficult

The themes listed in **Table II-4** were identified as reasons contributing to the difficulty of the cognitively challenging steps listed in the previous section. These reasons are of different nature: cognitive, physical, psychological, social, logistical, and institutional.

Acute stress	The stress caused by an acutely unwell and potentially rapidly deteriorating patient can hinder rational decision-making. *"[…], but if this patient is peri-arrest, I am likely to be quite buzzed. That's likely to be influencing my wider judgment."*
Cognitive limitations	Cognitive biases were often mentioned, like the anchoring bias in which a clinician's decision-making process is influenced by either a suggestion given at handover (*"Look I think [the patient has] abdominal compartment syndrome"*) or their initial opinion on the case. The difficulty of finding the overall pattern in the available information and information overload were other aspects mentioned in relation to cognitive limitations.
Communication and handover	Information can be lost or misinterpreted during handover between shifts, groups of medical professionals, medical specialties, or seniority level. For example, the junior doctors being at the frontline of patient care, and potentially failing to recognise important signals, consultants can remain unaware that a critical situation is unfolding.
Competing duties	In every given clinical shift, clinicians have multiple tasks, duties and commitments happening in parallel. *"I think a major cognitive bottleneck is what else is going on in my job, my wider workload."*
Consideration for colleagues	This is different from fear of the hierarchy or ego-related blockade and refer to the positive inclination not to disturb colleagues as long as possible, even more so when out of hours. *"So the juniors, they sort of soldier on themselves either because they don't want to bother you or they don't know that they should […]"*
Dealing with uncertainty	There is often no single ground truth. *"So it's not like a black and white rule; there is a greyscale of decision-making […]"* Different senior clinicians or different specialists might have different answers to the same question or different opinions on the best way forward with a case. This can lead to situations where the available options are mutually exclusive (e.g. going to interventional radiology for embolization vs going to theatre for open surgery).
Diversity of presentations	The same pathology can present differently depending on the patients.
Empathy toward patients	Most clinicians care for the wellbeing of their patient at a personal level. They want to avoid missing something which could negatively impact their care or exposing them to harm during diagnostic test or therapeutic intervention. This empathy toward the patient can influence the (rational) decision-making process. *"Fear of the consequences of operation might affect a novice."*
Emotional response	Emotions can affect judgement. The topic of fear was commonly cited, namely fear of missing something important, fear of the consequences for the patient, or fear of the system and potential consequences for the clinician. *"[…] because people are scared of calling a consultant unless they absolutely have to."* Acute stress in front of a rapidly deteriorating patient can also trigger an emotional response.
Hierarchy	Senior clinicians felt that juniors have to fight harder to convince people around them, even when they are correct. *"Whereas if you are a junior, you get challenged a lot more."* Juniors also sometimes don't trust themselves to question the treatment delivered by a consultant.

Limited know-how availability	Lack of senior support (mainly out of hours) and lack of specific expertise depending on the physical location of the patient were cited as limiting factors. *"Out of hours I think there is a whole separate problem. People are more hesitant to call for help as there is less senior presence available in hospitals; so out of hours adds a whole different element of complexity to patient care."*
Limited knowledge	Six aspects were cited to explain limitations in knowledge when dealing with postoperative complications: i) junior doctors occupy the frontline positions, ii) medical school does not always prepare for the practical reality in hospital, iii) juniors still lack experience, iv) juniors lack of surgical exposure (having assisted to the operations leading to complications could facilitate the understanding of the pathophysiological mechanisms and provide important prior information), v) there is a lack of teaching, and vi) as patients and shifts constantly change, it is difficult to follow up on cases and there is a lack of feedback/opportunities to learn from outcomes. *"I think what sometimes happens is that the people parachute in, the seniors arrive, and then either the patient disappears, either to scan, theatre or ICU, and then they don't know what has happened, so they don't have a good understanding of feedback loop."*
Physiological limitations	Constraint to which clinicians are subjected by their own body. *"It's also 1am, so I'm likely to be tired, hungry, disoriented."*
Pride	Ego, perceived or real expectations, and fear of being judged by others (e.g. as incompetent) can hinder rational decision-making, blind to limitations and delay escalation. The fear of missing important information can also lead to an over prescription of tests. *"[Because of] the fear of missing something, you end up going down [the spiral] of testing for everything."* Notably, seniors seemed to have more realistic expectations from their juniors that these latter have from themselves. *"I don't think the juniors are expected to come up with an answer. All they need to do is recognise a setback." "The bottom line is juniors don't have that experience to make a call. And you can't expect them to."*
Responsibility and liability	Two aspects were discussed by participants: i) responsibility in terms of long-term management (*"[The juniors are] more focussed on the end of their shift […] and then it will be someone else's problem. Whereas [if] you are the operating surgeon, it's your complication, then that is going to live with you forever, or you will be dealing with it over multiple days, weeks potentially."*), ii) legal liability for their actions (*"[This phone call with a specialist] would be the kind of the get out of jail free card if she deteriorated."*)
Structural limitations	Limitations inherent to the hospital organisation, such as lack of staff, limited availability of technical expertise (e.g. CT scan and radiologist) or material resources (e.g. blood culture kit). *"You really have to like, you know, beg, plead, kill someone to get a scan."*
Technical limitations	Limitations caused or difficulties created by technology. Two examples were given: i) false positive test results, and ii) electronic prescription system non-functioning or refusing to validate a specific investigation.

Table II-4: reasons for being difficult. Descriptions and explanatory quotations for the themes identified as reasons making the different challenging steps difficult. Quotations are reproduced from interview transcripts.

Of note, two senior participants expressed the view that managing surgical complications at junior level is after all **not that complicated**, either because such complications fall into big groups whose diagnosis is most of the time straightforward, or because what is expected of juniors is mainly to take immediate actions and call for help.

Common errors

The themes presented in **Table II-5** were identified as common errors made by juniors when attending postoperative complications. A discrepancy between the number of identified reasons why challenging steps are difficult overall and the number of identified common errors made by junior clinicians can be observed.

Cognitive traps	This describes juniors falling into cognitive traps, such as anchoring bias (see **Table II-4**), confirmation bias (selecting and interpreting information to confirm one's existing belief), or skipping rational decision making steps, hence jumping to conclusion or actions. *"There is this idea about this type 1 and type 2 thinking, where there is a mechanism where you can jump to conclusion essentially."*
Not anticipating enough	Not anticipating issues delay appropriate care and can lead to suboptimal use of resources. *"Following the chess analogy a bit more: they see a move that deals with the immediate issue and they don't look then and think about what the next step is."*
Not recognising limitations	This includes juniors' own limitations, especially in term of knowledge (like for example prescribing tests whose results they can't fully interpret), and the limitations of a chosen course of action (like for example not recognising that the current management plan is not delivering the expected results and perseverating in the same wrong direction).
Oversight	Oversights were described either in terms of missing an important detail amongst all the information available (e.g. urine output slowly deteriorating) or taking things for granted (e.g. *"They just took it for granted that the negative pressure dressing was intact; it was functioning and therefore that was fine."*)
Tunnel vision	Often mentioned, this theme refers to a reduction in clinicians' cognitive scope either because they are focusing on a limited set of information (e.g. the usual diagnosis or only the current state of the patient omitting to consider prior data) or on a limited number of tasks, important but not unique or not even essential at all. *"Everyone has been a junior once, right? Their first instinct is go and put in the drip, isn't? You still have that urge […], if in doubt, go and put a drip in, so that you look busy."*

Table II-5: common errors. Descriptions and explanatory quotations for the themes identified as common errors made by juniors. Quotations are reproduced from interview transcripts.

Of note again, two of the participants (a senior and a junior) highlighted that errors are not only negative but also **part of the learning process**. *"I am not sure you can become a doctor without making these sorts of judgment errors; and there are just some people that think about their errors more than others."*

Strategies

The themes listed in **Table II-6** were identified as strategies used to overcome the difficulties encountered when managing postoperative complications. Some can be actively trained whereas others mainly come with time and experience. As already mentioned, there was no clear difference between strategies used by junior and senior clinicians.

Adapting behaviour to the audience	Most junior clinicians adapt their narrative or decision making according to their interlocutors (*"I know that in order for [a CT scan] to happen overnight I need to say X, Y or Z or I need to impress on the radiologist that X, Y and Z is happening."*) or supervisors (*"I certainly make decisions differently based on if I know one of my consultants is on call in comparison to another one of my consultants is on call. Because I know they make different decisions."*)
Building up experience	By accumulating exposure to different type of operations (as well as their postoperative management) and seeking feedback about the final outcome of their cases, clinicians develop an understanding of what the expected clinical trajectory is after a given type of intervention and of how to best react to deviations from the norm. *"I think what helps is just the experience of the operation […]; and knowing that recovery from liver surgery is different from recovery from pancreatic surgery; and what the expected norms are; and even sometimes having been there at that operation knowing whether that was straightforward or difficult."*
Correlating information with clinical examination	Clinicians interpret results from other investigations (e.g. lab tests or imaging) in the context of the clinical examination's findings, and give these later higher weights in their decision-making process. *"I would want the positive results I have to also correlate with what I have found clinically."*
Clinical trajectories	Clinicians consider a patient's clinical trajectories rather than their status at a single timepoint. *"So the big picture […], in the back of our mind, we have always got the patient's trajectory in mind."*
Commitment	Commitment gives clinicians the drive to overcome difficulties.
Communication with patient	Good communication with patients improves patients' understanding of their care and increase adherence. Communication also helps clinicians to better understand some of the patients' psychology, which can in turn help contextualise the response they are having to surgery and detect clinically still invisible variations from the expected course.
Contextualisation	With experience, clinicians interpret information in their context and use this contextualisation to differentiate signal from noise.
Escaping the cognitive traps	The participants described two strategies to escape some of the cognitive traps: i) standing back and thinking outside the box (*"I think you benefit as an intensivist from being able to stand back a bit. […], because it means you can open your mind up to things that are not directly surgical […]"*); and ii) restarting the analysis and decision-making process about a case from the beginning (*"It is actually just about having someone come along and start again at the beginning."*).
Intuition	Even if not necessarily a conscious strategy and often prone to biases, several participants mentioned intuition (or gut feeling) as an important aspect of their decision-making. *"I am not sure these are the answers that you want to hear, but I think a lot comes to intuition and data only gets you so far."*

Human resources	Clinicians use the human resources available. This include being aware of the human support available, drawing in colleagues with complimentary expertise when needed, trying to *"divide and conquer"* tasks by sharing them between team members, and sharing decision-making with peers. In cases involving several healthcare professionals, a clear definition and understanding of each individual's competencies were quoted as an important factor. *"[Stopping haemorrhage] is the kind of thing we have to do, because other people, other specialists, are taking care of the other elements in this chain […] It is a team approach."*
Immediate safety first	Clinicians focus their initial management on stabilisation and patient safety. *"So for me, it is all about initial safety of the patient."*
Investigations	Appropriate imaging and lab tests results are an important support in the decision-making process.
Logistical resources	Awareness of the available logistical resources is crucial for the management of complications. *"Yes, so I think the big picture is: where are you? […] Is this the right site and location to be managing this patient?"*
Patient history	The medical and surgical histories, as taken from the patient, read from the patient notes, received during handover, or obtained directly from assisting in theatre form the basis of clinical decision-making. Information about the ceiling of care complement this theme.
Rule out approach	Clinicians often use a rule out approach to the management of complications, in which the most serious, common or simple complications are investigated first and other, more complex or unusual complications are only fully considered at a later stage, once the former have been ruled out. *"Doing the bare basics, covering simple things first, would lead you to diagnosing something a bit more complicated later on."*
Set guidance and processes	Clinicians use set processes and recognised guidance, such as the A to E assessment, checklists, and clinical guidelines.
Task management	Good task management, including awareness of available resources and how to use them, anticipation, prioritisation, and parallelisation of tasks, helps minimise the impact of logistical hurdles. *"Are there ways of working smart, accomplishing more…? Yes I think that boils down to thinking ahead. […] So what I will do is, I will try to front-load my assessment of that patient and do a sort of blend of simultaneous assessment and management, in order to achieve task in a more timely manner."*
Threshold for action	When unsure about the best management plan (and especially when opting for conservative management), clinicians set a threshold for action. *"We decided after a certain threshold that this is not acceptable bleeding, we have to take him to theatre to explore the wound."*
Trial and error approach	Clinicians sometimes try out a treatment plan (including conservative and non-interventional) to confirm or invalidate a diagnosis. This strategy necessitates regular re-assessment, as it is uncertain whether the treatment is adequate.
Using likelihood-based thinking	Knowledge of condition prevalence and pre-test probability are key to make rationale decision, based on facts and logic, about a case. *"[…] deciding which things are more or less likely, again I think with experience, you start to understand pre-test probability and where investigation findings are positive […]."*

Table II-6: coping strategies. Descriptions and explanatory quotations for the themes identified as coping strategies to address the challenges of postoperative complications management. Quotations are reproduced from interview transcripts.

Desired support needs

The themes presented in **Table II-7** were identified as desired applications of digital technology to support clinicians in the management of postoperative complications. A certain heterogeneity on the digitalisation level of the study centres could be observed, which was also reflected in the answers of the participants. The desired computerised support hence ranged from relatively simple information management tools already used in many centres to yet undeveloped systems with advanced data interpretation and recommendation capabilities.

Centralised source of information	The computerised support should be a centralised source of information, facilitating retrieval of all the data available about the patient. Several participants highlighted that having an electronic health record at all would already be a major step forward, even without any additional information processing modalities. This centralised source of information would be useful to access information both from the bedside and remotely (e.g. on call consultants).
Centralised source of knowledge	The computerised support should be a centralised source of knowledge, such as clinical guidelines, checklists, evidence-based resources, information about the expected postoperative trajectories for given types of operation. The computerised support could also facilitate clinical feedback by providing a practical way of tracking patient evolution beyond shift end or discharge.
Integrating information and providing differential	Computerised support should, by integrating available information (both current and prior), provide a suggested differential diagnosis. This would help closing the performance gap between juniors and seniors, by simulating experience. *"If there was something that was able to recognise features of a patient's history, or background, that led you down a particular avenue, that might be helpful. Because that would essentially do what I do naturally, with experience."*
Intelligent reminder	The computerised support should broaden clinicians' perspective on the case by flagging up missing items (e.g. no troponin results available), changes in the patient condition (e.g. white blood cell count increase compare to the previous day), or important parameters which could have been overlooked in the plethora of clinical information (e.g. the patient has had no urine output in the last 6 hours). *"And […] having an algorithm there that […] can point out the other things that you have missed, I think is […] a good sort of safety net in a sense."*
Logistics management tool	Computerised support should provide logistical support to clinicians, like suggesting priorities in task execution (*"It might give you the sort of logical next steps in an order that is sensible."*), progress tracking on prescribed investigations (*"I guess similar to like ordering Domino's, […] so to have you know a task on the computer and you can see its progress in real time and whether it had stalled at a particular step […]*), or assisting in supply delivery (*"If there was an automated system where you could literally type in what you want [and] it could provide the equipment, ensure that you are getting the correct equipment"*).
Monitoring	The computerised support should monitor patient in the background and flag up any significant changes. *"I think you can only ever get snapshots of patients unless you stand next to them, although the utility of potentially standing next to them is kind of low or diminishing returns because you stop noticing changes […]"*

Next investigations recommendation	The computerised system should provide recommendation about the next investigations to be performed in order to refine the differential diagnosis. *"I think it would probably be a very good basis of, of giving you somewhere to start from."*
Severity assessment tool	*"[…] the million-dollar question is, you know, is this a serious thing or is this not a serious thing."* Based on evolution prediction, computerised support should provide a severity assessment of the case. This would help prioritising the patients and back juniors in their decision to escalate. *"Juniors won't be seen as incompetent, it's no longer them making the call. [They] now have the tool to go to their senior and say: 'hey, based on this, I need to call you' […] You take the fear factor out of the juniors."* A severity assessment algorithm could also reassure clinicians who decided in favour of conservative management.
Treatment recommendation	The computerised system should provide recommendation about the most appropriate definitive treatment plan. This would help not only to treat the patients to the best standards of care, but also to anticipate needs (e.g. operating room).
Tool for patient self-reporting of information	Computerised support could be used by patients to report information (e.g. evolution of symptoms during the night or complications occurring in the community) without having to go through a medical professional.

Table II-7: desired computerised support. Descriptions and explanatory quotations for the themes identified as potential applications of digital technology to support clinicians in the management of postoperative complications. Quotations are reproduced from interview transcripts.

II.4 Discussion

12 interviews with junior and senior clinicians were conducted to better understand the challenges, errors made, strategies used and desired computerised support when attending postoperative complications. A thematic saturation analysis was performed and confirmed that most of the relevant themes were identified through the selected participant sample. The results of this study will provide rigorously obtained perspectives to inform future development of CDSS in the field of perioperative surgery. In the context of this thesis, the study findings will be used to tailor the development of the proposed algorithm to junior surgeons' needs.

Unsurprisingly, the results of the study demonstrated that the management of postoperative complications is a multifactorial problem, comprising of several challenging steps, with often common difficulties and to which clinicians respond with similar coping strategies. This finding questions the current ML-based CDSS paradigm, in which algorithms are most of the time developed to address a single question (e.g. is this nodule malignant?). Considering that each mathematical model comes with its own error rate, concatenating single question models to answer complex medical management challenges would result in an overall system producing a constant flow of errors. Therefore, it is my opinion that future ML-based CDSS should focus on answering multiple clinical questions from a single learned classification, regression or clustering.

In terms of desired support needs, many identified themes reflect currently available tools or systems, namely centralised information sources (e.g. EHR/EPR), centralised knowledge sources (e.g. NICE guidance, UpToDate[215,216], medStandards[217], HeadToToe[218], local antibiotic therapy guidelines), monitoring and severity assessment tools[33] (e.g. Streams[219], HYPE trial[220], SEND[221], HAVEN[36]), diagnostic recommendation systems (e.g. ImageChecker, Isabel, DXplain)[54,85], and treatment recommendation systems (e.g. AI clinician[222], radiotherapy planning[101]). It is also worth mentioning detection algorithms (e.g. lung nodules[223], colon

polyps[224]), which could be affiliated to the category of intelligent reminders, even though current detection algorithms mainly focus on reducing the number of missed items during constrained tasks. Logistics management tools are also in development (e.g. resources use prediction and resources allocation[225]), even though the current algorithms mainly operate at an institutional level. Nonetheless, the study also identified some computerised support needs not currently considered (to the best of my knowledge), such as next investigations recommendation, tailored logistics management support at patient/clinician level, opportunity for inpatients to self-report information, and simplified case outcome tracking (feedback module).

Despite current efforts made to develop adequate support in other fields of medicine and surgery[226–229], the latest generation of CDSS, based on ML algorithms, is either not aimed at postoperative patients, or only includes them amongst general ward patients, without accounting for surgical specificities. Moreover, there is still a long development process between the evaluation of a CDSS in research settings and its deployment at the frontline of patient care, so that the current tailored CDSS offer in surgical wards is almost non-existent. The identified support modalities were mentioned by participants as being desired, but not yet available technology in their work settings. It is also noteworthy that, in many cases, the participants didn't wish for advanced CDSS, but rather having HER/EPR in their hospital at all (*"So, to summarise then, I would like stuff to be on computer, full stop."*).

On a related note, the analysis of the identified themes highlighted that a significant proportion of the difficulties faced by clinicians when managing postoperative complications are not cognitive in nature, but rather organisational, cultural, or logistical. Hierarchy (and the fear thereof), staffing issues, healthcare system structure as well as physiological and psychological factors affecting the clinician play an important but unquantifiable role in decision making, alongside purely cognitive considerations. CDSS developers should remain mindful of this wider context when designing new systems and ensure these latter do not add to the logistical complexity for the sake of cognitive simplicity.

It is therefore also necessary to discuss the limitations of CDSS in the management of postoperative complications, and in clinical settings more generally. Even if I did not mention this topic during the interviews, seven participants spontaneously brought it to the conversation. They made a clear point that, in many situations, the solution *"I'm afraid for your study, [is] not technology"* and that they can on the contrary see *"how a computer hinders [them] in this process"*. The reasons mentioned were: time needed to use the systems, alarm fatigue, cumbersome and rigid display of information, prescriptive rather than supportive CDSS, inability of CDSS to contextualise, CDSS inertia to changes in knowledge or practice, inappropriate recommendations adding to a case's complexity or simply reinforcing the status quo, and uncertainty about liability when using CDSS. One participant also highlighted that, as long as CDSS continue to be developed on retrospective data, the field will have to deal with an incomplete, or even biased, representation of reality.

The traces left in the electronic patient record reflect in many cases decisions which have already been made and the original set of information accessible to the clinicians at the decision timepoint can only be inferred. Many relevant pieces of information are also simply not registered at all so far, such as clinical signs and symptoms, which play a crucial role in clinicians' decision-making strategies. These findings reinforce a growing consensus in the field, and my personal opinion, that CDSS should be primarily developed to enhance, rather than replace human decision-making in clinical settings.

II.4.1 Strengths and limitations of this study

Even if the current results are specific to postoperative complications or surgical cases, the methodological approach used could be applied to other medical specialties or clinical problems in the future. The data collection spread over a range of expertise and study sites. This work will also be, to the best of my knowledge, the first attempt to systematically investigate surgeons' cognitive support needs in the perspective of developing a ML-based CDSS to address the challenges of postoperative complications.

Despite the relatively small number of interviews, the thematic saturation analysis demonstrated that most aspects relevant to the study questions have been covered by the participants.

This study should be considered in the context of several limitations in the study design. First of all, a trade-off between ease of conducting the study and exhaustiveness had to be found. In the context of the COVID-19 pandemic, many design choices were aimed at reducing the burden on practicing clinicians taking part in the study. The interviews were for example not repeated, and participants might have answered differently to the same questions if asked on a different occasion. The participants were also not asked to review the transcript of their interview, which could have improved the quality of data collection, but is time consuming.

Second, the conduct of the interview is interviewer dependent, and even if the impact of interviewer inputs was reduced using a semi-structured interview protocol, I might have brought some of my own preconceptions into the conduct of the interviews and follow up questions asked.

Third, the interviews' analysis and identification of themes is as well influenced by the researcher conducting them. As noted by Braun and Clarke[211], themes do not "emerge" and are not "discovered" in a *passive* way from the transcripts, but are *actively* identified by the person conducting the analysis, who brings their own biases and assumptions into the process. Therefore, the final identified themes might not always reflect the participants' original opinions. A way to reduce the main researcher's influence on theme identification would be to conduct a parallel review of the interview transcripts by two independent researchers. However, and although originally planned, this was not performed for this study in an effort to reduce workload on research colleagues in the context of multiple parallel projects.

The methods used themselves also have some limitations. ACTA was not specifically developed to assess tasks relying implicitly on multi-operator collaboration[203], and is sensitive to the sources of variability mentioned above (i.e. interviewers, interviewees, interviews). The

Guest *et al.* method to assess thematic saturation assumes interview questions and sample characteristics remain constant over time[213]. Interviews were semi-structured, with the core questions remaining the same throughout all interviews and the participants were selected amongst fixed and defined categories of medical professionals. However, some variability cannot be excluded. Moreover, the Guest *et al.* method was not designed to assess the detection of differences between groups or study sites. It is therefore possible that the 12 interviews, despite reaching thematic saturation overall, missed some of the more granular differences between the junior and senior groups. Finally, a common pitfall when using thematic analysis is to succumb to anecdotalism[211], as described by Bryman[230]. This was addressed by providing a clear argumentation for each theme (see **Table II-4** to **Table II-7**), but could not be excluded.

II.4.2 Conclusion

In conclusion, this study identified the cognitive challenges, related difficulties but also coping strategies, faced and used by clinicians when attending postoperative complications. The results also highlight computerised support needs desired by practicing clinicians in this context. These findings provide a basis for further research in computerised support for perioperative care and will inform the objectives and design of the algorithm presented in Chapter IV of this thesis.

CHAPTER III

Machine learning-based clinical decision support systems and clinician diagnostic performance - a systematic review

This chapter is adapted from:

Vasey, B. et al. Association of Clinician Diagnostic Performance with Machine Learning-Based Decision Support Systems: A Systematic Review. JAMA Netw. Open 4, (2021) [85].

III.1 Introduction

Once the needs of clinicians, and more specifically junior surgeons, when facing postoperative complications were identified, it was necessary to gain a better overview of previously described machine learning (ML)-based clinical decision support systems (CDSS) and how they were evaluated. As mentioned previously, the algorithm proposed in this thesis was intended to integrate, by design, a component of human decision making, hence making the evaluation of the assisted human decision the main focus of evaluation. The objectives of the systematic review presented in this chapter were therefore to:

1. Investigate the impact of the use of ML-based CDSS on clinician diagnostic performance.
2. Identify existing ML-based CDSS used to improve the management (more specifically the diagnosis) of postoperative complications and already evaluated on their impact on human decision-making.
3. Investigate if specific CDSS design or implementation strategies (e.g. type of mathematical model used, presentation of outputs, timing of decision support) are associated with better performance.
4. Collect evidence to test the hypothesis, made in chapter II, that junior clinicians would benefit more from ML-based CDSS than their more experienced peers.
5. Collect information on how the human-CDSS interaction and user feedback were evaluated.
6. Investigate the influence of the human intelligence on the overall human-CDSS system performance (i.e. the change in performance between stand-alone CDSS output and assisted human performance)

ML-based CDSS are a category of Software as Medical Devices (SaMDs) designed to support health professional decision making by providing recommendations (for example diagnostic or therapeutic) learned by an algorithm from clinical data during a training process. Despite their name, CDSS have so far been evaluated mainly *against* human experts but rarely on their impact when used *with* human clinicians as an adjunct to human intelligence. Demonstrating that computers can be as good as humans for specific medical tasks has some useful applications, notably for large population screening where patients may otherwise not be able to see a healthcare professional in due time. Nevertheless, this approach neglects an important factor of any medical encounter, namely the human clinicians present.

As long as humans continue to hold the ultimate responsibility of signing off a diagnosis or a treatment plan, it will be their interpretation of a CDSS output, and not the output itself which will have an impact on patient care. Human decision making is known to be influenced by numerous external factors and cognitive bias[231]. It would therefore be unwise to assume without further evidence, that a human operator would always fully follow CDSS recommendations without questioning them. In order to investigate the actual clinical potential of a new ML-based CDSS, its performance should therefore be evaluated when used in interactive collaboration with a human clinician, and not solely on its stand-alone performance in silico (i.e. on a test dataset) against human experts, as commonly reported in studies and analysed in systematic reviews[60].

Previous systematic reviews have investigated the impact of CDSS on clinician performance or its surrogate clinical outcomes.[54,55,53,56,232] However, most of the included studies described hard-coded systems (i.e. systems in which all information necessary to calculate the recommendations is embedded within the systems by their developers, as opposed to modified by users or learned) or diagnosis generators based on manually curated knowledge repositories, hence not fully representing the true promise of ML - to become better than its creator by “learn[ing] without being explicitly programmed”.[59]

It is also unclear if the addition of human intelligence (i.e. the interpretation of the CDSS outputs by its user, leading to final decisions which may or may not reflect the CDSS outputs) usually increases or decreases the CDSS stand-alone performance (i.e. the performance achieved by the computer outputs without subsequent human intervention). Equally unclear is whether any specific output presentation or timing of support is consistently associated with better user performance.

At the time of starting this work, no systematic review had investigated the impact of using an ML-based CDSS on clinician diagnostic performance, nor the association between the presence of human intelligence, the CDSS output presentation and the timing of support with overall system performance.

III.2 Methods

The systematic review was conducted and reported according to the recommendations outlined in the PRISMA guidelines.[233] A study protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO) on the 24th of June 2019, with registration number PROSPERO 2019 CRD42019140075 and can be found in **Annex III-1**. I screened all abstracts and full texts for eligibility, extracted data and assessed risk of bias on all included studies, and performed the analysis. Benjamin Beddoe (Faculty of Medicine, Imperial College London), Nicole Bilbro (Department of Surgery, Maimonides Medical Center), Neale Marlow (Nuffield Department of Surgical Sciences, University of Oxford), Elliott Taylor (Nuffield Department of Surgical Sciences, University of Oxford), and Stephan Ursprung (Department of Radiology, University of Cambridge) contributed as independent second reviewers to duplicate each step of this process.

III.2.1 Search strategy

A search strategy built around four additive concepts ("machine learning", "decision support system", "clinician" and "performance evaluation") was designed with the support of a specialist librarian and can be found in **Suppl. Note III-1**. The strategy sought to identify all articles whose primary goal was to evaluate the change in clinician diagnostic performance when they worked with an ML-based CDSS, compared to their performance without CDSS. The original search was conducted on MEDLINE, Embase, and PsycINFO, using the Ovid search engine, for a period from the date of the database creation to May 31st 2019. The time window was subsequently limited to January 1st 2010 to May 31st 2019, due to changes in the designation of the targeted CDSS overtime and difficulties in retrieving some of the inclusion criteria in earlier studies (see section III.4.1 for more details). The initial search was conducted on May 20th 2019, and the last search to identify possible late indexation was conducted on June 1st 2020. For the main search and analysis, only peer-reviewed literature published in

English was considered. A grey literature search including conference abstracts, the World Health Organization International Clinical Trials Registry Platform, conference abstracts (from 2017 onward), and the Cochrane Central Register of Controlled Trials was also performed using an adapted search strategy (**Suppl. Notes III-2** to **III-4**). The initial research proposal was to include both diagnostic and therapeutic CDSS. However, after an initial screening, it was decided to limit the study to diagnostic CDSS, in order to reduce the heterogeneity of the included studies and focus on the larger of the two groups.

III.2.2 Screening

All retrieved titles and abstracts were independently screened by two reviewers, with at least one reviewer reviewing all the abstracts for consistency. Conflicts were adjudicated by a third reviewer. Full text articles were obtained and independently reviewed for eligibility by two reviewers, with the main reviewer reviewing all the articles for consistency. Conflicts were resolved by consensus. Inclusion criteria were:

- **study type:** primary, peer-reviewed research
- **population:** medical doctors from any specialties and levels of seniority, in both in- and outpatient settings, facing a clinical diagnostic decision.
- **intervention:** interactive use of a decision support system based on clinical data and machine learning algorithms to improve diagnosis or diagnostic investigations planning. Machine learning algorithms were defined as algorithms that have the ability to independently learn, from clinical data, knowledge unknown to their programmers and to generate outputs that have not been explicitly programmed. A CDSS was considered diagnostic if its output produced qualitative information (e.g. benign vs malignant) about the nature of a lesion or if the detection of a lesion was in itself sufficient to pose a diagnosis and influence a therapeutic choice (e.g. the presence of pulmonary emboli).
- **comparator:** medical doctors without the aforementioned decision support system. This includes studies where the same individuals had to perform a task with and without the decision support system.
- **outcome:** any metrics assessing performance, usability, trust, or other components of human-computer interaction.

Exclusion criteria were:

- case reports and case series
- monitoring, alert, or detection systems
- systems based on validated scores only
- systems based on natural language processing only
- systems based on manually curated knowledge repositories or set diagnostic rules
- systems based on models considered as general medical statistics, such as linear regression and logistic regression[a].
- studies using medical students as study population.

The abstract screening and full-text review were conducted using the Covidence software[234]. To avoid missing any algorithms published under commercial names, one round of forward and backward reference searching was conducted for all the included studies. An additional search was done using the names of 26 recently FDA approved algorithms.[235]

III.2.3 Data extraction

Data were extracted using a tailored extraction table (see **Suppl. File III-1**). The following data were extracted if present:

- study population: number, specialty, seniority
- patient population: type of medical conditions, number of different hospital sites
- dataset: type of sample, sample size and number of events (for training and validation sets), independence of training and test sets
- experiment: task to be performed, experimental design, number of cases per physician, timing of support, gold standard comparison, familiarity with the system.
- main purpose of the decision support system

[a] Depending on the definition they use, some authors may consider linear and logistic regression as machine learning.

- system characteristics: mathematical model used, International Medical Device Regulators Forum (IMDRF) risk classification, type of support, attempts to increase the interpretability of the model (i.e. any additional information displayed by the system to contextualise the recommendation and the way it was generated).
- metrics of human performance: type and value of all the metrics used to assess human performance with and without the CDSS.
- metrics of computer performance: type and value of all the metrics used to assess the performance of the decision support system alone
- study funding: provenance
- existence of a published study protocol

The risk of bias for each included study was assessed using the QUADAS-2 tool as modified by Riches[236,54] and the ROBINS-I tool[237]. QUADAS-2 was used to assess the risk of bias of the CDSS's diagnostic accuracy validation, while ROBINS-I considered the publications as non-randomised comparative studies with the medical doctors as study participants. Data extraction and bias assessments were all conducted independently by at least 2 reviewers. Conflicts were resolved by consensus. Meta-bias was assessed by searching the WHO International Clinical Trials Registry Platform (ICTRP) and the Cochrane Central Register of Controlled Trials (CCRCT) for unpublished trials or evidence of selective reporting. Information about the origin of study funding and the presence of a protocol was also extracted and reported.

III.2.4 Analysis

A summary table of all metrics used to assess clinician performance was created. To allow more meaningful comparison between the included studies and interpretation of the results, the analysis was then limited to the ten most commonly used metrics. For each study (and each of the considered metric), the main reported result (i.e. subgroup results were not considered) was reported in a summary table. When results were presented both at case and at finding level (a case can have several findings, such as pulmonary emboli), only the case

level results were considered. For studies reporting on two different CDSS or the same CDSS in two different use modalities, several main results were considered for each metric. Narrative summaries were produced for all the six objectives of the systematic review. Subgroup analyses were performed to evaluate separately: studies with evaluation settings representative of a clinical environment (consecutive or non-augmented random patient sample and access to the usually available clinical data at the time of decision making), the effect of clinicians' experience level (experienced vs. novice), the mathematical model used, the type of output presentation to the clinician (single output vs. output + context information), and the reader paradigm (concurrent vs. second reader). In the concurrent reader paradigm, the model recommendations are displayed at the same time as the clinical data, while in second reader paradigm recommendations are displayed after the observer had a chance to make their initial decision based on the clinical data. Given the heterogeneity of the included studies no meta-analysis could be performed. Qualitative data will be presented descriptively as recommended in the PRISMA guidelines. All studies were included in the analysis, irrespective of their risk of bias.

III.3 Results

III.3.1 Description of the included studies

8112 titles were identified, of which 2774 were duplicates and 184 not in English. 5154 abstracts were screened of which 156 were selected for full-text review. Of the 156 studies assessed against inclusion criteria, 22 were included. 15 additional publications meeting the inclusion and exclusion criteria were retrieved from other sources, including forward/backward references search, trade names search, related literature references search and grey literature search. A total of 37 publications were extracted and assessed for risk of bias.

Figure III-1 summarises the search, screening, and inclusion processes.

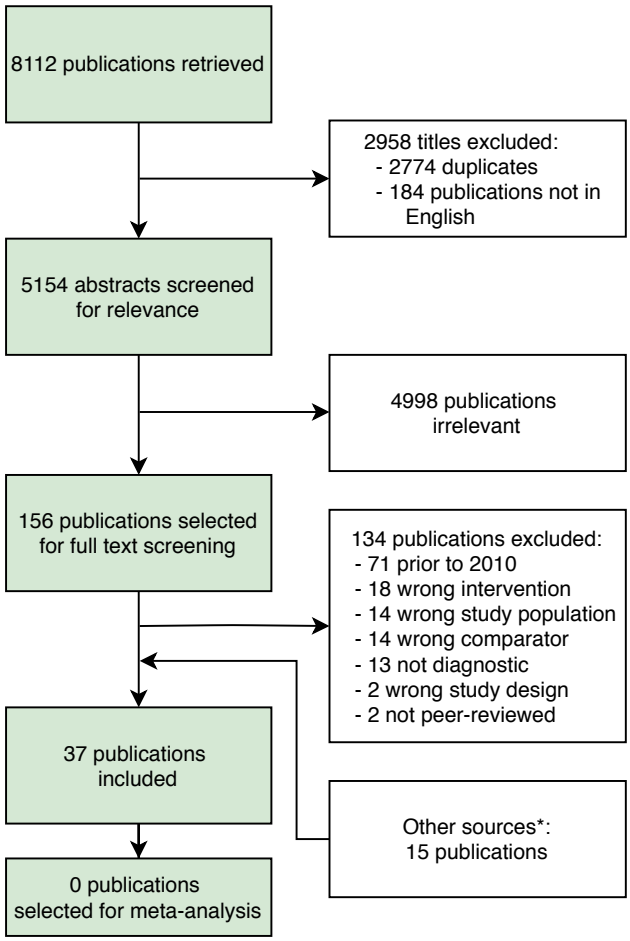


Figure III-1: PRISMA flowchart (adapted from Vasey *et al.*). *other sources included forward/backward literature search, reference search from relevant literature, trade name search, and conference abstracts or entries in the Cochrane Central Register of Controlled Trials.

First author	Year	Medical condition	algorithm used	IMDRF risk category	N° of sites	N° of users	Test set sample size ‡	Cases read / user	reader paradigm	private sector funding
Aissa[238]	2018	melanoma	ClearRead CT* (CNN)	4	1	3	46	46	first	NA
Aslantas[239]	2016	bone metastasis	perceptron-based ANN	4	2	1	130	130	NA	no
Bargallo[240]	2014	breast cancer	SecondLook*	4	NA	4	21321	8100§	second	NA
Barinov[241]	2019	breast cancer	cCAD* (ANN)	4	multiple	3	500	450 & 500	first & second	NA
Bartolotta[242]	2018	breast cancer	S-Detect* (CNN)	4	NA	4	300	300	second	NA
Bien[243]	2018	knee MSK injury	CNN	3	2	9	120	120	first	NA
v.d. Biggelaar[244]	2010	breast cancer	SecondLook*	3	1	2	1048	524§	first	no
Blackmon[245]	2011	pulmonary embolism	VA10 PE* (SVM)	4	NA	2	79	79	second	NA
Cha[246]	2018	bladder cancer	CNN	4	NA	12	123	123	second	No
Chabi[247]	2012	breast cancer	B-CAD* v.2	4	1	4	160	160	second	NA
Cho[248]	2017	breast cancer	S-Detect* (CNN)	4	1	2	119	119	second	yes
Choi J.-H.[249]	2018	breast cancer	S-Detect* (CNN)	4	1	4	200	100§	second	no
Choi J. S.[250]	2019	breast cancer	S-Detect* (CNN)	4	1	4	253	253	second	no
Cole[251]	2014	breast cancer	ImageChecker* v.1.0 & SecondLook* v.1.4	4	multiple	15 & 14	300 & 300	300 & 300	second	no
Endo[252]	2012	pulmonary nodule	Euclidian distance clustering	4	1	3	30	30	NA	NA
Engelke[253]	2010	pulmonary embolism	PE-CAD*	4	NA	4	58	58	second	NA
Giannini[254]	2017	prostate cancer	SVM	4	NA	3	89	89	first	no
Hwang[255]	2019	thoracic pathology	CNN	4	1	15	200	200	second	no
Lindsey[256]	2018	wrist fracture	CNN	3	1	24	300	300	second	yes
Park[257]	2019	breast cancer	S-Detect* (CNN)	4	1	5	100	100	second	no
Rodríguez-Ruiz[258]	2019	breast cancer	Transpara* v.1.3.0 (CNN)	4	2	14	240	240	NA	yes
Romero[259]	2011	breast cancer	Image Checker* v.5.4 (CNN)	4	1	2	9389	4695§	second	NA
Samulski[260]	2010	breast cancer	Image Checker* v.8.0 (CNN)	4	NA	9	120	120	second	no
Sanchez Gomez[261]	2011	breast cancer	SecondLook* v.1.1	4	NA	6	21855	3643§	second	no
Sayres[117]	2019	diabetic retinopathy	CNN	3	multiple	10	1796	1796	first	yes
Shimauchi[262]	2010	breast cancer	Bayesian ANN	4	2	6	60	60	second	NA
Sohns[263]	2010	breast cancer	Image Checker* v.2.3 (CNN)	4	NA	2	303	303	first	NA
Steiner[116]	2018	breast cancer	CNN	4	2	6	70	70	first	yes
Stoffel[264]	2018	breast cancer	ViDi Suite* v.2.0 (ANN)	4	1	4	33	33	first	no
Sun[265]	2014	atrial thrombus	ANN	3	1	5	130	130	second	no
Sunwoo[266]	2017	brain metastasis	k-means clustering + ANN	4	1	4	60	60	second	no
Tang[267]	2011	ischemic stroke	ANN	4	multiple	6	71	40	second	no
Taylor[268]	2018	parkinsonian syndromes	SVM	3	1 & multiple	2	55 & 100	55 & 100	second	no
Vassallo[269]	2018	lung metastasis	SVM	4	1	3	225	225	second	no
Watanabe[270]	2019	breast cancer	cmAssist* (CNN)	4	1	7	122	120	second	yes
Way[271]	2010	lung malignancies	linear discriminant analysis	4	1	6	256	NA	second	no
Zhang[272]	2016	lymph node malignancies	SVM	4	1	10	178	178	second	no

Table III-1: characteristics of included studies (adapted from Vasey *et al.*). *commercial name of proprietary algorithm, ‡for the observer test with CDSS support, ¶each case was either seen once or multiple times (with/without assistance) depending on the study, §on average, ANN = artificial neural network, CNN = convolutional neural network, IMDRF = International Medical Device Regulators Forum, MIC = multiple instance classifier, MSK = musculoskeletal, NA = not available, SVM = support vector machine.

All included studies described CDSS based on imaging modalities, with breast and pulmonary pathologies being the most represented medical conditions. 31 studies (84%) assessed a CDSS belonging to the International Medical Device Regulators Forum's risk category 4, the highest[142]. 20 studies (54%) investigated CDSS technology with a designated trade name at the time of publication. None were related to the management of postoperative complications.

Table III-1 gives an overview of the included studies' characteristics. The median number of participants in the observer test was 4 (interquartile range (IQR): 3 to 8, mean: 7), each reading a median of 123 different cases (IQR: 79 to 300, mean: 635). 18 (49%) studies were based at a single centre, 10 (27%) were multicentric and 9 (24%) did not specify. The full description of study characteristics can be found in **Suppl. File III-1**. The median proportion of events (i.e. target condition or pathology) in the test set was 44% (IQR: 32% to 54%). Only 13 studies (35%) reported a consecutively or randomly selected case sample. 3 studies reported on more than one CDSS, or used the same CDSS in different modalities

III.3.2 Studies quality and risk of bias

Using QUADAS-2, 28 studies (76%) were rated at high risk of bias in at least one of the four core domains and none were considered at low risk of bias in all four core domains. "Patient selection" and "index test" (i.e. the diagnostic tool evaluated) were the two domains most contributing to the studies high risk of bias. Using ROBINS-I, six studies (16%) were rated at serious or critical risk of bias, one due to cofounding, two due to deviation from the intended interventions, and three due to likely selection of the reported results. Only one study (3%) was considered at low risk of bias in all seven domains[265]. **Figure III-2** shows the risk of bias assessment for each category of the two tools. The detailed risk of bias marking for each study can be found in **Suppl. Table III-1** and **Suppl. Table III-2**.

The grey literature search identified 1 randomised clinical trial protocol with expected completion date outside of the review's search period, 1 conference abstract which led to a publication after the review's search period[273], and 1 conference abstract which did not lead to any publication. 6 studies (16%) reported private sector funding and 12 (32%) gave no or unclear information about their source of funding. Only 2 studies (5%) referenced a study protocol[265,272].

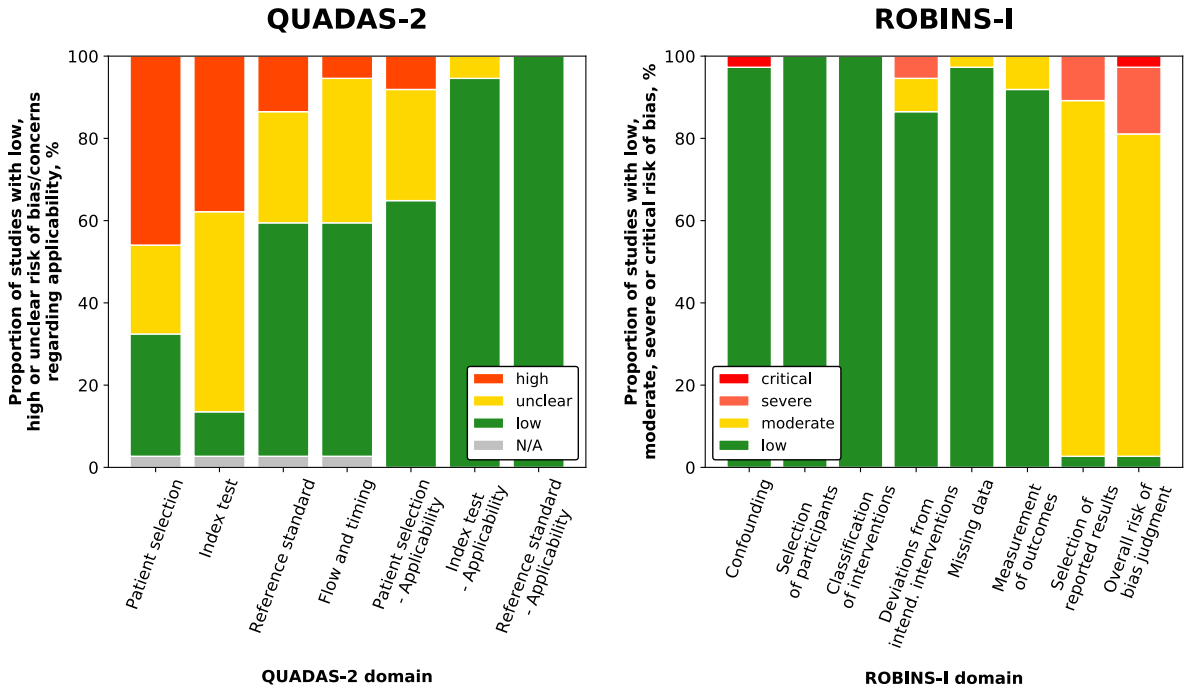


Figure III-2: risk of bias assessment. Distribution of the risk of bias scores over the QUADAS-2 and ROBINS-I domains. 100% represents the 37 included studies.

III.3.3 Association of CDSS use with clinician diagnostic performance

The ten most common categories of metrics used to quantify human performance were: sensitivity (81%), specificity (70%), area under the receiver operating curves (AUC, 51%), accuracy (38%), interobserver agreement (30%), positive predictive value (PPV, 30%), negative predictive value (NPV, 30%), reading time (22%), rate of recall for further investigation (11%), and the positive predictive value of further investigations (8%). **Table III-2** presents a full list of evaluation metrics with their occurrence.

Metric used	Occurrence	Metric used	Occurrence
Sensitivity/detection rate or number	30	True positive fraction for a given false positive fraction's interval	1
Specificity/number of false positive	26	Positive predictive value at x % prevalence	1
Area under the curve (ROC, JAFROC)	19	Negative predictive value at x % prevalence	1
Accuracy (binary, standard deviation or percentual scoring error)	14	Subjective "obviousness score"	1
Interobserver agreement/variability (Kappa, Kendall's tau, ICC, Blant & Altman, standard deviation of estimates)	11	Accuracy (multi-reader congruent diagnosis)	1
Positive predictive value	11	Sensitivity (multi-reader congruent diagnosis)	1
Negative predictive value	11	Specificity (multi-reader congruent diagnosis)	1
Reading time/time to decision in second	8	Failure to detect at least one nodule	1
Rate/number of patients recalled for further investigations	4	Detection of at least one false positive	1
Positive predictive value of further investigations	3	Confidence	1
Correct clinical management	1	Severity stratification	1
Lesion stage/class (radiological, pathological, or clinical)	2	Overestimates	1
Number of discarded computer flag	1	Underestimates	1
Diagnostic odd ratio	1	Change in recommended action	1
% of a specific subgroup amongst the diagnosed lesions	1	Complete agreement of management recommendations	1

Table III-2: metrics used to evaluate the impact of ML-based CDSS on human performance. The number of occurrences represent the number of studies using the metric for at least one analysis. ROC = receiver operating characteristic, JAFROC = jackknife free-response receiver operating characteristic.

Out of the 107 results with reported statistical significance, 54 (50%) reported a significant increase in the observed metric overall or for half or more of the individual participants, 4 (4%) a significant decrease and 49 (46%) no significant difference overall or no clear pattern in individual participants' results. 36 additional results were reported without mentioning statistical significance. Most studies defined statistical significance at $p<0.05$ with some applying correction for multiple comparisons. The AUC, accuracy and interobserver agreement were mainly increased, the sensitivity and positive predictive value were similarly increased or unchanged, while the specificity, negative predictive value, reading time, rate of recall for further investigation and PPV of further investigations remained unchanged in most cases. **Table III-3** shows a summary of CDSS use's impact on the ten most used evaluation metrics.

In the six studies[240,242,244,249,259,261] evaluating CDSS in a representative clinical environment (see definition in III.2.4), and when considering the ten most used evaluation metrics as previously listed, 20 results were reported with an assessment of the statistical significance. 16 (80%) showed no difference in performance, while 4 (20%) showed an increase in either sensitivity, AUC, interobserver agreement or PPV (see **Table III-4**).

	Results reported with statistical significance			Results reported without statistical significance			
	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≦ 50% of the participants	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≦ 50% of the participants	Total CDSS* evaluated
Sensitivity	10	11	1	9	1	0	32
Specificity	6	11	1	3	2	5	28
Area under the curve	13	7	0	1	0	0	21
Accuracy	8	4	0	5	1	0	18
Interobserver agreement	7	2	0	2	0	0	11
Positive predictive value	5	3	0	2	0	2	12
Negative predictive value	3	5	0	3	1	0	12
Reading time	2	2	2	0	1	1	8
Recall for further investigations	0	2	0	1	0	0	3
PPV of further investigations	0	2	0	0	0	1	3

Table III-3: association between ML-based CDSS use and clinician diagnostic performance (adapted from Vasey *et al*.). Number of results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to clinicians alone. PPV = positive predictive value. *Three studies reported on more than one CDSS (or used the same CDSS in different modalities).

	Results reported with statistical significance			Results reported without statistical significance			
	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≦ 50% of the participants	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≦ 50% of the participants	Total CDSS evaluated
Sensitivity	1	3	0	2	0	0	6
Specificity	0	2	0	1	0	1	4
Area under the curve	1	1	0	0	0	0	2
Accuracy	0	2	0	0	0	0	2
Interobserver agreement	1	1	0	0	0	0	2
Positive predictive value	1	1	0	1	0	1	4
Negative predictive value	0	2	0	1	1	0	4
Reading time	0	0	0	0	0	0	0
Recall for further investigations	0	2	0	1	0	0	3
PPV of further investigations	0	2	0	0	0	1	3

Table III-4: association between ML-based CDSS use and clinician diagnostic performance for the six ML-based CDSS evaluated in representative clinical environment. Number of results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to clinicians alone. PPV = positive predictive value.

A detailed list of the included studies results, including effect size and statistical significance, can be found in **Suppl. Table III-3**.

III.3.4 Influence of system design and implementation strategy on clinician diagnostic performance

25 studies (67%) used a second reader paradigm for the CDSS support, 8 (22%) used a first reader paradigm, 1 (3%) used both and 3 (8%) did not specify. 29 (78%) studies described CDSS based on neural networks and 8 (22%) on other mathematical models. 22 CDSS (59%) displayed only a single main output, whereas 15 (41%) displayed additional explanatory information to contextualise the main output. The additional explanatory information displayed included: selected input features, heatmaps of features importance, the underlying probability distributions of the outputs, and per pixel/voxel disease probability maps (**Suppl. Table III-4**). **Table III-5**, **Table III-6** and **Table III-7** present the results of the subgroup analysis according to the reader paradigm, mathematical model used, and presence of explanatory outputs, respectively. The total number of results in each cell is equal to the number of results in the corresponding cell in **Table III-3**. More of the studies presenting a second reader CDSS appear to show a significant increase in performance than those presenting a first reader CDSS. No clear pattern can be seen in the other two subgroup analyses.

	Results reported with statistical significance			Results reported without statistical significance			
	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Total CDSS* evaluated
Sensitivity	1 / 9	3 / 8	0 / 1	4 / 5	0 / 1	0 / 0	8 / 24
Specificity	0 / 6	4 / 7	0 / 1	0 / 3	0 / 2	3 / 2	7 / 21
Area under the curve	1 / 12	3 / 4	0 / 0	0 / 1	0 / 0	0 / 0	4 / 17
Accuracy	1 / 7	2 / 2	0 / 0	0 / 6	0 / 0	0 / 0	3 / 15
Interobserver agreement	2 / 5	1 / 1	0 / 0	1 / 1	0 / 0	0 / 0	4 / 7
Positive predictive value	0 / 5	2 / 1	0 / 0	0 / 2	0 / 0	0 / 2	2 / 10
Negative predictive value	0 / 3	2 / 3	0 / 0	0 / 3	0 / 1	0 / 0	2 / 10
Reading time	0 / 2	0 / 2	2 / 0	0 / 0	1 / 0	0 / 1	3 / 5
Recall for further investigations	0 / 0	0 / 2	0 / 0	0 / 1	0 / 0	0 / 0	0 / 3
PPV of further investigations	0 / 0	0 / 2	0 / 0	0 / 0	0 / 0	0 / 1	0 / 3

Table III-5: association between ML-based CDSS use and clinician diagnostic performance according to the reader paradigm (first reader/second reader). Number of results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to clinicians alone. PPV = positive predictive value. *Three studies reported on more than one CDSS (or used the same CDSS in different modalities).

	Results reported with statistical significance			Results reported without statistical significance			
	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Total CDSS* evaluated
Sensitivity	7 / 3	9 / 2	1 / 0	9 / 0	0 / 1	0 / 0	26 / 6
Specificity	5 / 1	9 / 2	1/ 0	2 / 1	2 / 0	4 / 1	23 / 5
Area under the curve	11 / 2	6 / 1	0 / 0	1 / 0	0 / 0	0 / 0	18 / 3
Accuracy	6 / 2	4 / 0	0 / 0	3 / 3	0 / 0	0 / 0	13 / 5
Interobserver agreement	4 / 3	1 / 1	0 / 0	2 / 0	0 / 0	0 / 0	7 / 4
Positive predictive value	5 / 0	2 / 1	0 / 0	2 / 0	0 / 0	1 / 1	10 / 2
Negative predictive value	3 / 0	4 / 1	0 / 0	2 / 1	1 / 0	0 / 0	10 / 2
Reading time	0 / 2	2 / 0	1 / 1	0 / 0	1 / 0	1 / 0	5 / 3
Recall for further investigations	0 / 0	2 / 0	0 / 0	1 / 0	0 / 0	0 / 0	3 / 0
PPV of further investigations	0 / 0	2 / 0	0 / 0	0 / 0	0 / 0	1 / 0	3 / 0

Table III-6: association between ML-based CDSS use and clinician diagnostic performance according to the mathematical model used (neural networks/other models). Number of results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to clinicians alone. PPV = positive predictive value. *Three studies reported on more than one CDSS (or used the same CDSS in different modalities).

	Results reported with statistical significance			Results reported without statistical significance			
	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Total CDSS* evaluated
Sensitivity	6 / 4	7/ 4	0 / 1	4 / 5	1 / 0	0 / 0	18 / 14
Specificity	1 / 5	6 / 5	1 / 0	1 / 2	2 / 0	3 / 2	14 / 14
Area under the curve	5 / 8	4 / 3	0 / 0	1/ 0	0 / 0	0 / 0	10 / 11
Accuracy	3 / 5	0 / 4	0 / 0	3 / 3	0 / 0	0 / 0	6 / 12
Interobserver agreement	4 / 3	0 / 2	0 / 0	0 / 2	0 / 0	0 / 0	4 / 7
Positive predictive value	1 / 4	1 / 2	0 / 0	1 / 1	0 / 0	2 / 0	5 / 7
Negative predictive value	1 / 2	1 / 4	0 / 0	2/ 1	1 / 0	0 / 0	5 / 7
Reading time	1 / 1	2 / 0	1 / 1	0 / 0	1 / 0	1 / 0	6 / 2
Recall for further investigations	0 / 0	2 / 0	0 / 0	1 / 0	0 / 0	0 / 0	3 / 0
PPV of further investigations	0 / 0	2 / 0	0 / 0	0 / 0	0 / 0	1 / 0	3 / 0

Table III-7: association between ML-based CDSS use and clinician diagnostic performance according to the outputs' level of support (single output/explanatory output). Number of results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to clinicians alone. PPV = positive predictive value. *Three studies reported on more than one CDSS (or used the same CDSS in different modalities).

III.3.5 Influence of clinician experience on the CDSS effects

19 (51%) studies provided information on the differences in performance between different user levels of experience. For 7 out of the 8 metrics with reported results, the use of ML-based CDSS was suggested to have a stronger influence (mostly an increase in performance) on less experienced clinicians than on their more senior peers, as shown in **Table III-8**.

	increase more for juniors	increase more for experts	no difference	decrease more for experts	decrease more for juniors	Total CDSS evaluated
Sensitivity	8	1	1	0	0	10
Specificity	4	2	0	2	2	10
Area under the curve	9	1	2	0	0	12
Accuracy	7	1	0	0	0	8
Interobserver agreement	2	0	1	0	0	3
Positive predictive value	2	1	1	0	0	4
Negative predictive value	3	0	1	0	0	4
Reading time	1	1	1	1	0	4
Recall for further investigations	0	0	0	0	0	0
PPV of further investigations	0	0	0	0	0	0

Table III-8: association between clinicians' level of experience and performance changes when using ML-based CDSS. Number of main results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to clinicians alone. PPV = positive predictive value.

III.3.6 Influence of human intelligence on the overall performance

27 studies reported on the CDSS stand-alone performance, permitting an investigation of the human intelligence contribution to the final decision-making. In all but one unclear case[268], the users overrode some of the CDSS recommendations. Out of the 75 results reported using one of the top ten metrics, the human contribution changed the overall results in 70 analyses (90%) and made no difference in 5 (10%). In 45 analyses (60%), human intelligence input was associated with an increased metrics value, in 25 (33%) with a decreased metrics value, and in 5 (7%) with no change. Only 3 results[117,258] were reported with statistical significance, of which 1 showed no significant change, and 2 a significant increase in accuracy (**Table III-9**).

	Results reported with statistical significance			Results reported without statistical significance			
	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Increase overall or for ≧ 50% of the participants	No change or unclear trend	Decrease overall or for ≧ 50% of the participants	Total CDSS evaluated
Sensitivity	0	0	0	11	2	10	23
Specificity	0	0	0	11	2	5	18
Area under the curve	0	1	0	5	0	7	13
Accuracy	2	0	0	6	0	1	9
Interobserver agreement	0	0	0	6	0	0	6
Positive predictive value	0	0	0	0	0	0	0
Negative predictive value	0	0	0	4	0	2	6
Reading time	0	0	0	0	0	0	0
Recall for further investigations	0	0	0	0	0	0	0
PPV of further investigations	0	0	0	0	0	0	0

Table III-9: association between human contribution and system performance (adapted from Vasey *et al.*). Number of results reported for the ten most commonly used metrics, comparing computer-assisted clinicians to stand-alone computer. PPV = positive predictive value.

III.3.7 Human-CDSS evaluation and Human Factors

Several study designs were used to assess the changes in performance of users with and without a CDSS. In Sequential Read, the users performed the diagnostic task in one of the study arms (i.e. with or without CDSS support), before waiting for a wash out period, and repeating the same task in the other study arm. Intra-reader variability can be assessed by repeating some of the cases in the same study arm after the wash out period. In Crossed Read, users alternated between the study arms within the same reading session, before waiting for a wash out period, and repeating the same reading session with inverted study arm attribution. In Historical Control, users performing the diagnostic task with CDSS support were compared to their (or different users') performance in the past, before the introduction of the CDSS. Finally, in Back-to-back Double Read (only possible in second reader mode) a case was first reviewed without CDSS support and the initial decision recorded, then directly reviewed again with CDSS support and the final decision recorded. Most studies were multireader (i.e. each case was reviewed by several readers) allowing for the assessment of inter-rater variability. None of the included studies were randomised controlled diagnostic trials[274].

Out of the 37 studies included, 15 (41%) attempted to increase the interpretability of the model by presenting some explanatory information to contextualise the model's final output and 13 (35%) included some form of training prior to starting data collection. This could be through theoretical training, practice cases, a wash in period, or by selecting users with previous exposure to the CDSS. In the 8 studies using practice cases as training, the median number of cases was 25 (IQR: 10 to 37.5, mean: 422.5). 4 studies (11%) attempted to collect user opinion about the CDSS, of which only 3 (8%) gathered this feedback through a formalised process (see **Suppl. Table III-4**). Van den Biggelaar *et al.* asked study participants to indicate on their case evaluation forms if the CDSS marks "added valuable diagnostic information to their own original evaluation" but did not report specifically on this outcome[244]. Taylor *et al.* designed mixed open and closed question interviews to "provide an insight into the CADx-

radiologist relationship [and] to assess the effects of the CADx software on clinician decision-making"[268]. The study participants reported a good correlation between their decision and the CDSS outputs, with a small to moderate impact on their reporting decision. They also considered it would be of small to moderate benefit if the CDSS would display more information on how it generates its decision, and thought CDSS could be of moderate to substantial benefit to support training and improve inexperienced clinicians' performance. Finally, Endo *et al*. invited the study participants to give direct feedback on the CDSS outputs by grading how relevant these were for the specific task at hand. 26 out of 30 best cases proposed by the system (87%) were judged visually similar to the test cases[252].

III.4 Discussion

This systematic review provides an overview of published literature, identifying two gaps, which the following chapters of this thesis will attempt to address. First, no ML-based CDSS focusing on the diagnosis of postoperative complications could be found. Second, there is a lack of methodological rigor in the evaluation of ML-based CDSS impact on human performance, most notably in the consideration of the human-computer interaction and human factors.

Expressing a straightforward judgment about the benefits of a CDSS is often difficult because it depends on other factors specific to each application, such as common clinical practice or the target conditions' prevalence. This is why the impact of CDSS on human performance was reported for each of the selected metrics, leaving the researchers to decide if specific changes are desirable in the context of their specialty. Overall, approximately half of the results where statistical significance was reported showed no difference between performance with and without CDSS support. When a difference was observed, the AUC, the accuracy and the interobserver agreement seemed to be the metrics most often influenced by CDSS use. When limiting the analysis to results of the 6 studies conducted in a representative clinical environment, most results showed no change in performance.

These results concur with the findings of several other studies assessing the impact of CDSS use for mammogram screening in large populations which showed no or arguable benefits in using these systems[275–277]. These studies were not included in the review, as the algorithms used were not described in enough detail to meet the inclusion criteria, but were likely based on machine learning models. These findings call for more caution when promoting the potential benefits of ML-based CDSS to improve patient care, and highlight the important differences between *in silico* and with-human performance evaluation.

CDSS use seems to be associated with a more marked impact on less experienced clinician performance compared to their senior colleagues, although such differences were rarely reported with statistical significance testing, and should therefore be interpreted with caution. CDSS use also appears to be associated with higher interobserver agreement between clinicians of all experience levels. These observations could support the development of CDSS which do not necessarily aim to “beat the experts” but instead endeavour to narrow the user performance variance, ideally skewing it towards the best performers.

It is very important to consider the results of this review in light of the risk of bias assessment. Most of the included studies were at high or unclear risk of bias, echoing the findings of a recent systematic review on the design and reporting standards of studies comparing deep learning model against clinicians[84]. This high risk of bias was mainly due to three aspects. First, the lack of prospectively or randomly selected case samples. Case samples randomly selected but enriched for target events were not considered as randomly selected for the risk of bias assessment, because they would likely not represent case samples encountered in real-life application. Second, the absence of clinical data otherwise available in real-life clinical settings when making a diagnosis. For example, knowledge about a patient family history and associated risk, would in most cases influence the classification of a suspicious lesion on a mammogram. Third, the absence of a protocol. Without a clear protocol it is very difficult to assess if the authors are reporting on the initially intended outcomes or have selected the most encouraging results to highlight the benefits of their CDSS.

It should also be noted that risk of bias assessment using the ROBINS-I tool only informs us about the risk of difference between the results of a non-randomised intervention study and its “target” randomised control trial. It says nothing about the generalisability of the findings[237]. An interventional study with the same individual clinician in both arms (once performing the diagnostic tasks with, and once without CDSS support) would have a minimum risk of bias, but would be almost meaningless in terms of external validity. Many studies had a very low number of users (median = 4) and made no mention of power calculations, undermining the

generalisability of their results. A confusion between statistical significance at patient level and statistical significance at practitioner level was also noted in the included studies. Bootstrapping the clinical cases reviewed to produce a p-value would, for example, not give any indication about the generalisability of findings to other clinicians using the same system. It would merely assess the likelihood of the same clinicians or group of clinicians to display similar changes in performance with a new sample of patients. Several included studies nonetheless offer good example of how to account for user variability by using mixed effects models[116,117,255,258,264] or the Dorfman-Berbaum-Metz method[271] when calculating significance.

In addition to the bias and statistical problems, a marked heterogeneity in the metrics used to assess CDSS, even amongst studies assessing CDSS with similar intended use, made a reliable comparison of these systems complicated, when possible at all. The issue of performance comparability is well known in the field, and led to the creation of data challenges, in which *in silico* algorithm performance is evaluated on common datasets[278,279]. However, these considerations have not yet been addressed at the stage of clinical evaluation and could benefit from standardised metrics, evaluation methods and reporting guidelines[105,280–282].

The findings in this review also highlighted a lack of consideration for the importance of human factors, despite clinicians being the end users of the systems tested. Only a third of the studies included user training or a wash in period. Given the likely existence of learning curves as reported by Rodríguez-Ruiz *et al*.[258], a failure to reach a stable performance level prior to the start of data collection would likely bias the study results, possibly underestimating the CDSS true effect. This echoes observations made in other fields, such as surgery[283], where the investigation of learning curves is part of innovation evaluation guidance[186,187]. Given the unique nature of AI-based decision support, in which two competing forms of intelligence have to collaborate, the notion of learning curves could be extended to trust curves. The evolution of trust could indeed be theorised as being potentially as important to the overall performance as the evolution of user skills in the handling of the CDSS. Two authors explicitly mentioned trust in their discussion[117,240], while several others considered trust related issues by

investigating the context in which users changed their mind after receiving CDSS support. User feedback was reported in only 4 studies, hindering any iterative improvement of the human-computer interaction. This strongly contrast with other safety-critical industries, like the aviation and energy sectors, where human factors principles, such as usability testing and cognitive task analysis, are routinely employed.

The need for more consideration of human factors was further supported by the observation that, in all but one study, where data about the algorithm stand-alone performance were available, the clinicians decided to override at least some of the CDSS recommendations. Even if the results could not make any conclusion on whether the contribution of the human intelligence was beneficial or detrimental to the overall decision-making performance, these elements make it clear that defining the effectiveness and safety profile of a new CDSS through computer simulation alone is insufficient. In real life situations, it is the clinician perception and processing of the CDSS outputs, and not the outputs themselves, which will have an impact on patient care. As long as clinicians have the final responsibility for a diagnostic or therapeutic choice, they will, consciously or not, factor in other variables in addition to the CDSS outputs, be influenced by legal considerations[114] and probably never fully trust all the CDSS recommendations. Therefore, the assisted decision-making process should be considered as the endpoint and focus of the evaluation, not merely the stand-alone CDSS performance.

III.4.1 Strengths and limitations

The methodology used for the review follows the best practice standards and every step of the process was independently performed by at least two reviewers. The review was the first to set human users, and not the algorithms themselves, at the centre of a systematic review on the utilisation of ML-based CDSS in clinical settings. It provided missing evidence about the actual potential of AI to improve clinician diagnostic performance, in contrast to predictions or promises, commonly portrayed by the media and private sector. As a unique collection of

studies evaluating the impact of ML-based CDSS use on clinician performance, the review also constitutes important material to inform the development of future guidance for the clinical evaluation of this technology.

Due to the heterogenous description of ML-based CDSS across medical specialties, the use of commercial names in some studies, and the fact that categorisation of this technology in specialised search engines is relatively recent ("machine learning" was added as a MeSH term on PubMed in 2016), it is possible that some relevant literature was not retrieved. A forward and backward literature search of the included studies, and an additional search for common or new commercial names were conducted, in an attempt to minimise this issue. This partly explains the high ratio of additional publications from other sources analysed, compared to publications identified through the initial literature search. Despite initially not limiting the search window in the past, it was decided, after starting the full text review, to confine it to studies from 1st January 2010 in all searched databases. The reasons for this change in the protocol were twofold. First, it was observed that the nomenclature describing the targeted ML-based CDSS became increasingly heterogeneous as time goes back, rendering a systematic retrieval and inclusion impracticable. Second, with a rapidly evolving industry, it was sometimes impossible to find information about early systems specifications due to changes in CDSS ownership and commercial names, which hindered the inclusion criteria's assessment. This protocol amendment was made before any data was extracted, excluding any conscious or undue selection bias.

Given the broad range of CDSS covered, inclusion criteria had to be precisely defined. Some of these definitions don't have broad consensus support in the literature and are therefore arguable. For example, differentiating between diagnostic (included) and detection (excluded) is not always trivial. In the context of this thesis, it was decided to consider a CDSS as diagnostic if its output produced qualitative information (e.g. benign/malignant) about the nature of a lesion, or if the detection of a lesion was sufficient in itself to pose a diagnosis and influence therapy (e.g. the presence of pulmonary emboli). In contrast, a pure detection tool

may identify a pulmonary nodule, but offer no suggestion about the likelihood of malignancy or next therapeutic steps. Finally, the qualitative evaluation of the included studies is a subjective process to some extent and its conclusions should be interpreted with the necessary caution. To reduce this subjectivity, the assessment of bias was conducted independently by at least 2 reviewers for each study, using validated tools.

III.4.2 Conclusion

This systematic review of the literature showed that there is so far no robust evidence that ML-based CDSS improves clinician diagnostic performance in representative clinical environments. It highlighted that most of the studies comparing clinician diagnostic performance with and without CDSS support are at high risk of bias, and have a low number of participants. It was also observed that, when the data were available, the users almost always overrode some of the CDSS recommendations, although no conclusion could be made about the influence of human intelligence on the overall performance. No reference to an existing ML-based CDSS focusing on the management of postoperative complication could be found. An association between user clinical experience level and the influence of ML-based CDSS was observed, indicating a stronger effect on junior clinicians. Little attention was given to training and the evaluation of human factors.

These findings support the initial aim of the thesis, to develop an algorithm to support junior clinicians in the management of postoperative complications. However, they also highlighted the low number of studies focusing on clinicians' performance improvement and the generally poor methodology used to evaluate this important aspect of the translation from *in silico* model development to real world impact on patient care. Guidance for reporting on the early stage of clinical evaluation could offer an actionable solution to the methodological problem identified, and encourage more researchers to undertake initial investigations in clinical settings.

CHAPTER IV

A patient similarity-based approach to postoperative complications: model development and retrospective testing

IV.1 Introduction

Building on Chapter II results and the findings that lists of possible actions (e.g. drug prescription, further diagnostic tests, escalation, etc) tailored to specific clinical situations are desired by clinicians, I intended to develop a clinical decision support system (CDSS) driven by artificial intelligence (AI), which could provide such recommendations in the context of postoperative complications. Through this process, I also sought to understand the challenges and limitations hindering the development of AI-based CDSS focused on the postoperative period.

AI is currently mainly based on machine learning (ML) algorithms for the processing of the acquired data and generation of recommendations or other intelligent outputs. ML is an heterogenous group of mathematical models, each with strengths and limitations depending on the task at hand and data presentation. Common approaches to clinical problems are to use either regression or classification models. For the work of this thesis however, I decided to take a clustering approach, which differs from classification in that it identifies similarities between objects (in this case a mathematical representations of patients), rather than optimising parameters' weights to make an accurate attribution to a defined category. The chosen approach will apply clustering at patient level, a concept described in the literature as patient similarity[284]. In other words, the desired model will build groups (the clusters) of similar patients, based on previous hospital admissions registered in a database. Each new case (the target patient) can then be attributed to the closest clusters and information about similar patients provided to the user, such as drug prescriptions or procedures booked, hence using past data to inform future decisions (see **Figure IV-1**). This approach has several advantages.

First and foremost, it mimics medical experience. Seasoned doctors often reason in terms of similar cases seen in the past. By imitating a familiar human thought process, the algorithm's underlying mechanisms will hopefully seem more understandable, increasing acceptability.

Second, it suits the nature of surgical complications, of which early stages often manifest with common features. Thus, the goal for clinicians, and therefore for a useful decision support tool, is not to identify a single diagnosis, but rather a list of potential causes and the appropriate courses of action to address them. By suggesting a cluster of similar patients, the algorithm will provide a spectrum of possible diagnoses, which the human physicians can further refine. Third, this approach is quite versatile and can be applied to most of the currently available datasets, independently of the number and quality of clearly recorded surgical complications outcomes, whose lack often prevent meaningful classification tasks for this specific subset of patients. Finally, this approach is well suited to situation of uncertainty, where a single truth cannot be inferred from the available information. AI support is shifted from providing a single correct recommendation to providing the most commonly chosen course of action in similar situation, a solution proposed by several authors as a promising avenue of research for clinical AI[285,286].

IV.1.1 Patient similarity in medicine

More generally, patient similarity can be seen as a way to bring more personalised medicine to the patient's bedside. Guidelines and evidence-based knowledge have so far built the basis of rational medical decision making. However clinical trials often exclude important groups of patients with common conditions, taking commonly prescribed medications or from large population subgroups based on age or sex[287]. Clinical trials, and hence the treatments guidelines derived from them, also do not always give detailed information about secondary effects or long-term safety[288]. Under these conditions, a source of knowledge based on individual patient's data and complementing existing guidelines has great potential to improve clinical decision making[289]. Patient similarity has also been proposed as a strategy to better identify clinical trials whose study population represents the target patient[290]. Furthermore, it has yielded promising results in identifying patient subgroups, for example in ICU or type 2 diabetes patients[291,292]

In 2018, Parimbelli *et al.* identified in a systematic review of studies using patient similarity models for precision medicine the four types of data most commonly used as inputs: molecular data, clinical data, imaging/biosignal data and laboratory data[293]. The same review highlighted cancer as the clinical domain to which the concept of patient similarity was applied most of the time, with a predilection for breast cancer. The study also interestingly pointed out the dynamism but also volatility typical of a still relatively new paradigm, with more than 25% of the referenced publications presenting a novel approach to translate similarity. The three main applications of patient similarity were treatment targeting, outcome-related subgroups identification and disease sub-phenotype discovery. The year before, Sharafoddini *et al.* published a scoping review focused on the use of patient similarity in prediction models based on health data[294]. The review covers 22 articles using a range of different mathematical approaches but does not provide clear evidence on the most efficient one. The main reason is probably the strong dependence between the method and the data type and structure. The most commonly used algorithms are those based on the nearest neighbourhood as measured with Euclidian or Mahalanobis distances. Only two studies compared the performance of similarity-based approaches to population-based predictive models, with both showing improved performance using similarity[295,296]. The publication pointed out the need for further research on predictors selection and informative missingness, or in other words, the fact that specific patterns of missing data can be informative for a model.

Comparable approaches are already used in other fields. In the music and movie industries, for example, collaborative filtering is commonly used to generate personalized recommendations[297,298]. Engineering sciences use “digital twins” for the diagnosis of jet engines[299]. By coupling sensors of an actual engine to a digital model of this engine running in accelerated time (the twin), weaknesses of the real system can be predicted and at-risk components changed before they actually break. Although still rarely used in clinical medicine due to a lack of powerful enough models, the concept of digital twins offers promising avenues of research for personalised medicine in the future[300].

IV.1.2 AI for the management of postoperative complications

Despite growing interest for the application of patient similarity to different fields of medicine, none of the papers included in the reviews mentioned above focused on surgical patients. A complementary, non-exhaustive, literature search also found no study using patient similarity to provide direct support regarding postoperative patient management. Maheshwari *et al* used topographical analysis to identify groups of similar patients based on the treatment they received after colorectal surgery. Although not aimed at directly influencing clinical management for individual patients, the authors claimed that their system could improve the overall care pathways by identifying successful strategies[301]. Zhou *et al.* presented DoseGuide, a system using several neural networks to cluster patients based on dynamic intraoperative data to predict postoperative pain. Even if the objective of the system is to improve the postoperative evolution of the patients, its application remains intraoperative[302].

More broadly, several publications proposed ML approaches to the prediction or detection of surgical complications. For example, Stam et al. published a systematic review of 15 ML-based algorithms aimed at predicting surgical complication after major abdominal surgeries[303]. The MySurgeryRisk algorithm was trained to predict the occurrence of a range of postoperative complications using generalised additive models and random forests[304]. Soguero-Ruiz *et al.* proposed an algorithm based on support vector machine to predict anastomotic leakage from free text (clinical notes, radiology reports, discharge letters, etc), blood tests and vital signs[305]. Pimentel *et al*. used a model based on abnormality detection to identify patient deteriorating after upper-gastrointestinal surgery[37]. In addition, numerous ML models have been proposed for deterioration detection in general ward[306]. Even if they are not specific to surgical patients, these models are often applied to this population as well.

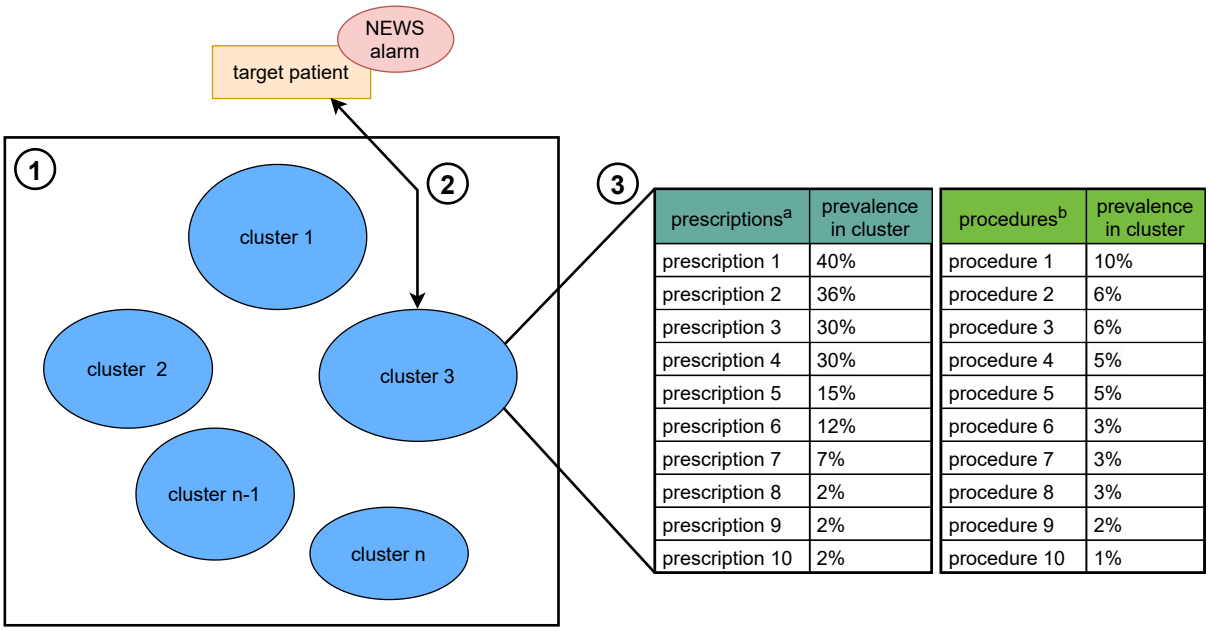
However, none of the systems identified so far provided ML-based bedside decision-making support regarding the immediate management of surgical complications, which is the target intended use of the model presented in this chapter.

IV.1.3 The HAVEN dataset

The data used to train and test the present model come from the Hospital Alerting Via Electronic Noticeboard (HAVEN) project. The project, started in 2014, was funded by the Health Innovation Challenge Fund, a collaboration between the Wellcome Trust and the Department of Health. HAVEN is a relational database of demographic, process and clinical data retrospectively collected in the Portsmouth Hospitals NHS Trust and Oxford University Hospitals NHS Trust (extended to the Royal Berkshire NHS Foundation Trust, Lancashire Teaching Hospitals NHS Foundation Trust and South Warwickshire NHS Foundation Trust since 2020, although not directly relevant to this thesis). HAVEN contains data on 346,934 hospital admissions between January 1st 2016 and December 31st 2017, with over 19 millions datapoints for vital signs and 24 millions laboratory values (including point of care testing). It led to the development of a ML-based algorithm for the early detection of in-hospital adverse events outperforming existing scoring systems[36] and of an age-specific early warning score, improving the detection of deterioration in patients between 16 and 45 years of age[307]. The HAVEN dataset was also used to evaluate the performance of the National Early Warning Score (NEWS) 2 compared to its predecessor in identifying patient at risk of in-hospital mortality or other adverse outcomes, showing no improvement or even significantly lower discrimination for patient considered at risk of type II respiratory failure[38].

IV.1.4 Objectives

The primary task of the present model is to develop a decision support system to improve the management of postoperative complications by providing prescription recommendations to the attending clinicians. To this end, the model should in a first step build clusters of similar patients from a sample of the HAVEN database (general surgery patients triggering a first NEWS alarm on the ward during the postoperative period). The model should secondly match any new patient triggering a first alarm during their postoperative stay on the ward (the target patient) to its most similar cluster (i.e. the group of past patients that the clustering algorithm identified as most similar to the target patient based on the provided input features). In a third step, the model should extract the desired information (the top ten drug prescriptions in the cluster) and display them to the attending clinician. (**Figure IV-1**). The aim is not to recommend a unique combination of prescriptions, but rather to display a shortlist of prescriptions for the attending clinician to consider. Extending on this primary objective, the secondary objective is to provide the same type of shortlist for the top ten most likely medical procedures or imaging exams to be booked in the same situation.



prescriptions[a]	prevalence in cluster
prescription 1	40%
prescription 2	36%
prescription 3	30%
prescription 4	30%
prescription 5	15%
prescription 6	12%
prescription 7	7%
prescription 8	2%
prescription 9	2%
prescription 10	2%

procedures[b]	prevalence in cluster
procedure 1	10%
procedure 2	6%
procedure 3	6%
procedure 4	5%
procedure 5	5%
procedure 6	3%
procedure 7	3%
procedure 8	3%
procedure 9	2%
procedure 10	1%

Figure IV-1: schematic representation of the model's functioning. (1) Clusters of similar patients are built from retrospective data (in this case, from patient in the training set). (2) New target patients (in this case patients on the test set), or more precisely their first contact point triggering a NEWS alarm, are matched with the most similar of the pre-established clusters. (3) A list of the ten most common prescriptions and a list of the ten most common procedures within this cluster are presented to the attending clinician to orient their decision making. [a] prescriptions within 24h of a NEWS alarm; [b] procedures performed on day 1 after the alarm (see IV.2.1 for more details).

IV.2 Methods

All analyses and graphs were produced using Jupyter Notebook v6.0.0 (running Python v3.7.4) and DBeaver v6.2.0 running MySQL 8+.

All relevant codes can be accessed in **Suppl. Files IV-1** and **Suppl. Files IV-2**.

IV.2.1 Data acquisition and extraction

The primary patient data used for the analysis were extracted from the HAVEN database. The HAVEN project was approved by the Health Research Authority research ethics committees in Portsmouth and Oxford (reference number 08/02/1394 and 16/SC/0264 respectively) as well as by the Confidentiality Advisory Group (16/CAG/0066). The use of the data for the present work was approved by the Critical Care Research Group (CCRG) Data Access Committee on the 24th of July 2019. All primary data are deidentified and stored on a secured server accessible through a virtual machine. The data in the HAVEN dataset were collected from different hospital IT systems: the SEND systems[221] for vital signs, laboratory reports, point of care testing (POCT) results, and administrative information (such as patient demographics). The exact devices used to measure laboratory and POCT values are heterogenous and not always identifiable. Overall, the HAVEN database contains 224,405 hospital admission with at least one visit to theatre. For the work of this thesis, the following selection criteria were applied at hospital admission level:

- Admissions in the Oxford University Hospitals NHS Trust (the John Radcliffe Hospital - a university teaching hospital, the Churchill Hospital - a specialist cancer hospital, and the Horton General Hospital - a district general hospital);
- Admissions between January 1st 2016 and December 31st 2017;
- Admissions of patients 16 years old and above with at least one theatre visit;
- Admissions with NHS treatment function code: general surgery (100), colorectal surgery (104), hepatobiliary and pancreatic surgery (105), upper gastrointestinal surgery (106), gastroenterology (301), hepatology (306), or accident and emergency (180) combined with a consultant specialty code in general surgery (100);
- No admissions with same day discharge.

General surgery patients were selected as this type of surgery has amongst the highest complication burden in high-resource healthcare settings[308]. The inclusion period was selected to start with the completion of the SEND system deployment in all wards of the included hospitals and end with the latest database update.

Input features

Table IV-1 summarises the variables used as input features to the model. All vital signs needed to calculate the NEWS were extracted, including the estimated FiO2 fraction (calculated based on previously published estimation, depending on the mask type[309–312]). Available laboratory variables which were recorded for less than 1% of the observation sets were not considered. Common laboratory features were laboratory features whose missingness did not exceed 66.66% in the target sample after fetching values in a 26h window period prior to a contact point (see section IV.2.2 for more details). A detailed account of laboratory features missingness can be found in **Suppl. Table IV-1**. Admin/demographic variables were either extracted or calculated: time since the end of the first operation, time since the end of the last operation (in case of multiple visits to theatre), the Charlson comorbidity index[313] (CCI), and age. Although two important variables, sex and operation type (emergency vs list) were not included because the clustering algorithms used do not perform well with binary data. The data used for this study can be accessed through request to the Critical Care Research Group Data Access Committee, Nuffield Department of Clinical Neurosciences, University of Oxford.

Three sets of features were tested as input:

1. All features
2. Vital signs + common laboratory values + admin/demographic variables
3. Vitals signs and associated variables only (as used to calculate the NEWS)

	Features
Vital signs + associated variables	Heart rate – HR [bpm], Respiratory rate – RR [rpm], Systolic blood pressure – SBP [mmHg], Diastolic blood pressure – DBP [mmHg], Oxygen saturation – SPO2 [%], Temperature – TEMP [°C], Consciousness – AVPU [A-V-P-U], Estimated fraction of inspired oxygen – ESTIMATED_FIO2_2 [%]
Common laboratory variables	Basophil count, blood – BAS [10^9/L], Base excess, blood – BASEX [mmol/L], Bicarbonate, plasma – BICAR [mmol/L], Calcium, plasma – CAL [mmol/L], Chloride, plasma – CL [mmol/L], Creatinine, plasma – CR [µmol/L], Estimated GFR, blood – EGFR [mL/min/1.73m2], Eosinophil count, blood – EOS [10^9/L], Glucose, blood – GLU [mmol/L], Haematocrit, blood – HCT [%], Haemoglobin, blood – HGB [g/dL], Lactate, plasma – LAC [mmol/L], Lymphocyte count, blood – LYM [10^9/L], Mean corpuscular haemoglobin, blood – MCH [pg], Mean corpuscular haemoglobin concentration, blood – MCHC [g/L], Mean corpuscular volume, blood – MCV [fL], Monocyte count, blood – MON [10^9/L], Mean platelet volume, blood – MPV [fL], Neutrophil count, blood – NEU [10^9/L], pH, blood – PH [no unit], Platelet count, blood – PLT [10^9/L], Potassium, plasma – POT [mmol/L], Red blood cell count, blood – RBC [10^{12}/L], Sodium, plasma – SOD [mmol/L], Urea, plasma – UR [mmol/L], White blood cell count, blood – WBC [10^9/L]
Further laboratory variables	Albumin, plasma – ALB [g/L], Alkaline phosphatase, blood – ALP [IU/L], Alanine aminotransferase, plasma – ALT [IU/L], Activated partial thromboplastin time, blood – APTT [s], Adjusted calcium, plasma – COCA [mmol/L], C-reactive protein, plasma – CRP [mg/L], Prothrombin international normalised ratio, blood – INR [no unit], Ketones, blood – KET [mmol/L], Methaemoglobin, blood – METHB [%], Osmolality, blood – OSM [mosmol/kg], Total bilirubin, plasma – TBIL [µmol/L]
Admin/demographic variables	Age – age [years], Charlson comorbidity index – cci [no unit], time since first operation – post_firstOP_time_hours [hours], time since last operation – post_lastOP_time_hours [hours]

Table IV-1: input features. Full name of feature – abbreviation [units]

Outcomes of interest

Several outcomes of interest were defined and are described in **Table IV-2**. Unplanned ICU admission, cardiac arrest, re-admission to theatre and death were considered critical events. This corresponds to an adaptation to surgical patients (addition of re-admission to theatre) of a definition otherwise largely accepted for general ward patients[35,36].

Outcome	Definition
Unplanned ICU admission	Any unplanned ICU admission in the 24 hours following the observation set
Cardiac arrest	Any cardiac arrest in the 24 hours following the observation set
Death	Death of patient in the 24 hours following the observation set
Re-admission to theatre	Any re-admission to theatre in the 24 hours following the observation set
Prescribed drugs	Any drugs prescribed in the 24 hours following the observation set
Prescribed procedures/imaging	Any procedures (including those performed in theatre) or imaging exams booked on the day following the observation set*

Table IV-2: outcomes description. *this is due to the metadata of the procedure/imaging exam records, which only include the day they were performed and not the exact time.

Prescribed drugs were grouped in categories according to their target conditions or pharmacological properties (e.g. antibiotics, COPD medication, etc), as previously done[314]. Likewise, booked procedures/imaging were grouped according to their nature. The details of the groupings and exclusions can be found in **Suppl. File IV-3**.

IV.2.2 Pre-processing

All observation sets (i.e. vital signs recordings during ward rounds) with at least six out of eight vital signs and associated values recorded were considered as contact points with the patient. Only observation sets registered postoperatively in the ward were considered, excluding observations from the ICU, perioperative rooms, and observations registered less than one hour after the end of the operation[b]. For each contact point, a vector was created with vital signs and laboratory values either populated or imputed. Missing values were considered missing-not-at-random[315] and the imputation process designed accordingly. Missing values were imputed according to the following rules in the order specified (subsequent rules where only used if the previous ones did not yield a result), where x is the feature value, t_0 is the time of contact, X_i is the set of all x for an hospital admission i, and X_{pop} is the set of all x in the included population:

a) for vital signs and associated variables:
    i. linear interpolation using the last and next available value[c]
    ii. if X_i is not empty: $x = x_{t0-1}$ (select the last available value)
    iii. $x = \text{median}(X_{pop})$ (takes the population median)

b) for laboratory values:
    i. if {xt} where t0-26h ≤ t ≤ t0 is not empty: x = xtmax (select last value in a 26-hour time window)
    ii. x = median(Xi) (takes the median value of the hospital admission so far)
    iii. x = random draw from a Gaussian distribution of the feature's physiological range

[b] Pre-operative laboratory values were also considered for the imputation of missing values.
[c] This is based on the assumption that vital signs are easy to collect and could rapidly be completed if missing in real-life use. Therefore, there is an argument for having the best possible estimation of the missing value in the dataset, justifying the use of future data.

The 26-hours collection time window defined in rule b-i was established based on common hospital practices and on an analysis of the interval between data collection (see **Figure IV-5**). Rule b-iii is based on the assumption that pathological lab values are overrepresented in clinical datasets because lab tests are mainly ordered when suspected to be abnormal. A population median imputation would therefore probably skew the data toward pathological values.

Additional min-max and delta features (defined below) were generated for vital signs and delta features only for laboratory values. The addition of these features is based on previous work by Shamout *et al.*[316] and Youssef *et al.*[309], showing increased performance for deterioration detection in similar contexts when using the information encoded in these features as discrete proxies for the variables' evolution over time. The min-max value of a vital sign for a given contact point was defined as the absolute difference between the lowest and highest value registered in a 24h window prior to the contact point. The delta value of a vital sign for a given contact point was defined as the difference between the vital sign's 12h median value at the time of observation and 6h prior. The delta value of a laboratory results for a given contact point was defined as the difference with the last recorded value within a 50h time window, scaled on a 24h period to allow better comparison. In case there was no comparator in this time period, or that the laboratory value was imputed, the delta value was defined as the difference with the median of the admission's previous values or the mean of the physiological range if no previous values were available.

For the clustering task and subsequent performance analysis, only a subset of the data was considered, namely "snapshots" of patient hospital admissions corresponding to the first suspicion of a postoperative complication. Indeed, these theoretically represent the first time a ward doctor is called to a given patient to define a care strategy (including not doing anything) for a suspected complication. For each contact point, the NEWS was calculated from the recommended formula[34]. Observation sets triggering a NEWS alarm, namely reaching an overall NEWS score of ≥5 or a score of three in any of the individual component[34], were

selected and, to avoid the over representation of the sickest patients, only the first alarm trigger in each hospital admission was included in the sample of interest[d]. These selected contact points will be referred to as "contact points of interest" from here on and build the instances (i.e. the rows) of the dataset used to train and test the developed model. The dataset was then split into a training (80%) and a test (20%) set using the GroupShuffleSplit() function from the Scikit-learn package[317], grouping by patient ID to avoid data leakage from patients having multiple hospital admissions. The training and test sets were then independently standardised using the StandardScaler() function from the Scikit-learn package[317]. The function is described as $z = (x - \mu) / \sigma$, where z is the standardised value, x the original value, μ the mean and σ the standard deviation.

IV.2.3 Dimensionality reduction

An imbalance between the number of features used to train a model and the number of instances included (too many features for a limited sample size) can lead to overfitting. In addition, as the number of dimensions increases the distribution of the pairwise distances between different data points tends to narrow (also called distance concentration), which is particularly a problem for the performance of distance-based clustering algorithms.

For these reasons, it was decided to reduce the number of features used for the clustering task. Features selection approaches such as forward features selection or regularisation methods are well fitted to supervised tasks but less so to unsupervised clustering. Therefore, general methods of dimensionality reduction were preferred. In practice, dimensionality reduction was performed before clustering and the dataset with reduced dimensionality used as input to the clustering algorithms.

[d] Only triggers happening within 30 days of the last operation were considered as being potential postoperative complications.

Principal component analysis (PCA)

PCA is a dimensionality reduction method based on the analysis of variance within the data. Informally, it defines a new orthogonal coordinate system through the data, whereby the newly-defined axes (the principal components or PCs) successively describe those directions through the data corresponding to maximum variance. The PCA will provide a number of principal components equal to the original dimension of the data, and where the resulting axes defined by the PCs is a rotation of the original axes. Each axis (i.e. PC) has a corresponding eigenvalue (of the data covariance matrix), with the first PC corresponding to the largest eigenvalue, and thus the largest contribution to the variance. Each subsequent PC corresponds to successively smaller eigenvalues, and therefore axes that have ever smaller contributions to the overall variance of the data.

Most of the variance should therefore be explained by the first few principal components, allowing to disregard the subsequent PCs with only a minimum loss of information. That is, the original data can be projected onto a subspace defined by the first k PCs, hence reducing the dimension to k. Choosing larger values of k corresponds to retaining ever more of the original dimensions of the PCA transformation, but where the assumption is that most of the variance (of relevance to the task at hand) can be captured by the top k PCs (i.e., those with the largest corresponding eigenvalues). More formally, the original data is then transformed into the new subspace using the formula $y = W' \times x$, where W' is the transpose of the selected eigenvector matrix of dimension $d \times k$ with d being the original dimension of the dataset and k the number of selected components and new dimension of the dataset.

This method works particularly well in datasets with strongly correlated variables and is a commonly-used tool for dimensionality reduction in data science, and in bioinformatics in particular. PCA dimensionality reduction also tends to preserve large pairwise distances between datapoints more than smaller pairwise distances. In this thesis, PCA dimensionality reduction was performed using PCA function from the Scikit-learn package in Python[317]. The number of principal components was chosen to explain a given proportion of the variance, which was considered as an optimisation parameter.

Autoencoder

An autoencoder is a deep learning tool built of two, most of the time symmetric, neural networks: the encoder and the decoder. It is an unsupervised learning approach designed for representation learning. In other words, it trains a first neural network to project the input data into a smaller dimension (the latent space) and then a second to reconstruct an output as close as possible from the initial input data. Weights of the models are iteratively updated to minimise a defined loss function, often a metric for the error of the reconstruction. Autoencoders offer the advantages of learning non-linear relationships between variables and are hence a good alternative to methods like PCA for dimensionality reduction. (Very formally, it can be shown that PCA is a special case of autoencoders under certain architectural constraints, though this is beyond the scope of this thesis. For the present work, autoencoders were considered to be a non-linear alternative to PCA.)

In this thesis, the autoencoder was programmed using the keras package in Python[318]. The network architecture can be seen in **Figure IV-2** and uses a fix number of neurons for the encoder and decoder parts, with a bottleneck at the latent space layer, needed for the dimensionality reduction function of the autoencoder. Several hyperparameters were tuned including the number of layers in both the encoder and decoder (2, 4, 16, 32), the number of neurons in each layer (64, 128, 256, 512) and the dimension of the latent space (in a first step: 4, 8, 16, 32; and in a second step: by increment of one fifth of the original input data dimension). Further details about the autoencoder parameters and optimisation can be found in **Suppl. Table IV-2a-c**. The best hyperparameters were determined by running a grid search on the three mentioned tuned parameters and using an 8-fold cross validation. The split between the training and validation sets was 7:1, bringing the overall training-validation-test split to 7:1:2.

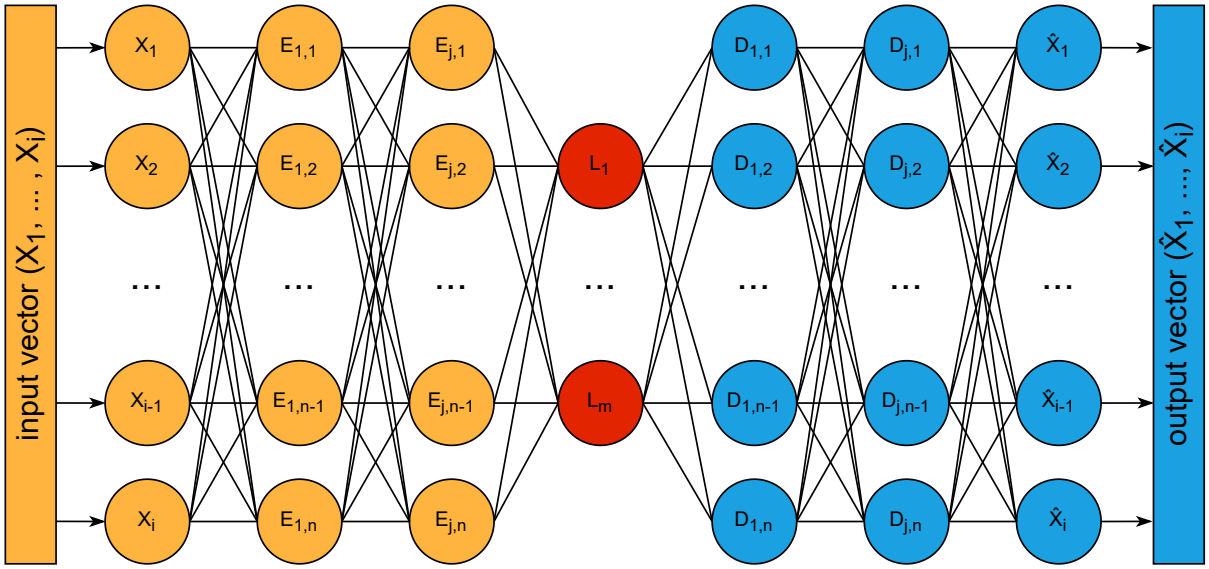


Figure IV-2: architecture of the autoencoder. On the left in orange is the encoder part, on the right in blue the decoder, whereas the red layer at the centre represents the latent dimension of the model. i = dimension of the input and output vectore, j = number of layer in the encoder and decoder, n = number of neurons in each layer, m = dimension of the latent space.

IV.2.4 Modelling

In order to build clusters of similar patients, or more exactly of similar presentations at the time of the first NEWS alarm, without having detailed labels of postoperative complications, two unsupervised clustering algorithms were tested.

k-means clustering

k-means clustering is an unsupervised learning approach aimed at building clusters of similar data points based on some selected distance metrics. The model first chooses k random centroids. Then it assigns every data point to the closest centroid. It subsequently takes the mean of all data points in each cluster, to define the new centroid positions. The whole process is repeated until cluster centroids reach stability or the maximal number of iterations set as hyperparameter is reached. k-means clustering is a relatively simple approach to implement, guarantying convergence and easily adapting to new examples.

In this thesis, k-means clustering was performed using the KMeans function from the Scikit-learn package in Python[317]. Euclidean distances were used to compute similarity and the maximum number of iterations was set at 100. As the position of the initial centroids can influence the final clustering, repeated initialisation (n=10) was performed and the best initial centroids in terms of final inertia selected.

Hierarchical clustering

Hierarchical clustering is an unsupervised learning approach generating a hierarchy in the grouping of datapoints, alike to the taxonomy trees it is mimicking. It can be divided into two types of algorithm: divisive and agglomerative. Divisive hierarchical clustering starts with one cluster comprising all datapoints and divides it into smaller ones, whereas agglomerative hierarchical clustering starts with as many clusters as there are samples and agglomerates them into larger clusters. The criteria for dividing or agglomerating clusters vary between models, but they are generally based on some distance-based metric, used to compute a similarity matrix between the clusters. Hierarchical clustering has the advantage of not making any prior assumption about the number of cluster and to be perfectly reproducible as no random initiation is needed.

In this thesis, agglomerative clustering was chosen, using the average Euclidean distance between the points of two clusters as metrics. Analyses were performed using the AgglomerativeClustering function of the Scikit-learn package in Python[317].

Cluster number optimisation

Clustering algorithms do not define an optimal number of clusters and this number has to be optimised by the developers based on their objective for the clustering task and using methods specific to each clustering algorithm. In this thesis, the optimal cluster number for the k-means algorithm was selected using the performance of a custom proportion score (see IV.2.5) on a validation set through 8-fold cross validation (the training set was divided 7:1 into a training set and a validation set for the cluster number optimisation). Other conventional approaches, such as the so-called elbow method and silhouette score, were also applied, but did not yield satisfactory results (see **Suppl. Figure IV-1**). For hierarchical clustering, the dendrogram method was used. A dendrogram is a tree-like diagram in which leaves represent individual samples and nodes groups of different sizes. In this case, the leaves were the contact points of interest and nodes the different clusters. The length between two nodes was proportional to the dissimilarity between two clusters (see **Figure IV-8**). The optimal number of clusters was determined by analysing the dissimilarity between clusters. In practice, the optimal number of clusters was given by the number of vertical lines (greater than n = 2) intersected by a horizontal line running through the tallest section of uninterrupted vertical lines on the dendrogram.

IV.2.5 Retrospective testing on the hold-out test set

To evaluate the model, its performance was tested on the hold-out test set. Using Euclidean distance to cluster centroids, each instance in the test set was matched to the closest cluster in the training set. The prescriptions received (within the next 24h) and procedures booked (on the next day) for each instance of the test set were then compared to the top ten prescriptions/procedures of the attributed cluster.

The performance of the model was tested using one standard and one custom score. The first is defined as being recall at position 10, representing the number of relevant recommendations

in the top 10 prescriptions of the cluster (i.e. prescriptions received/procedures booked for the instance of the test set also present in the cluster's top ten) divided by the number of prescriptions received/procedures booked for the instance of the test set.

The second was designed to account for the proportion of patients in the attributed cluster receiving each of the top ten prescriptions/procedures

$$\text{score proportion} = \frac{\sum_{i=0}^{n} P(i)}{n}$$

where i is a prescription/procedure in the list of new prescription/procedure received by an instance of the test set, n the number of new prescriptions/procedures in this list and P the proportion of patients in the corresponding cluster receiving the prescription/procedure i (see **Figure IV-1**). Any new prescription/procedure received by the instance in the test set but not included in the cluster's top 10 was given a probability of 0, in order to penalise heterogenous clusters.

Performance of the model was compared to a random selection of 10 prescriptions/procedures amongst all prescriptions/procedures received by any patient in the training set and to a selection of the 10 most common prescriptions/procedures within the training set.

IV.3 RESULTS

IV.3.1 Population and data description

The selected dataset included 191,382 observation sets, from 6,590 hospital admissions and 5,996 patients. Of these observation sets, 21,012 triggered a NEWS alarm (NEWS score of five or greater, or of three in a single component), from 2,667 hospital admissions and 2,496 patients. After sampling, 2,667 first alarm observation sets were included for clustering and performance analysis (**Figure IV-3**). 68 of the included observation sets were followed by a critical event within 24h (PPV = 0.03). The training set comprised 2130 observation sets, of which 40 were followed by a critical event (PPV = 0.02), and the test set 537 observation sets, of which 28 were followed by a critical event (PPV = 0.04). The number of prescriptions received within 24h (median = 1; range = 0 to 14) and procedures booked on the next day (median = 0; range = 0 to 5) were similar in the training and test sets. **Table IV-3** gives an overview of the population characteristics. The median age of the patients triggering at least one NEWS alarm and the median age of the patient experiencing at least one critical event were higher than the overall population median.

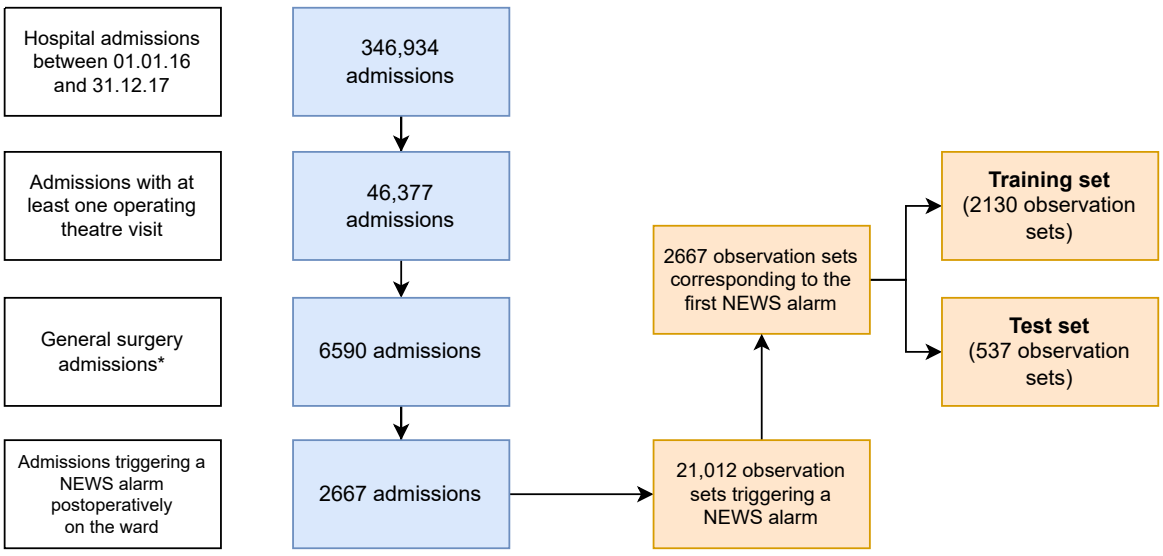


Figure IV-3: data selection flowchart. The numbers refer to the data available in the HAVEN database for the Oxford University Hospitals NHS Trust. Hospital admissions are preferred units as a same patient can have several admissions. Observation sets were considered as contact points and complemented with additional features as described under section IV.2.2. * defined according to the criteria stated under section IV.2.1.

Patients experiencing at least one critical event had a higher Charlson comorbidity index than the rest of the population. Not all critical events were preceded by a NEWS alarm and amongst those which did, not all where within 24h of the first alarm trigger. The PPV of the first NEWS alarm is lower than the overall NEWS alarm PPV.

	All admissions	Admissions with at least one event	Observation sets triggering a first NEWS alarm	Admissions with at least one event within 24h of a first NEWS alarm
Number of hospital admissions	6590	338	2667	68
Number of patients	5996	327	2496	68
Number of observation sets	191,382	40,740	2667	68
Number of observation sets triggering NEWS alarms	21,012	7083	2667	68
Number of observation sets followed by an event <24h	4052	4052	68	68
Positive predictive value (PPV)	0.07	0.20[a]	0.03	N/A
NEWS sensitivity	0.35	0.35[a]	N/A	N/A
Number of admissions with at least one event, including:	338	338	68	68
- cardiac arrest	0	0	0	0
- unplanned ICU admission	134	134	40	40
- death	11	11	1	1
- readmission to theatre	277	277	36	36
Median age (IQR)	56 (40-70)	62 (50-73)	62 (47-72)	65.5 (53.5-73)
Median CCI (IQR)	0 (0-4)	3 (0-8)	0 (0-5)	0 (0-8)

Table IV-3: characteristics of the study population, stratified by NEWS alarm status. CCI = Charlson Comorbidity Index; ICU = Intensive Care Unit; IQR = Interquartile Range; NEWS = National Early Warning Score; [a] theoretical value, not applicable in practice.

Figure IV-4 shows the distribution of NEWS values amongst observations triggering an alarm (values below five mean that one of the NEWS individual components received a score of three), the distribution of the number of alarm trigger per hospital admission, the distribution of all alarm triggers over time, and the distribution of the first alarm triggers over time. Of the total number of alarm triggers, the vast majority (score <=6, 16,902/21,012, 80%) are reviewed by ward doctors not necessarily part of a critical care team. Over half of the admissions had three or less alarms during their stay (1478/2667, 55%). Half of the alarm triggers (10,676/21,012, 51%), and the vast majority of the first alarm triggers (2371/2667, 89%) happened during the first 48 postoperative hours.

An analysis of the intervals between two observation sets or laboratory value recordings (see **Figure IV-5** showed that the periodicity of vital signs and laboratory results recording is stable over the length of the hospital admissions. It also showed that a threshold of 26 hours for the forward imputation of missing laboratory values was adequate and reflecting the clinical periodicity of data collection. An analysis of drug prescriptions frequency (**Figure IV-6**) showed that contact points triggering an alarm seem followed by an increase in the number of prescriptions, independently of the occurrence of a critical event. As expected, an increase in the number of prescriptions was also observed for contact points followed by a critical event.

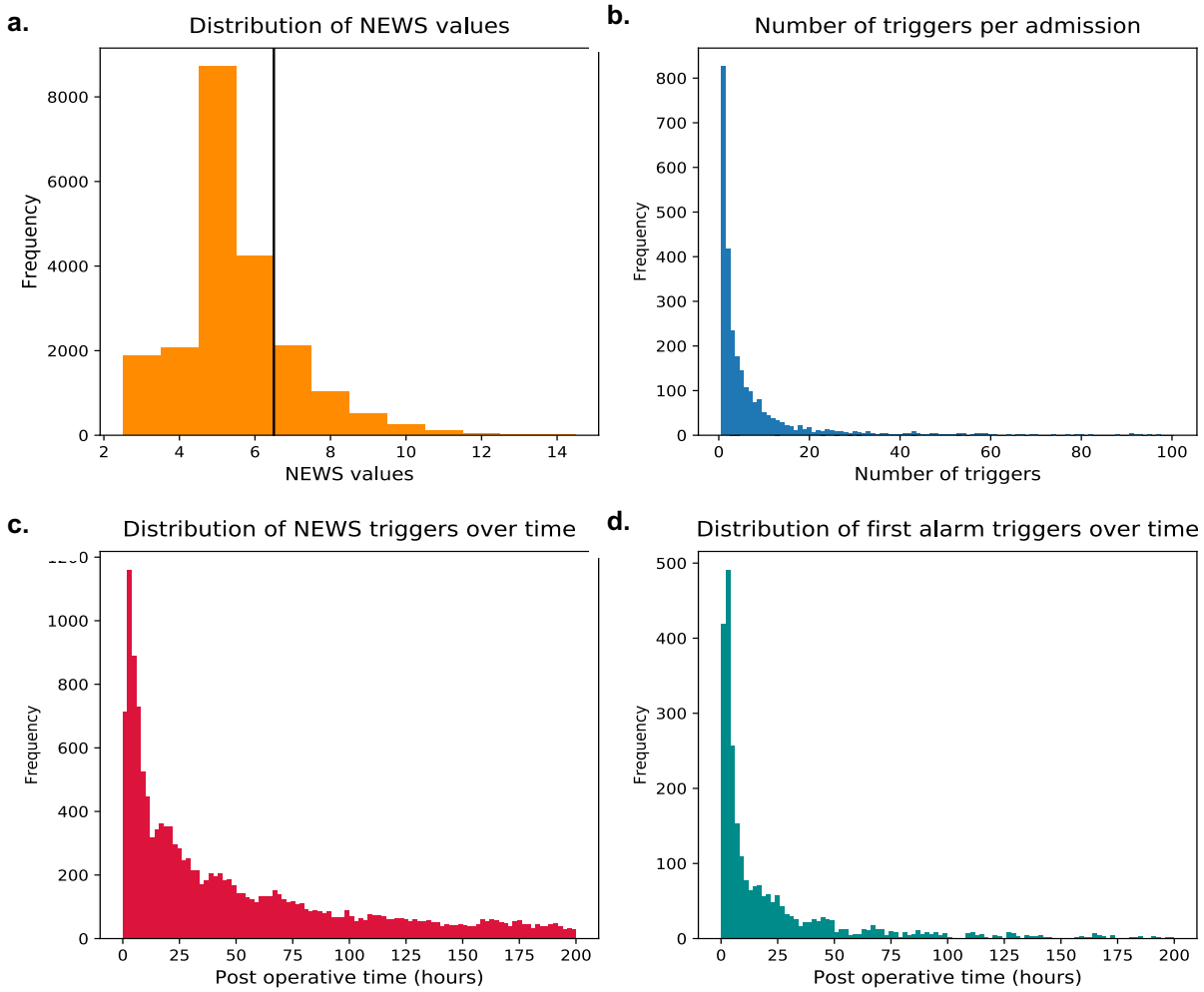


Figure IV-4: dataset characteristics. a. distribution of NEWS value amongst all observation sets triggering an alarm. All alarms with a score <=6 are mostly reviewed by junior ward doctors. Alarms >6 should in theory be reviewed by a critical care team. b. distribution of the number of triggers per hospital admission. One patient can be represented in several hospital admission. c. distribution of all NEWS alarms over the postoperative time. d. distribution of the first NEWS alarm per hospital admission over the postoperative time.

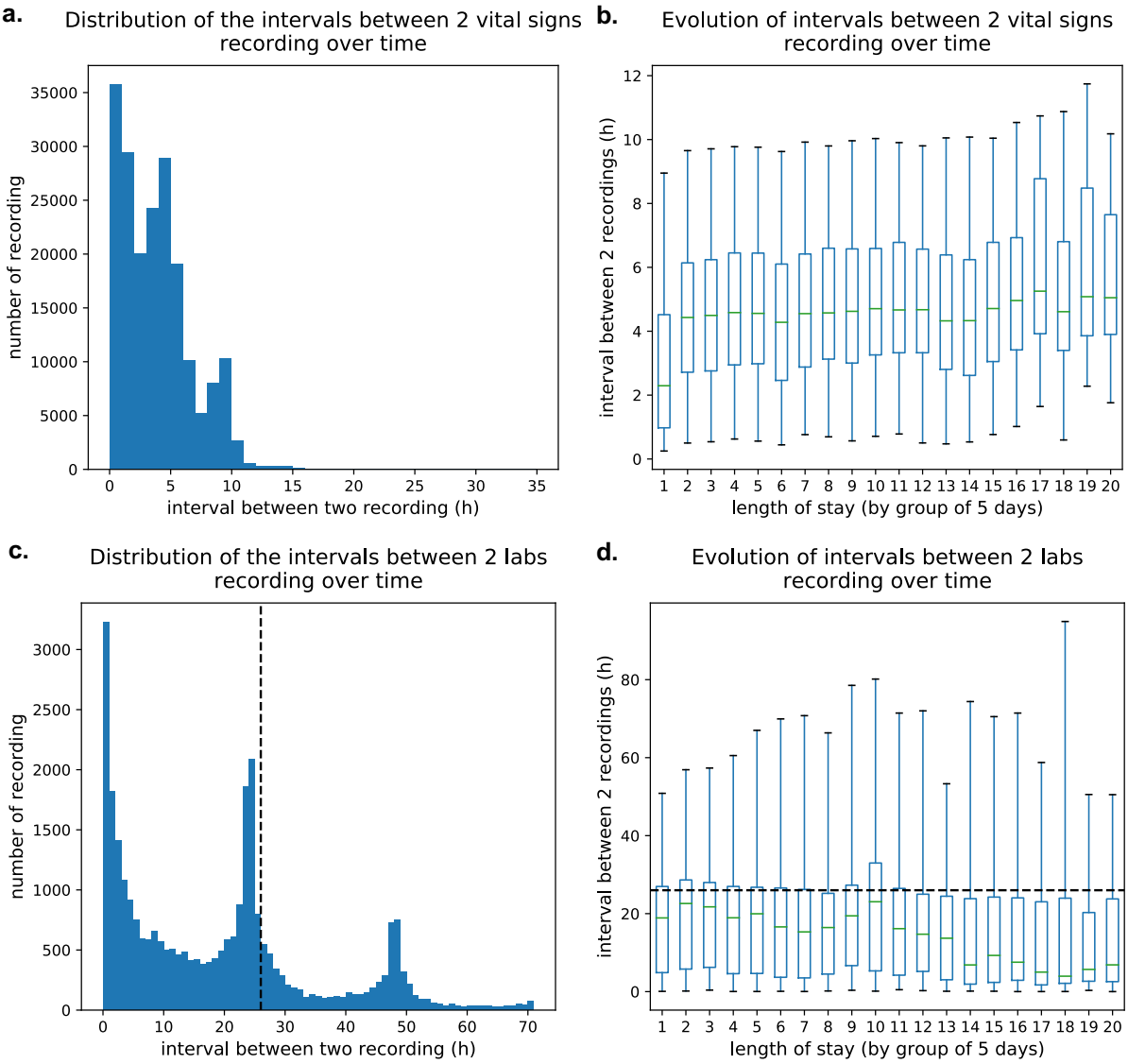


Figure IV-5: description of the intervals between the timepoints of input data collection.
(a.) frequency of the intervals for vital signs recording; (b.) distribution of the intervals over the time spent in hospital (grouped by 5-day periods) for vital signs; (c.) frequency of the intervals for laboratory values (any types); (d.) distribution of the intervals over the time spent in hospital (grouped by 5-day periods) for laboratory values (any types). The threshold for imputing a laboratory value by the value of its previous occurrence was 26 hours.

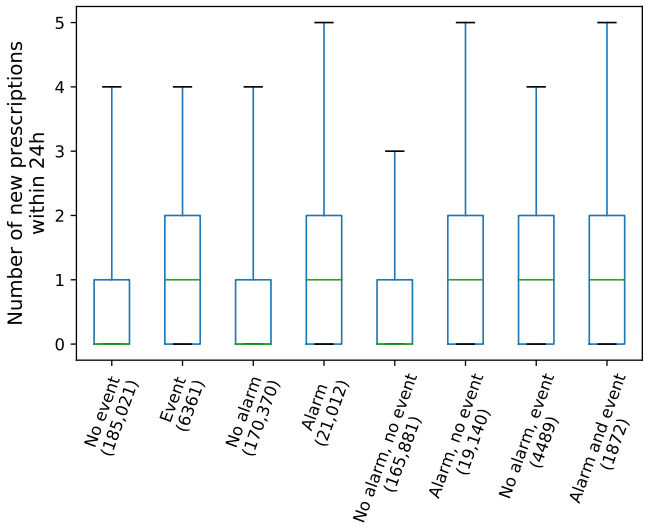


Figure IV-6: distribution of the number of new prescription within 24 hours according to the occurrence of NEWS alarms and critical events following a contact point. (number of contact points)

IV.3.2 Model optimisation

Figure IV-7 presents an example of the cluster number optimisation for the k-means model, with different combinations of dimensionality reduction algorithms (PCA and autoencoder) and dimensions of the latent space (given either by the number of selected PCA components or the size of the autoencoder's central layer, see section IV.2.3). The complete optimisation figures for all input feature sets are available in **Suppl. Figures IV-2a-c** and **Suppl. Figure IV-3a-c**. The number of principal components and of latent variables appear to have only a limited impact on performance. Balancing the results of the grid search optimisation, suggesting that a higher dimensional latent space yield better results (see **Suppl. Table IV-2a-c**), and the need to reduce dimensionality to avoid overfitting and distance concentration, the number of dimensions in the latent space of the autoencoder was fixed at just below 40% of the input feature set dimension. Based on similar reflexion, the number of components for the PCA transformation was set to explain 80% of the variance, echoing common practice in the field. The graph could overall be divided into two phases: a first phase of sharp, regular increase in score and a second phase of flatter, oscillating increase in score. The optimal cluster numbers were selected at the junction between these two phases and are summarised in **Table IV-4**.

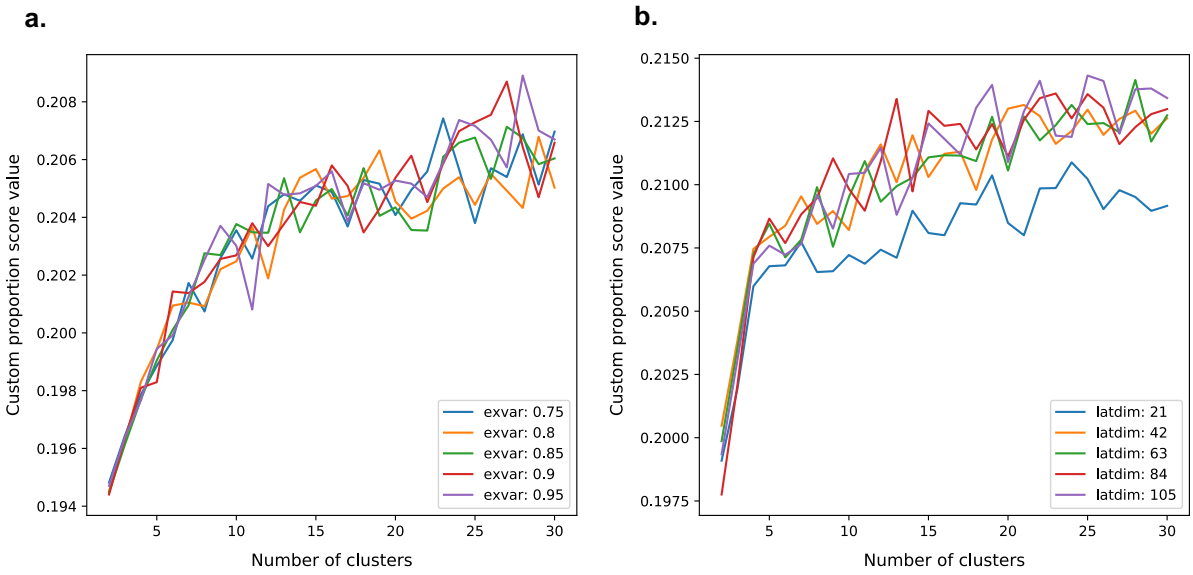


Figure IV-7: cluster number optimisation for the k-means algorithm using the custom proportion score. Input feature set = all variables. (a.) for a combination of PCA and k-means; (b.) for a combination of autoencoder and k-means; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

Figure IV-8 displays an example of hierarchical clustering model dendrograms after dimensionality reduction through PCA and autoencoder. As explained in the methods section, the optimal number of clusters is given by the number of vertical lines intersected by a horizontal line running through the tallest section of uninterrupted vertical lines (represented here in blue). **Table IV-4** summarises the optimal cluster numbers for the different algorithm combinations and input feature sets. The complete optimisation graphs for all input feature sets are available in **Suppl. Figure IV-4a-c**.

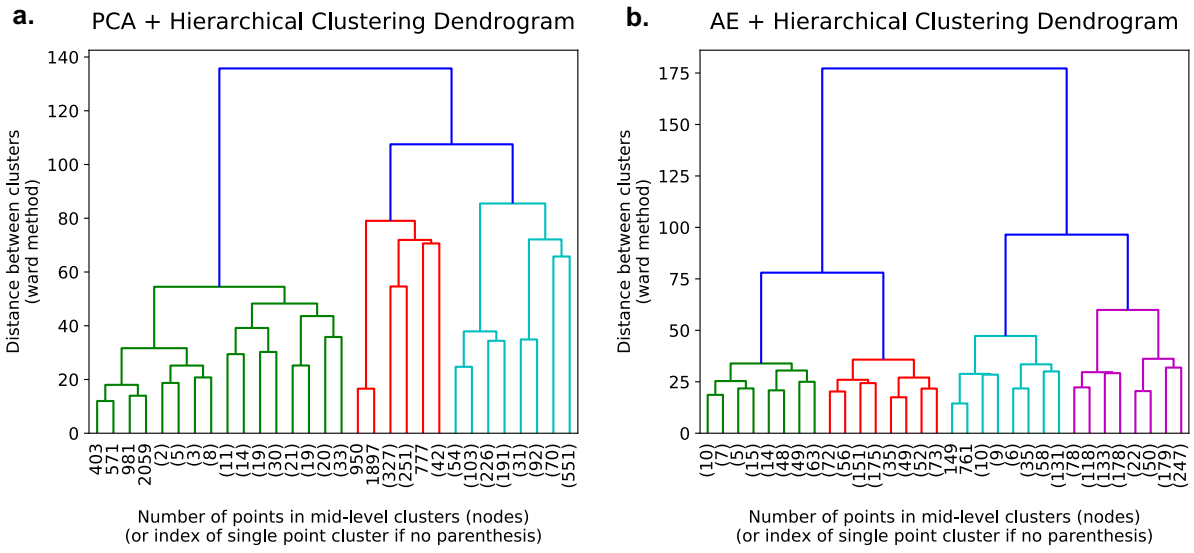


Figure IV-8: cluster number optimisation for the hierarchical clustering algorithm using the dendrogram method. Input feature set = vital signs + common laboratory variables. (a.) for a combination of PCA and hierarchical clustering; (b.) for a combination of autoencoder and hierarchical clustering. The colour coding represents the optimal number of cluster obtained through the dendrogram method.

a.

	k-means clustering	hierarchical clustering
PCA (exvar = 0.8, ncomp = 47)	6	3
autoencoder (latdim = 42)	4	4

b.

PCA (exvar = 0.8, ncomp = 35)	5	3
autoencoder (latdim = 30)	5	4

c.

PCA (exvar = 0.8, ncomp = 12)	9	4
autoencoder (latdim = 8)	7	3

Table IV-4: optimal number of clusters for each combination of dimensionality reduction and clustering algorithm. (a.) for the input set including all features; (b.) for the vital signs + common laboratory values + admin/demographic variables input set; (c.) vital signs only. PCA = principal component analysis; exvar = explained variance; latdim = dimension of the latent space; ncomp = number of components.

IV.3.3 Drug prescription recommendations

Out of the 2667 observation sets triggering a first NEWS alarm, 1760 (66.0%) were followed by the prescription of between 1 and 14 new drug groups (median 2, 894 different combinations) within 24 hours of the alarm. In comparison, out of the 111,781 observation sets from the same hospital admissions but not triggering a NEWS alarm, 47,290 (42.3%) were followed by the prescription of between 1 and 12 new drug groups (median 2). **Table IV-5** shows the 10 most prescribed drug groups and their occurrence within the training set.

Table IV-6 summarises the performance of different clustering pipelines for each set of input variables. For both recall at position ten and the proportion score (see section IV.2.5), the best performing model fed the vital signs + common laboratory variables feature set to a hierarchical clustering algorithm after dimensionality reduction through autoencoder and achieved an overall performance of 79.5% for recall at position ten and 20.9% for the proportion score. In comparison, a random list of ten prescription would yield a recall at position ten of 16.1% and a proportion score of 3.0%. A selection of the ten most prescribed drug groups (including no new drugs) would yield a recall at position ten of 78.3% and a proportion score of 18.7%.

Prescription	Prevalence	% hospital admissions in the training set receiving the prescription
No new drugs	723	33.9%
Volume >1l	545	25.6%
Pain (WHO level 2 & 3)	387	18.2%
Antacid	303	14.2%
Potassium	291	13.7%
Antibiotic	212	10.0%
Anticoagulant	195	9.2%
Magnesium	127	6.0%
COPD & asthma	127	6.0%
Antiemetic	123	5.8%

Table IV-5: top 10 drug prescription categories in the training set. COPD = chronic obstructive pulmonary disease; WHO = World Health Organization

a.

	All (n=106)*	vitals + common labs§ (n=78)*	vitals (n=24)*
random list	16.1%	16.1%	16.1%
top 10	78.3%	78.3%	78.3%
PCA + k-means	78.7%	78.8%	77.9%
AE + k-means	78.3%	78.4%	79.2%
PCA + hierarchical clustering	78.3%	79.5%	78.8%
AE + hierarchical clustering	79.5%	79.5%	78.7%

b.

random list	3.0%	3.0%	3.0%
top 10	18.7%	18.7%	18.7%
PCA + k-means	20.2%	19.9%	19.7%
AE + k-means	20.4%	20.2%	19.7%
PCA + hierarchical clustering	19.0%	19.6%	19.1%
AE + hierarchical clustering	19.8%	20.9%	19.1%

Table IV-6: results summary for the primary objective for (a.) recall at position 10 and (b.) score proportion. *including additional input features min-max and delta, as described in section IV.2.2; §only variables with less than 2/3 of missing values. PCA = principal component analysis; AE = autoencoder.

Table IV-7 presents the size, top ten prescription groups and proportion of patients experiencing a critical event within 24 hours of the alarm for each of the four clusters built by the hierarchical clustering algorithm after dimensionality reduction with the autoencoder (best performing combination for task 1). Clusters 1 and 3 seem to be higher risk clusters with the proportion of admission developing a critical event within 24h approximately doubled compared to Clusters 2 and 4. Cluster 1 has a higher rate of antibiotics prescription and could reflect infectious complications. Cluster 4 has the lowest rate of new prescription (the most common prescription being fluid) while being a low risk cluster, which could reflect a cluster of patients with favorable post-operative evolution and few complains, triggering false positive alarms. Clusters 2 and 3 have the highest rates of new prescriptions with 75.9% and 86.6% of the patients receiving at least one, respectively. The prescription rates of COPD medication and antacids could indicate clusters of patients with more risk factors for complications such as smoking or obesity. The overall top ten prescription groups (see **Table IV-5**) represent the vast majority of prescribed drugs in all clusters.

Prescription	% hospital admissions receiving the prescription	Prescription	% hospital admissions receiving the prescription
Cluster 1 (size=398; risk=4.0%)		**Cluster 2 (size=700; risk=1.6%)**	
No new drug	33.9%	Volume	26.1%
Volume	27.4%	Pain (WHO level 2 & 3)	25.9%
Potassium	19.1%	No new drugs	24.1%
Antibiotics	17.6%	Antacids	20.1%
Pain (WHO level 2 & 3)	14.8%	Potassium	10.9%
Antacids	9.5%	Anticoagulants	10.6%
Antiemetics	6.0%	COPD	9.4%
Magnesium	6.0%	Anaemia	8.6%
Anticoagulant	5.3%	Pain (WHO level 1)	7.4%
Antipsychotic/COPD /transfusion	4.5%	Antibiotics	7.1%
Cluster 3 (size=197; risk=3.0%)		**Cluster 4 (size=835; risk=1.8%)**	
Volume	33.5%	No new drugs	47.1%
Magnesium	32.0%	Volume	22.4%
Antacid	27.0%	Pain (WHO level 2 & 3)	13.5%
Anticoagulant	22.3%	Potassium	12.7%
Pain (WHO level 2 & 3)	17.3%	Antibiotics	9.1%
Potassium	16.8%	Antacids	8.5%
No new drugs	13.2%	Anticoagulants	6.7%
COPD	11.2%	Pain (WHO level 1)	4.2%
Albumin	9.1%	Antiemetics	4.2%
Anaemia	8.6%	Anaemia	3.1%

Table IV-7: top 10 prescription groups for each of the four clusters built by the best performing algorithm combination (autoencoder + hierarchical clustering with vital signs and common laboratory values as input features). The risk indicates the percentage of patients in each cluster experiencing a critical event within 24h of the first alarm. COPD = chronic obstructive pulmonary disease; WHO = World Health Organization.

IV.3.4 Procedure/imaging exam recommendations

Out of the 2667 observation sets triggering a first NEWS alarm, 161 (6.0%) were followed by the booking of between 1 and 5 new procedures (median 1, 63 different combinations) on the next day (the metadata of procedures recording only contained the date of the procedure, see section IV.2.1). In comparison, out of the 111,781 observation sets from the same hospital admissions but not triggering a NEWS alarm, 5657 (5.1%) were followed by the booking of between 1 and 5 new procedures (median 1).

Table IV-8 shows the 10 most booked procedures and their occurrence within the training set. **Table IV-9** summarises the performance of the different clustering pipelines for each set of input variables. For recall at position ten, the best performing model fed the vital signs + common labs feature set to a k-means clustering model after dimensionality reduction through

an autoencoder and achieve an overall performance of 98.0%. For the proportion score, the best performing model fed the feature set containing all the variables to a hierarchical clustering model after dimensionality reduction through PCA and achieve an overall performance 89.2%. In comparison, a random list of ten procedures would yield a recall at position ten of 17.5% and a proportion score of 15.1%. A selection of the ten most booked procedures (including no procedure) would yield a recall at position ten of 97.7% and a proportion score of 87.9%.

Procedure	Prevalence	% hospital admissions in the training set receiving the prescription
No procedure	2005	94.1
CT scan	51	2.4
Urinary catheter	14	0.7
CT PE	10	0.5
Urinary catheter removal	9	0.4
Approach to organ under image control	9	0.4
Laparotomy	4	0.4
VAC dressing	4	0.2
(9^{th} and 10^{th} position empty – ex equo: intravenous nutrition, echocardiography, gastroscopy, irrigation of organ, thoracic drainage)		

Table IV-8: top 10 procedure prescriptions categories in the training set. CT = computer tomography; PE = pulmonary embolism; VAC = vacuum-assisted closure.

a.

	All (n=106)*	vitals + common labs§ (n=78)*	vitals (n=24)*
random list	17.5%	17.5%	17.5%
top 10	97.7%	97.7%	97.7%
PCA + k-means	97.8%	97.8%	97.3%
AE + k-means	97.8%	98.0%	97.3%
PCA + hierarchical clustering	97.7%	97.8%	97.7%
AE + hierarchical clustering	97.9%	97.7%	97.4%

b.

random list	15.1%	15.1%	15.1%
top 10	87.9%	87.9%	87.9%
PCA + k-means	87.9%	88.0%	88.0%
AE + k-means	87.9%	88.0%	88.0%
PCA + hierarchical clustering	89.2%	88.4%	88.6%
AE + hierarchical clustering	88.3%	88.4%	88.0%

Table IV-9: results summary for the secondary objective for (a.) recall at position 10 and (b.) score proportion. *including additional input features min-max and delta, as described in section IV.2.2; §only variables with less than 2/3 of missing values. PCA = principal component analysis; AE = autoencoder.

IV.4 DISCUSSION

In this chapter, the proof of concept of a ML model based on patient similarity to improve the management of patients triggering a NEWS alarm postoperatively was tested and presented. Several combinations of input feature sets, dimensionality reduction models and clustering algorithms were evaluated. The different models matched test patients to pre-established clusters and provided a list of the ten most commonly prescribed drug groups or booked procedures in the corresponding clusters. All the models tested performed markedly better than a random selection of prescription or procedures. However, when using recall at position ten (a common performance metrics for information retrieval algorithm, see section IV.2.5), only a few of the tested model combinations outperformed the recall at position ten of a recommendation made of the ten most common prescription/procedure overall in the training set, and only did so by a marginal increment (see **Table IV-6** and **Table IV-9**). This is likely due to the high prevalence of the most common prescription/procedure groups (see **Table IV-5** and **Table IV-8**) in all clusters. Nonetheless, this metric does not account for the position of the recommendations in the list, nor the proportion of the patients in the cluster receiving the recommended prescriptions/procedures, even though these two pieces of information can play an important role in influencing the user decision to follow or not a recommendation.

This is the reason why a custom score accounting for the proportion of patients in the cluster receiving each of the prescriptions/procedures was developed. This score, although not commonly used, gives a better estimate of the model recommendations' usefulness to a clinician by compiling, for every test patient, the mean of the proportions of patients in the attributed cluster receiving each of the prescriptions/procedures also received by the test patient. The best achieved performance was 20.9% using hierarchical clustering after dimensionality reduction with an autoencoder. This performance was 2.2 percentage point higher than the performance of recommendations based on ten most commonly prescribed drug groups overall (representing a 12% relative increase). Nonetheless, this result remains

far from the 100% an ideal system would display by forming homogenous cluster of patients with the same condition and management (assuming all patients with the same condition are treated similarly), is of unknown robustness to changes in the composition of the patient population and remains of unclear clinical significance. A better score would have been to calculate the proportion of the patients in the cluster receiving the exact same combination of drugs that the test patient, although this was not practically implementable given the high number of drug group combinations (n=894) compared to the number of samples receiving a prescription (n=1760).

A comparison of these results to existing literature is made difficult by the scarcity of studies using similar approaches to prescription recommendation, let alone in the field of surgery. Wang *et al.* developed a recommendation tool integrated in the EHR of the study site and displaying the most commonly prescribed physician orders for a group of patients similar to the new target patient, in order to speed up the electronic prescription process. Tests on a set of 12,818 patients across medical specialties yielded an average precision of 0.88 and 0.80 at position 5 and 10 respectively.[319] However, these results are not directly comparable as the authors discarded target patients with less than the average number of prescription. Because this was not done in the present study, precision at position could not be appropriately evaluated and recall at position was measured instead. Panahiazar *et al.* used k-means, hierarchical clustering and a supervised clustering algorithm to produce heart failure therapy recommendations and measured performance by calculating the area under the receiver operating characteristic (AUROC) curve. The best performance was achieved by a supervised learning model (AUROC = 0.74), followed by hierarchical clustering (0.71) and k-mean (0.69)[320].

When analysing the composition of the clusters formed by the best performing model for drug prescriptions on the proportion score, two relatively large, yet low risk, clusters are formed alongside two smaller clusters, yet with a higher proportion of patients experiencing critical events. Infectious complications seem to feature more prominently in one of the clusters,

whereas two other clusters appear to englobe patient with more chronic comorbidities such as COPD. It can also be observed that, whereas changing in order and frequency, the top ten prescription lists of each cluster remain mainly composed of the most frequent prescriptions overall, including no prescription at all. This is not surprising given the small number of clusters selected but impede clinical usefulness and a better evaluation of the system performance. This observation is even more relevant for the secondary objective, with 94% of the contact points of interest having no procedure prescribed on the next day.

There also seems to be a discrepancy between the optimal number of clusters from a mathematical and clinical perspective. Indeed, most surgeons would agree that surgical complications occur in more than four types of presentation, while all the mathematical cluster optimisation techniques used most of the time favoured a cluster number between two and six. This can be due to the absence of important, yet not collected, variables in the dataset, such as clinical signs (e.g. abdominal guarding) or symptoms (e.g. localisation of pain), allowing clinician to further refine their thinking but invisible to most algorithms so far. It could also come from mathematical limitation of the algorithms, such as the tendency of k-means to build spherical clusters or the impossibility for the hierarchical clustering algorithm to change lower-level cluster attribution once performed. This observation casts doubts on the clinical relevance of the built clusters, or at least on the adequacy of their granularity.

Therefore, improvements along two main axes remain needed. First, refining the accuracy and clinical meaningfulness of cluster generation. With more data, models with additional parameters, such as Gaussian Mixture Models, could be applied, allowing for more complex cluster shapes. Integrating mathematical penalties based on the clinical outcomes of interest into the clustering models could also be explored in order to increase the clinical appropriateness of the formed clusters. Another way of improving the recommendation relevance would be to filter out the contact points receiving no new prescriptions/procedures before the clustering and matching phases. Although diverting from the original rationale of having a unique and understandable training process, as well as requiring more data to

robustly train two sequential models, this approach could potentially offer more granular outputs to the users. Second, further work is warranted to improve the quality and completeness of the data used both as inputs and as outcomes of interest. More granularity in the chronology of procedure prescription, for example, would allow to capture immediate action taken on the same day, to which the present model was blind. Extracting and organising non-numerical data, such as clinical signs and symptoms, would probably unveil new clustering possibilities as well. It is a common saying in medical schools that 80% of a diagnosis is made by listening to the patients. If this estimate is true, even the best models currently learn on less than a quarter of the relevant data potentially available.

Once the performance improved, the presented system could offer advanced support to junior clinicians attending surgical complications on the ward, by presenting them with a tailored list of drug and procedure to consider in any given situation. This would act as much as an intelligent reminder system as a window into a patient possible clinical evolution based on the recorded trajectories of past patients. If deployed more broadly, such system (given it was trained on data from well-performing centres) could also help narrowing the performance spectrum across physicians by bringing more clinicians at the performance level of reference centres.

Finally, it should be noted that, even when achieved, good model performance on retrospective data does not necessarily equal clinical usefulness. In order to appraise the actual potential of the developed model, or more likely its future iteration, to improve patient care, its performance would still need to be evaluated in the clinical context of its intended use. Such investigations are necessary given the variability and many biases introduced by human users and the use of a system within existing clinical workflows (see Chapter V and VI). Evaluating the system in clinical scenarios and then in shadow mode during actual patient care was the original plan for the remaining chapters of the thesis, but had to be reconsidered given the model mixed performance results and restricted access to medical personal as well as clinical setting during the pandemic.

IV.4.1 Strengths and limitations

By using unsupervised clustering, the presented approach offers a scalable and easily adaptable method to generate clinical recommendations. Once the clustering performed, it is indeed relatively simple to match new cases to generated clusters and to interrogate these clusters on various type of information. This work focuses on drugs and procedures prescription within 24 hours, but the same model could be used to inform decision making about other clinical aspects such as long-time therapy choices or risk of readmission to hospital for example. By keeping humans in the loop and presenting a weighted spectrum of possible options, the described method also has the advantage of being potentially useful in practice even if based on incomplete representation of a clinical situation and not perfectly accurate. This could allow earlier implementation as for systems requiring high level of accuracy to be acceptable from a performance and safety perspective.

Despite the lack of important clinical information such as clinical signs and symptoms as well as the lack of granularity in the description of procedures, the model was trained on one of the most detailed databases available for ward patients in the United Kingdom, using eight vital signs, 37 laboratory tests, four further demographic or administrative features, as well as the complete list of drug and procedure prescriptions received by patients during their hospital stay. When appropriately handled, this depth of data allows to consciously explore, or indirectly benefit from, yet unknown connexion between clinical features.

The mathematical approach of the presented model was selected to reflect the way clinicians train themselves, namely by building clinical experience through the accumulation of cases and linking new patients to those previously seen. The outcomes of interest selected also reflect the Clavien-Dindo scale, a commonly accepted classification of surgical complications based on the therapeutic actions needed to address the undesired changes in a patient postoperative evolution[13,14]. These two points would hopefully improve the clinical acceptance of the system amongst practitioners during later phases of implementation.

Nonetheless, the results presented should be interpreted in the context of several limitations. First, despite the originally large number of records in the HAVEN dataset, the model was trained on a much smaller sample due to inclusion criteria. Even if encompassing over 2000 contact points of interest, the size of the training set must be appraised in relation to the number of input features (114 for the full input feature set). A disbalance between the number of features and the number of samples can lead to model overfitting and was partially addressed by the use of dimensionality reduction algorithm prior to clustering. Plans were made to extract more data from the trust data warehouse, extending the period covered in the dataset by at least 2 years and probably doubling the number of included samples, but were later abandoned due to the outbreak of SARS-CoV-2 in early 2020.

Second, the data contained in HAVEN were constrained by the way they were originally collected from the hospital systems and later extracted from the trust data warehouse. Therefore, some important metadata were sometimes missing. The most limiting example was the timing of booked procedures. The time stamp of these variables only recorded the date of the procedure, but not the exact hour and minute. In practice, this prevented the use in the model of procedures performed on the same day as the NEWS alarm trigger, because it could not be differentiated if the procedure took place before or after the alarm. This limitation probably prevented the use as outcome of many procedures booked in immediate reaction to a complication, such as urgent imaging. Here again, plans were made to update the metadata extraction methods in order to obtain more granular information about procedure timing, but were later abandoned due to the outbreak of the pandemic.

Third, the ML models used (k-means and hierarchical clustering) do not accept missing data. It was therefore necessary to input all missing data for the input features, which in many instances represented over 50% of the data. Despite attention given to develop an imputation strategy as close to the clinical reality as possible and to select input feature sets based on their missingness, it is likely that some of the imputed values did not faithfully represent the related patient's clinical status. Data missingness still represents a major challenge for the

exploitation of EHR data by ML algorithms and the transparent appraisal of their results[321]. Because it included laboratory variables, which are less regularly collected than vital sign observations, this study is no exception, although efforts were made to use imputation methods aligned to the clinical reality and transparently reporting them.

Fourth, when developing ML models, a certain number of hyperparameters need to be define and not all of them can be systematically optimised. Several design choices were made during model development, such as the type of distance used to calculate similarity, the number of splits used in the k-fold cross validation, or the epochs used to train the autoencoder. These reflect informal exploration carried out during the coding phase or rational choices based on underlying theory, but not necessarily the best performing set of parameters.

Fifth, in absence of low-level diagnostic labels and due to the impossibility to link back hospital admissions in the HAVEN database to the full patient records, the clinical meaningfulness of the clusters remains to be more thoroughly investigated. The clustering approach was chosen to present clinicians with an interpretable mathematical model and demonstrating that clusters reflect known patterns of complications would reinforce its face value. On the other hand, one of the hopes of using ML for personalised medicine is that models will one day be able to recognise patterns and similarities that the human eyes don't, making the quest for model interpretability and strict parallel to current clinical practice questionable[94]. Such practice would need models with high accuracy to become acceptable, though, which was still not achieve in the present work.

Sixth and finally, no external validation was performed. Patient population and clinician practice vary between centres. An external validation would be necessary to move beyond the proof-of-concept stage and offer evidence of model generalisability. Here again, plans were made to access data from a partner institution, but later abandoned due to the poor prospects of testing the model in clinical settings due to the outbreak of the pandemic and the reorientation of the thesis objectives.

IV.4.2 Conclusion

To conclude, I developed and tested on an internal hold-out dataset an unsupervised clustering algorithm recommending prescriptions and procedures for patients triggering a first NEWS alarm during the postoperative period on the ward according to their similarity with previous patients in the same situation. Despite several limitations and mixed performance results, the presented work is a proof of concept that a ML model based on patient similarity using unsupervised clustering could generate more tailored recommendations regarding the clinical management needed by patients triggering a NEWS alarm postoperatively. Such recommendations could provide advance support to junior clinicians facing postoperative complications on the ward and bring best standard practice of major centres to smaller hospitals. However, further research is warranted to improve the performance of the clustering model and validate the clinical relevance of the clusters built. More importantly, thorough evaluation in simulated clinical scenarios and then in live clinical settings will be needed before claiming any added value of the approach.

CHAPTER V

DECIDE-AI: a Delphi process

This chapter is adapted from:

Vasey, B. et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat. Med. (2021)[105].

-

Vasey B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. (2022)[322].

-

Vasey B. et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ (2022)[323].

V.1 Introduction

The algorithm presented in the previous chapter, even if overcoming the discussed limitations, would have been by design of limited clinical utility on its own. Like many other clinical AI applications, it would have deployed its full potential only when used in adjunction to human intelligence. The next logical step would therefore have been to assess its performance when used by human surgeons in simulated scenarios. This option being unrealistic (see section I.1), the focus of the thesis was redirected toward determining what ought to have been assessed, if scenarios, or more accurately early clinical trials, would have been possible. Indeed, despite an exponential growth in the number of AI algorithms published in the medical literature in the recent years, little guidance exists on how AI systems should be evaluated when first implemented, at small-scale, in live clinical settings. This apparent neglect of early-stage evaluation, coupled with an overemphasis on the technical aspects of the proposed algorithms (in contrast to the factors affecting the interaction with their human users), is in several authors' opinion[105] one of the likely explanations for the so-called AI chasm. The AI chasm was defined by Keane *et al.* as the gap between the number of algorithms developed and showing promising performance in simulation and the number of AI systems actually delivering clinical impact in real life [104].

As clinicians occupy the central role in patient care, and will likely keep occupying it in the near future, it is essential to focus the development and evaluation of clinical AI system on their potential to augment rather than replace human intelligence. Integrated within a traditional medical decision-making process, AI-based decision support systems pose unique challenges, such as their frequent lack of interpretability (the so-called "black box" problem) or their tendency to sometimes produce unexpected results. To account for these specificities, the field needs an approach which sets humans at the centre of the design and evaluation process, if it wishes to bridge *in silico* (i.e. through computer simulation) algorithm development and evaluation to commonly used bedside application. Four main aspects justify paying special attention to the early stage of clinical evaluation in particular.

First, a need to replicate the performances obtained in silico during real life clinical evaluation at small-scale, before proceeding to more comprehensive and summative evaluation. Human decision-making processes are complex and subject to many biases. It cannot, and maybe shall not, be expected that human users will exactly follow all of the algorithm recommendations. This applies even to directive models (i.e. prescribing a specific course of action), and is especially relevant in cases, probably the majority, where users remain accountable for their decisions[114]. In order to accurately evaluate an algorithm's impact and avoid the research waste of conducting expensive large-scale trials with decision support systems unlikely to improve user performance, it is essential to assess the actual impact of an algorithm on its users' decisions at an early stage. Additionally, consideration should be given to the difference between the development and target patient population, ensuring the algorithm's relevance in the implementation settings. Therefore, the assisted human performance in the target clinical environment, and not merely the algorithm's stand-alone outputs, need to be reported as outcomes.

Second, researcher duty to ensure patient safety. Because it cannot be assumed that users' decisions will mirror the algorithm's recommendations, safety data from *in silico* studies cannot simply be extrapolated to use in real life conditions. It is crucially important to test the safety profile of new algorithms when used to influence human decisions. Skipping this step and moving directly forward to large scale-trials would expose a considerable number of patients to an unknown risk of harm. This is both ethically unacceptable and potentially detrimental for future research on this technology. Suboptimal safety standards led to disastrous consequences in the early days of pharmacological trials; it is imperative we do not as a field repeat the same mistakes with clinical AI.

Third, the importance of human factors (ergonomics) evaluation and the benefits of iterative evaluation-design cycles. Human factors evaluations are commonly conducted in other safety critical fields such as aviation[324], the military[325] or energy sectors[326] and are already part of the regulatory process for medical devices[129,130]. They assess the impact of a device or procedure

on their users' physical and cognitive performance. Beyond the regulatory focus on safety, human factors are now receiving growing attention in the context of clinical AI, for their importance in system design and potential to improve system integration and clinical utility[134–136,327]. Technical requirements often evolve as a system starts being used, and users' expectations of a system also vary in the initial exposure period. For example, users might ask for an additional feature to contextualise the algorithm recommendations or better integration within existing clinical workflows[328], which in turn will require developers to adapt the display interface or compatibility with other systems. From a practical and economic viewpoint, the sooner human factors evaluation occurs, the more (cost) effective it is likely to be. By contrast, iterative design modification is difficult and inappropriate during large-scale trials, causing a serious risk of invalidating any conclusions, as the intervention tested would have been modified during trial.

Fourth, the stepping stone role of early-stage evaluation. Large-scale clinical trials are complex and expensive endeavours requiring careful preparation. A well-thought-out design is essential to produce valid and meaningful conclusions and needs preliminary information about the intervention under evaluation, such as, for example, the most appropriate outcomes for the trial, the expected effect size, the optimal participant inclusion and exclusion criteria, or the evolution of the users' trust in the algorithm over time. Not all such background information can be inferred from *in silico* evaluation and some data must be collected in small-scale prospective studies in order to be available to investigators at the time of drafting trial protocols.

As an increasing number of AI-based decision support systems move from development to implementation, better guidance on the reporting of early-stage clinical evaluation, including human factors, safety, preliminary clinical utility data, and preparatory steps for larger trials, is needed. Clear and transparent reporting on these aspects could not only avoid preventable harm to patients and research waste, but also play a key role in transforming AI from a promising technology to an evidence-based component of modern medicine. Guidelines

already exist (SPIRIT/CONSORT-AI), or are under development (TRIPOD-AI, STARD-AI), for the reporting of *in silico* development/validation of AI systems and their evaluation in large comparative studies[167,168,281,282]. but none focuses on the crucial stage of clinical AI development and evaluation mentioned in the previous paragraphs. In order to develop an informed and acceptable guideline, an international consensus-based process was initiated to determine the key information items that should be reported during the Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence (DECIDE-AI). For each information item, a justification of its importance and elaboration on its application (so called Explanation and Elaboration, E&E) were also developed. To improve the resulting guideline and increase its chances of adoption, a qualitative evaluation of the final items and E&E sections was conducted and the final documents modified according to its findings. The scope of DECIDE-AI are studies reporting on the early-stage, live, clinical evaluation of AI-based decision support system at small-scale. This scope can be compared to these of phase I/II trial for drug development or IDEAL stage IIa/IIb for surgical innovation (see **Figure V-1**)[186,187,329].

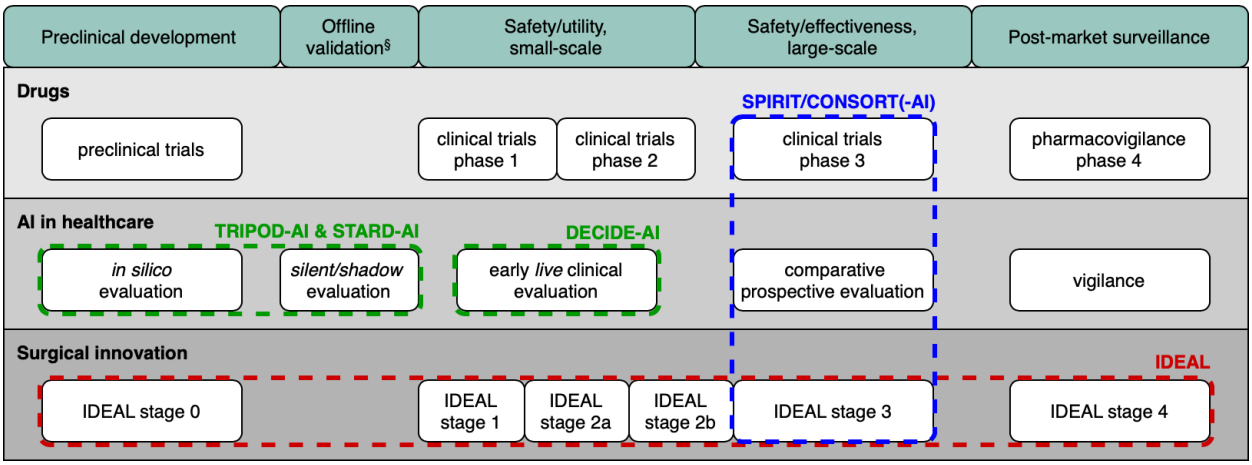


Figure V-1: comparison of drugs, AI in healthcare and surgical innovation development pathways. Adapted from *Vasey B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. (2022)*[322]. The dotted lines represent reporting guidelines, some of which are study design specific (TRIPOD-AI, STARD-AI, SPIRIT/CONSORT-AI), others stage-specific (DECIDE-AI, IDEAL). § the offline validation stage only apply to AI in healthcare.

V.2 Methods

For the present thesis, I led the development of the DECIDE-AI guidelines through an international, expert consensus-based process and in accordance with the EQUATOR network's recommendations for guidelines development[330,331]. A Steering Group composed of 12 invited members from various areas of expertise, was convened to oversee the guideline development process. Steering Group members were selected to obtain expertise in most of the necessary areas of research relevant to the guideline (e.g. machine learning, human factors, clinical trial methodology, reporting guideline development, entrepreneurship, etc.) as well as to have representation from the lead authors of other key existing or upcoming AI reporting guidelines (CONSORT-AI, SPIRIT-AI, TRIPOD-AI, STARD-AI)[167–169,171]. An effort to increase geographical representation was also made in a second step in order to increase the Steering Group variety of opinions and to diversify expert recruitment in later stages. The composition of the Steering Group can be found in Annex V-1. The project was reviewed by the University of Oxford Central University Research Ethics Committee (approval number R73712/RE003) and registered with the EQUATOR Network.

V.2.1 Delphi process

A modified Delphi process was conducted, with two rounds of participants feedback and one virtual consensus meeting. Delphi processes are sequential rounds of questionnaires (during which expert participants share their opinion on a set of questions or items), separated by centralised analysis of these answers and controlled feedback to the participants[332,333]. Proposed for the first time in 1963 by Dalkey *et al.* as a means to agree on which strategic industrial sites were most likely to be targeted by Soviet strikes[334], Delphi processes have since become popular tools to reach experts consensus in a methodological way. A modified Delphi approach was also selected by most reporting guideline development projects relevant to DECIDE-AI (i.e. IDEAL, TRIPOD-AI, STARD-AI, SPIRIT-AI, CONSORT-AI)[167,168,192,281,282].

The DECIDE-AI Delphi comprised two rounds of participants feedback (described in this chapter) and one final consensus meeting (presented in Chapter VI).

Participant recruitment

The invitation strategy aimed to ensure a broad range of stakeholder expertise and appropriate geographical representation. Experts were internationally recruited through four main channels:

1. invitation to experts recommended by the Steering Group. Twenty target stakeholder groups were defined before starting the recruitment process (administrators/hospital management, allied health professionals, clinicians, engineers/computer scientists, entrepreneurs, epidemiologists, ethicists, funders, human factors specialists, implementation scientists, journal editors, methodologists, patient representatives, payers/commissioners, policy makers/official institution representatives, private sector representatives, psychologists, regulators, statisticians, and trialists). When drafting the invitation list, it was aimed to include at least five experts from each of these groups.
2. invitation to authors of publications identified through the initial literature searches,
3. call to contribute published in a commentary article about the project in a non-UK medical journal with international readership and global reach[105], and consideration of any expert contacting the Steering Group from their own initiative.
4. invitation to experts recommended by the Delphi participants (snowballing).

Experts from the groups 1., 2. and 4. Were first contacted via email and presented the DECIDE-AI project through a dedicated two-page document (see **Annex V-1**). A special version of this document was developed for patient representatives (see **Annex V-2**). Upon acceptance, they were added to the Delphi participants list. The opportunity to recommend further experts was only offered to participants during the first Delphi round. Experts from group 3, as well as any other expert contacting us from their own initiative, were first screened to assess if they possessed adequate expertise (i.e. either a background in or recent publications about clinical AI or clinical evaluation). Upon successful assessment, they were

also added to the participants list. Experts were blinded to the names of other participants during the Delphi rounds. Informed consent was obtained electronically from all participants in the Delphi process before each round (see **Annex V-3**).

Initial item list development

I developed an initial list of candidate items based on (i) the studies included in the systematic literature review presented in Chapter III[85], (ii) an additional literature search about existing guidance for AI evaluation in clinical settings (see **Suppl. Note V-1**), (iii) literature recommended by the Steering Group and other experts in the field[84,135,174,181,192,335,336], (iv) institutional documents[92,142,144,337,338], and (v) expert opinion. The generated items were then reviewed and approved by the Steering Group. Short and descriptive explanation paragraphs were produced to introduce possibly difficult concepts (related to specific items) to the participants.

Round 1 - questionnaire design and distribution

The Delphi surveys were designed and distributed via the REDCap web application[339,340]. Data collection took place between February 22nd and March 25th 2021. REDCap is a secure web platform developed by a multi-institutional consortium led by Vanderbilt University, Nashville, Tennessee. It is a tool for the creation and management of online surveys and databases. REDCap was selected as one of the University of Oxford's recommended software for data collection due to its ease of use, widespread implementation (and hence abundance of training materials available online), and level of data protection.

The first-round questionnaire was first created in development mode and piloted with test participants (one medical student, one junior surgeon and one engineering DPhil candidate) independent of the Delphi participants. It consisted of four open-ended questions inviting the Delphi participants to comment on the key roles of early-stage clinical evaluation as described previously (real life performance, safety, human factors, preparatory step for larger trials). The

participating experts were then asked to rate, on a 1 to 9 scale, the importance of items in the initial list proposed by the Steering Group. 1 to 3 were defined as 'not important', 4 to 6 as 'important but not critical' and 7 to 9 as 'important and critical'. In other words, not important items should not be included in the guidelines; important but not critical items should be reviewed case by case by the consensus group; important and critical items should be included. An "I don't know" option was also available. Participants had the option to consult their answers to the four open-ended questions while scoring the items. They were also invited to comment on existing items (one comment box for each section) as well as on the guideline generally if they wished so. Additionally, they were offered the opportunity to suggest new items and justify them. Finally, the participants were asked to report their main affiliation, their workplace geographical localisation (only one choice) and the stakeholder group(s) they mostly affiliated with (several answers possible). They also had to describe, in free narrative, their level of experience and type of expertise with AI and machine learning, as well as with clinical evaluation and technology implementation. Participants were blinded to the other experts' answers during the round. A printout of the questionnaire is presented in **Annex V-4**.

Round 1 – analysis

The results of the first round of Delphi were analysed differently according to their nature. All analysis were performed independently by two reviewers and conflict resolved by consensus. Only complete set of answers submitted before the round closure deadline were included in the analysis.

- Free text narrative answers to the four open-ended questions: an inductive thematic analysis[209,211] was conducted in order to extract reporting item candidates potentially overlooked during the development of the initial list (see section V.2.2).
- Item scores: summary statistics were produced for each item, overall and broken down by stakeholder group (stakeholder groups with less than five representatives were aggregated to protect the confidentiality of their answers), including: median, interquartile range (IQR), percentage of participants scoring ≥7 (defined cut-off for inclusion), and percentage of participants scoring ≤3 (defined cut-off for exclusion).

Stakeholder groups with a median score difference from the overall item score of ≥2 points were highlighted.

- Item comments: item comments were collated and a narrative summary was produced for each item. In order to consider all comments equally and account for underrepresented groups, comment frequency was not calculated, nor reported.
- Proposed new items and their rational were considered and one of five actions was taken: i) select as new item, ii) include concept within existing item, iii) add the concept and rational to the provisory explanation of an existing item, iv) consider the concept and rational as a comment to inform the discussion around an existing item, or v) do nothing because the concept is already covered by an existing item. The exact wording of selected new items was defined by the research team.
- General comments were collated and a narrative summary produced.
- Personal information: Geographical origin and stakeholder group affiliation were reported in tabular form. The free text answers about participant expertise and experience were translated into general categories.

Item scores analysis and graphs production were performed on Python (Python Software Foundation, v.3.8.5, including standard packages). “I don’t know”, “Not sure” and NA answers were excluded from the analysis, but quantified as blank answers. For each item, a first round summary was produced, reporting on the same page: the summary statistics for all participants, any stakeholder group highlighted as having a median ≥2 points away from the overall median, a graphic comparing the median score and IQR overall and for each stakeholder group, the total number of comments received, and the narrative summary of comments received. This approach was defined by the Steering Group to inform the participants as transparently as possible about the round’s results, while reducing the influence of stakeholder group imbalance.

Based on this per item summary of the first-round results, the summary of new items proposed, and the results of the thematic analysis, a revised item list was drafted. Items from the initial list were either: carried over without modification, carried over with a wording modification, split, merged, or dropped. New items were also introduced to the revised list. A

change of wording represented a precision or development of the same broad topic, whereas mergers and split meant a change in the item focus. Two reviewers independently proposed their suggested action for each item and conflict were resolved by consensus. The resulting items were reordered and rearranged in main and sub-items. The first-round summary and revised list were circulated to the Steering Group for approval.

Round 2 – questionnaire design and distribution

The second round of Delphi was also distributed through the REDCap web platform. Data collection took place between May 2nd and May 26th 2021. The same design was used as for the first round, with modifications to account for the revised items. The following documents were included for the participants to consult and inform their answers:

- An executive summary of the first round (see **Annex V-5**)
- The per item summary of the first round described in section V.2.1 (see **Annex V-6**)
- The results of the thematic analysis (see **Annex V-7**)
- The summary of proposed new items (see **Annex V-8**)
- An overview of the revised item list (see **Suppl. Table V-1**)

Participants were asked to score the items of the revised list with the same instructions as during the first round. They could again submit item-specific and general comments, but could not propose new items. Given the substantial modifications made to the list between the first and second round, participants were also asked if the revised list was in their opinion qualitatively better than the initial list (“yes”, “no”, “similar”, “I don’t know”). Additionally, and in order to guide the future work of the consensus group, participants were asked how many items the final guideline should ideally have. The workplace geographical localisation, the stakeholder group(s) affiliation, and expertise of new participants were also collected. Participants were blinded to the other experts’ answers during the round. A printout of the questionnaire is presented in **Annex V-9**.

Round 2 – analysis

Item scores, comments and personal information were analysed in the same way as during the first round. The number of items each participant scored ≥7 (defined cut-off for inclusion) was compared to the number they reported as ideal target for the final guideline. All analysis were performed independently by two reviewers and conflict resolved by consensus. Only complete set of answers submitted before the round closure deadline were included in the analysis. “I don’t know”, “Not sure” and NA answers were excluded from the analysis, but quantified as blank answers. For the question about the ideal number of items to be included in the guideline, numbers below ten were considered as typing mistakes and excluded from the analysis. As for the first round, a per item summary was produced with the same features and circulated with the Steering Group for review. No modifications were made to the item list after the second round.

V.2.2 Thematic analysis

An inductive thematic analysis was performed, with identification of theme at a semantic level and with a high degree of granularity in order to produce a detailed description of the corpus of narrative answers to the four open-ended questions. No prevalence cut off was required for the consideration of a theme. Given the time constraint between the two rounds of Delphi, a streamlined version of thematic analysis was preferred. No familiarisation stage was performed. Coding and themes identification were performed independently by two reviewers, and conflict were resolved by consensus.

No thematic saturation analysis was performed, because it would have had no consequence on the data collection, the first round being already closed at the time of analysis. Moreover, given the small number of participants affiliating to some stakeholder groups and the relatively short length of most answers, the results of a thematic saturation analysis performed on the whole cohort would have been of only limited informative value about the true underlying

coverage of the subject from every stakeholder group's perspectives. The thematic analysis was performed using the NVivo software (QSR International Pty Ltd., v1.0).

Each identified theme was then either: i) selected as new; ii) included in an existing item; iii) added to the provisory explanation of an existing item; iv) considered as already covered by an existing item, out of scope or otherwise not appropriate for inclusion.

V.3 Results

V.3.1 Participant characteristics

Round 1

138 experts received a link to the first round of Delphi (95 identified from Steering Group recommendation, 12 from their publications, 21 by expression of interest following our commentary article or otherwise contacting the Steering Group from their own initiative, and 10 through snowballing), of whom 123 completed the questionnaire (89%), 83 identified from Steering Group recommendation, 12 from their publications, 21 by expression of interest following our commentary article or otherwise contacting the Steering Group from their own initiative, and seven through snowballing). The participating experts represented 16 countries and spanned all 20 of the defined stakeholder groups. **Table V-1** and **Table V-2** present an overview of the participants geographical origin and stakeholder group(s) affiliation.

Country	Number of participants	Country	Number of participants	Country	Number of participants
United Kingdom	64 (52%)	Italy	3 (2%)	Finland	1 (<1%)
United States of America	19 (15%)	Australia	2 (2%)	Portugal	1 (<1%)
The Netherlands	12 (10%)	France	2 (2%)	Singapore	1 (<1%)
Germany	5 (4%)	Austria	1 (<1%)	Spain	1 (<1%)
Canada	5 (4%)	Belgium	1 (<1%)		
Republic of Korea	4 (3%)	Brazil	1 (<1%)		

Table V-1: geographical origin of the participants in the first round of Delphi. (n = 123)

Stakeholder group	Number of participants	Stakeholder group	Number of participants
Clinicians	56 (46%)	Policy makers/official institutions staff	7 (6%)
Engineers/Computer scientists	43 (35%)	Regulators	7 (6%)
Methodologists	22 (18%)	Administrators/hospital management	6 (5%)
Implementation scientists	17 (14%)	Ethicists	6 (5%)
Statisticians	14 (11%)	Trialists	5 (4%)
Human factors specialists*	14 (11%)	Private sector representatives*	4 (3%)
Entrepreneurs*	12 (10%)	Patient representatives	3 (2%)
Epidemiologists	10 (8%)	Funders	2 (2%)
Journal editors	10 (8%)	Payers/Commissioners	1 (<1%)
Allied health professional	7 (6%)	Psychologists	1 (<1%)

Table V-2: self-reported stakeholder group affiliation in the first round of Delphi. 123 participants = 100%. Participants can be affiliated to more than one stakeholder group. * In total, 21 private companies of different sizes were represented.

Participants brought a wide range of expertise, both in clinical AI and clinical evaluation to the Delphi process, as summarised in **Table V-3**.

Expertise related to AI	Number of participants	Expertise related to clinical evaluation	Number of participants
Clinical evaluation of AI system	54	Implementation of other types of innovation	19
Development/training of models	48	Human factors evaluation of other technologies	16
Human factors/HCI/safety of AI technology	35	Clinical trials design, conduct or analysis	16
Implementation/translation of AI technology	34	Development of other technologies	7
Published or Reviewed AI recommendations/guidance	12	Ethics	6
Big data/images analytics	11	Innovation/business methodology	6
Advisory role for regulatory affairs	9	Risk management/quality/assurance	6
General AI methodology	8	Health economics	5
AI reporting guidelines development	3	Medical statistics	4
AI in low resource settings	2	Patient-centred outcomes	3
Editorial assessment of AI papers	2	Patient experience	3
		Systematic reviews	3
		Guidance development – non AI	1

Table V-3: expertise of the participants in the first round of Delphi.

Round 2

162 experts were sent the second round of Delphi questionnaire (97 identified from Steering Group recommendation, 13 from their publications, 31 by expression of interest following our commentary article or otherwise contacting the Steering Group from their own initiative, and 19 through snowballing), of whom 138 completed the questionnaire (85%, 85 identified from Steering Group recommendation, 12 from their publications, 28 by expression of interest following our commentary article or otherwise contacting the Steering Group from their own initiative, and 13 through snowballing). 110 of them had also completed the first round of Delphi (continuity rate = 89%). Out of the 28 new participants, ten were identified from Steering Group recommendation, one from their publications, nine by expression of interest following our commentary article or otherwise contacting the Steering Group from their own initiative, and eight through snowballing) completed the questionnaire. The participating experts represented 17 countries and spanned 19 out of 20 of the defined stakeholder groups. **Table V-4** and **Table V-5** present an overview of the participants geographical origin and stakeholder group(s) affiliation.

Country	Number of participants	Country	Number of participants	Country	Number of participants
United Kingdom	73 (53%)	Australia	3 (2%)	Finland	1 (<1%)
United States of America	22 (16%)	France	3 (2%)	Kenya	1 (<1%)
The Netherlands	13 (10%)	Italy	3 (2%)	Portugal	1 (<1%)
Germany	5 (4%)	Austria	1 (<1%)	Singapore	1 (<1%)
Canada	5 (4%)	Belgium	1 (<1%)	South Africa	1 (<1%)
Republic of Korea	3 (2%)	Brazil	1 (<1%)		

Table V-4: geographical origin of the participants in the second round of Delphi. (n=138)

Stakeholder group	Number of participants	Stakeholder group	Number of participants
Clinicians	65 (47%)	Policy makers/official institutions staff	7 (5%)
Engineers/Computer scientists	45 (33%)	Administrators/hospital management	6 (4%)
Methodologists	29 (21%)	Trialists	6 (4%)
Statisticians	20 (14%)	Ethicists	5 (4%)
Implementation scientists	19 (14%)	Private sector representatives*	5 (4%)
Epidemiologists	14 (10%)	Patient representatives	4 (3%)
Entrepreneurs*	13 (9%)	Regulators	4 (3%)
Journal editors	13 (9%)	Funders	3 (2%)
Human factors specialists*	12 (9%)	Psychologists	1 (<1%)
Allied health professional	9 (7%)	Payers/Commissioners	0 (<1%)

Table V-5: self-reported stakeholder group affiliation in the second round of Delphi. 138 participants = 100%. Participants can be affiliated to more than one stakeholder group. * In total, 26 private companies of different sizes were represented.

Participating experts brought a wide range of expertise and experience to the Delphi process, as summarised in **Table V-6**. The names and affiliations of the Delphi participants can be found in **Suppl. Note VI-2**.

Expertise and experience related to AI	Number of participants	Expertise and experience related to clinical evaluation	Number of participants
Clinical evaluation of AI system	59	Clinical trials design, conduct or analysis	21
Development/training of models	59	Implementation of other types of innovation	20
Implementation/translation of AI technology	37	Human factors evaluation of other technologies	14
Human factors/HCI/safety of AI technology	32	Development of other technologies	9
Published or Reviewed AI recommendations/guidance	12	Health economics	7
Big data/images analytics	11	Ethics	5
General AI methodology	10	Risk management/quality/assurance	5
Advisory role for regulatory affairs	6	Innovation/business methodology	4
AI reporting guidelines development	4	Medical statistics	4
AI in low resource settings	2	Patient-centred outcomes	4
Editorial assessment of AI papers	1	Systematic reviews	3
		Patient experience	2
		Guidance development – non AI	1

Table V-6: expertise and experience of the participants in the second round of Delphi. AI = artificial intelligence.

V.3.2 Initial item list

Table V-7 shows the initial item list used during the first round of Delphi.

Item n°	Recommendation
Title	
1	Identify the study as early-stage, exploratory or first-with-human clinical evaluation of an artificial intelligence or machine learning based decision support algorithm.
Abstract	
2	Provide a structured summary of the study, including mention of: [will be completed according to the outcomes of the Delphi]
Introduction	
3	Describe the target conditions and the patient population that would benefit from the algorithm, including information on the target conditions' prevalence and their impact on the healthcare system.
4	Describe the intended use of the algorithm, including its position in the care pathway and the conditions under which it would be used, the impact in terms of patient care it intends to achieve and the current state of the art practice.
5	Refer to the algorithm's development and validation studies and name the algorithm, including the version number. State the algorithm's expected performance from development and validation studies.
6	Identify the dataset used to develop the algorithm and provide information on its relevance to the test environment, including the target conditions' prevalence when appropriate.
7	Describe the current stage of development of the algorithm in terms of the essential questions which have been and remain to be answered about it (both from a scientific and a regulatory perspective).
8	State the study objectives.
Methods	
9	Provide a reference to any study protocol.
10	Specify the primary and secondary outcome measures.
11	Describe the study design using standard methodological terminology.
12	Describe precisely how users and patients were selected. If only a subgroup of users took part in the human factors evaluation, describe how these were selected.
13	Justify the sample sizes (for both users and patients).
14	Identify the hardware and software platforms used during the study. Describe the data needed by the algorithm as inputs, the data provided by the algorithm as outputs and the minimal computational resources needed.
15	Describe how the algorithm was used, at which stage of the decision-making process and who held the responsibility for the final clinical decision.
16	Describe precisely the environment in which the algorithm was tested, including the availability of the algorithm's input data and which additional clinical information (i.e. not provided by the algorithm) was accessible to the users.
17	Describe how the patient data were acquired (including from which sources), how they were processed and how missing or low-quality data were handled.
18	State what measures were taken to protect patient privacy and data security.
19	Describe the control group in sufficient detail to allow replication.

20	State the predefined statistical analysis plan and any additional exploratory analyses performed (state if the chosen approach accounts for both user and patient variability).
21	State any pre-specified subgroup analyses and their rationale.
22	Define the algorithm safety requirements, how these were established and how compliance to these requirements was evaluated.
23	Describe how algorithm recommendation/output errors were defined and how they were identified.
24	Describe any attempts to familiarise users with the algorithm, including any training received.
25	Describe the human factors tools, methods or frameworks used to evaluate usability, situation awareness and any other relevant human factors considerations. Justify this choice.
26	Describe any attempt to understand the user acceptance of the algorithm as well as user deviations from the algorithm's recommendations or intended use.
27	Describe any involvement of patients in understanding their opinion on the algorithm and how the algorithm's outputs could influence their care.
28	Describe the methodology used to collect and analyse data for the health economic assessment of the algorithm's use.
Results	
29	Describe the user population baseline characteristics (number, number of centres, specialty, seniority, previous experience with digital support, etc.)
30	Describe the patient population baseline characteristics (number, number of centres, age, sex, ethnicity if relevant, comorbidities, prevalence of the target conditions, etc.).
31	Report on the proportion of intended users who had exposure to the algorithm (implementation reach), on the number of instances the algorithm was used (implementation dose) and on the users' compliance with the intended implementation (implementation fidelity).
32	Report on the prespecified outcomes for the algorithm-assisted users (both overall and at an individual user level) as well as for the control group.
33	If applicable, report on the prespecified outcomes which would have been observed had all the algorithm's recommendations been strictly followed.
34	Describe any instances where the algorithm gave an erroneous recommendation/output. Report their rate of occurrence and detail their potential impact on patient care.
35	Report on the compliance with the safety requirements.
36	Describe any instances where users decided to override the algorithm's recommendation or to follow an erroneous recommendation.
37	Report on the evolution of users' trust in the algorithm (evolution of the overrides of the algorithm's recommendations with time) and on the learning curves (evolution of the users' performance with time).
38	Report the number of users involved in the human factors evaluation, their characteristics and the use cases examined.
39	Report on the usability evaluation, including time to task completion and display interface evaluation, using method-specific metrics.
40	Report on the situation awareness evaluation and on the users' perspective on the algorithm's interpretability.
41	Report on the outcomes of any other human factors evaluation, including the user acceptance of the algorithm and any induced changes in the care pathway.
42	Summarize all changes made to the algorithm or its hardware platform between the prototype used at the beginning of the study and its final version. Report the timing of these modifications and the changes in outcomes observed after each design-evaluation cycle.
43	Report the patients' opinion on the algorithm and whether they would accept their care being influenced by it.
44	Report the results of the health economic assessment of the algorithm's use and identify any trade-offs in the care pathway.

Discussion	
45	Discuss if the obtained results support the intended purpose of the algorithm in real world healthcare settings, including how the outcomes would translate into patient benefit, or if an alternative use could be more appropriate.
46	Explain what was learned about the reasons for human deviation from the algorithm's recommendations or intended use, and what this tells us about achieving better alignment.
47	Discuss the algorithm's errors and identify any underlying pattern or algorithmic bias. Explain how these can be mitigated.
48	Discuss what the results suggest about the safety profile of the algorithm.
49	Discuss the human factors results and comment on the evolution of the algorithm/hardware platform design. Discuss the need for additional technical requirements or product design improvement before large-scale summative evaluation.
50	Comment on the evolution of users' trust in the algorithm and on the learning curves. State when they reached a stable state.
51	Highlight any performance difference in user or patient subgroups and discuss the merits of limiting further evaluation to a specific group of users or patients.
52	Discuss the feasibility and appropriateness of large-scale summative evaluation in light of the obtained results.
Statements	
53	Disclose the source of funding for the study and authors' relevant conflicts of interest.
54	Disclose code and data availability.

Table V-7: initial item list.

V.3.3 Round 1 – item scores and comments

123 set of answers were submitted during the first round of Delphi. One set of score was excluded from the analysis due to a suspicion of scale inversion and two sets of answers could not be included in the analysis because they were submitted after the deadline. Round 1 yielded 6419 item scores and 61 blank scores, 185 item-related comments (for a total of 7724 words, some comment providing individual feedback about several items or about a whole section), and 43 general comments (for a total of 1929 words). **Figure V-2** shows an example of score medians and interquartile range (IQR) for items 1 to 18, broken down per stakeholder group categories, as they were used for the analysis and presented to participants. The full results for all items in Round 1 can be found in **Suppl. Figure V-1a-c**. **Table V-8** shows the percentage of participants voting to include or exclude each item as well as any stakeholder group scoring the item ≥2 points above or below the overall median. **Annex V-6** presents a per item summary of the Round 1 results (including statistic summary, graph of median and IQR for each stakeholder group and a narrative summary of comments related to the item).

Item	Median score	25th perc.	75th perc.	% vote include	% vote exclude	Stakeholder divergence	Item	Median score	25th perc.	75th perc.	% vote include	% vote exclude	Stakeholder divergence
Item 1	8	6	9	71	7	TR	Item 28	6	5	7	43	12	-
Item 2	8	7	9	79	3	-	Item 29	8	7	9	88	1	-
Item 3	8	7	9	88	0	-	Item 30	9	8	9	96	0	-
Item 4	9	8	9	93	1	-	Item 31	8	7	9	81	3	-
Item 5	8	7	9	78	6	-	Item 32	8	7	9	89	2	-
Item 6	8	6	9	72	6	-	Item 33	7	6	8	66	10	-
Item 7	7	6	8	60	5	-	Item 34	9	7	9	91	0	RE
Item 8	9	8	9	97	0	-	Item 35	8	7	9	82	0	-
Item 9	8	7	9	76	3	OTH	Item 36	9	7	9	88	3	-
Item 10	9	8	9	88	1	-	Item 37	7	6	8.25	65	5	-
Item 11	8.5	7	9	83	1	-	Item 38	7	6	9	67	5	TR
Item 12	9	8	9	93	0	OF, RE	Item 39	7	7	8	76	3	-
Item 13	8	6	9	72	3	-	Item 40	7	6	8	65	8	TR
Item 14	8	6	9	72	5	-	Item 41	7	6	8	70	5	-
Item 15	9	8	9	92	2	-	Item 42	8	7	9	78	3	OTH
Item 16	8	7	9	85	1	OTH	Item 43	7	5	8	56	13	PS, RE, TR
Item 17	8	7	9	87	0	-	Item 44	6	5	7	37	15	RE, AHP
Item 18	7	6	9	64	3	JE, AHP, AD	Item 45	8.5	7	9	87	3	-
Item 19	8	6.75	9	75	2	-	Item 46	8	7	9	86	3	-
Item 20	8	7	9	80	3	OTH	Item 47	9	7	9	89	3	RE
Item 21	7	6	8	70	3	-	Item 48	8	7	9	90	2	-
Item 22	8	7	9	84	2	-	Item 49	7	6	8	71	4	-
Item 23	8	7	9	90	0	-	Item 50	7	6	8	60	8	PS, TR
Item 24	8	7	9	82	1	-	Item 51	7.5	7	8.25	78	6	-
Item 25	7	6	8.5	71	1	RE, TR	Item 52	8	7	9	82	3	OF
Item 26	8	6.5	9	75	4	-	Item 53	9	8	9	98	0	-
Item 27	7	5.5	8	55	10	TR	Item 54	8	7	9	80	4	TR

Table V-8: summary statistics of the first round of Delphi. A score ≥7 was defined as a recommendation to include and a score ≤3 as a recommendation to exclude. A stakeholder group divergence was defined as a median score ≥2 points above or below the overall median. AD = Administrators; AHP = Allied Health Professionals; JE = Journal Editors; OF = Policy Makers & Official Institutions; OTH = Others (in Round 1: Funders, Patient representatives, Payers & Commissioners, and Psychologists); PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

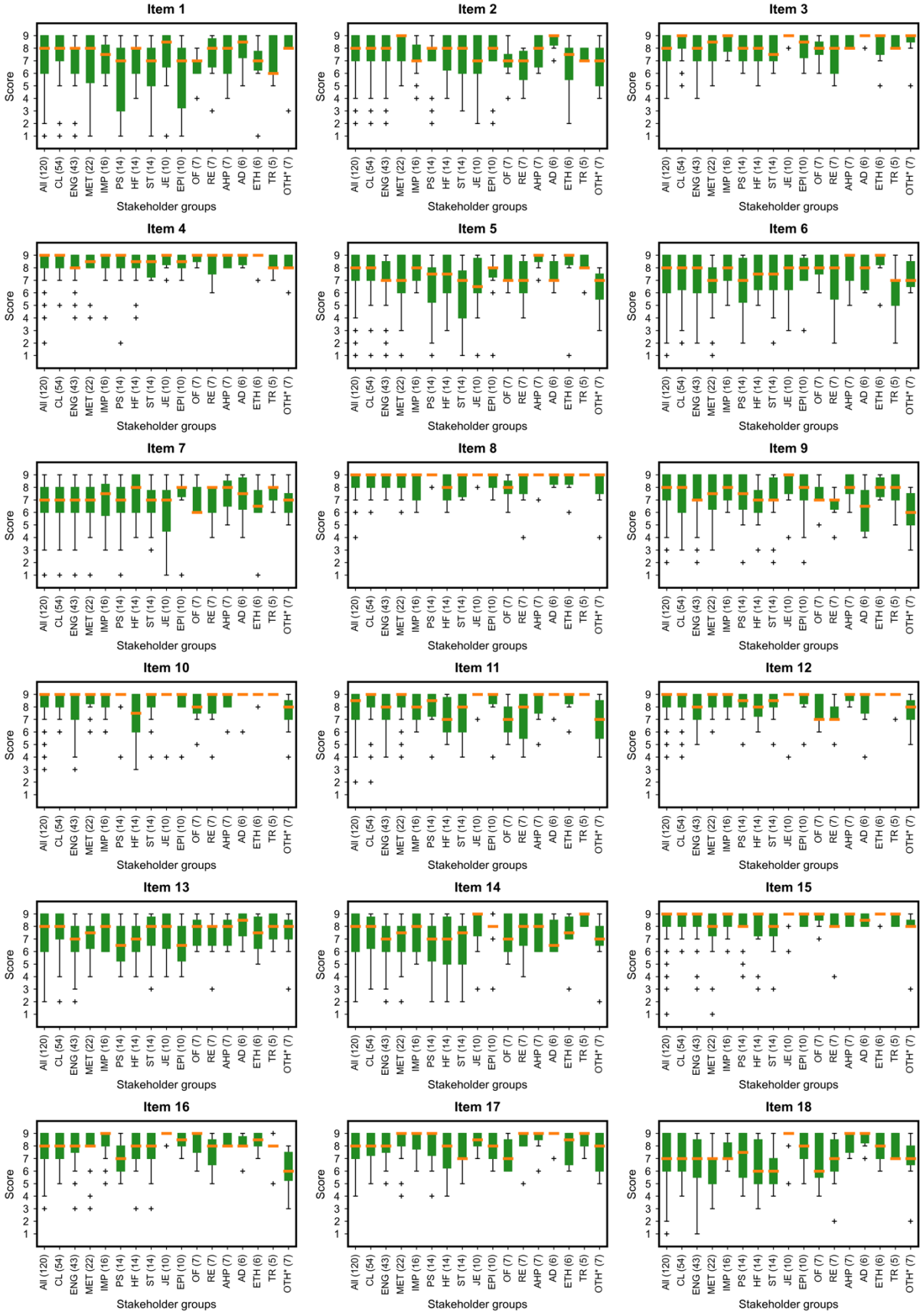


Figure V-2: median scores and IQR of item 1 to 18 during the first round of Delphi, overall and by stakeholder group. Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. *Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

V.3.4 Round 1 – new items proposed

64 propositions for new items were submitted during the first round of Delphi. Four were selected as new items (two were merged for a total of three new items), 16 included in existing item, 12 added to the provisory explanation of an existing item, 16 added to the comment of existing items, and 16 were already covered by an existing item or standard editorial requirements. **Annex V-8** presents a detailed summary of the propositions, their rationale, and action taken. The three newly selected items related to:

- Ethics methodology and consultations
- Algorithmic bias and subgroup analysis (both patients and users)
- Strength and limitations

V.3.5 Round 1 – thematic analysis

Round 1 yielded 482 valid and 2 blank answers to the four open-ended questions (for a total of 43,986 words of narrative answers). Reviewer 1 identified 143 themes/sub-themes and Reviewer 2, 71 themes/sub-themes. After consensus, 109 themes were kept. Nine were selected as new items (two merged into one item and two already proposed as new item, for a total of six new items), 14 included in existing item, 25 added to the provisory explanation of an existing item, and 61 were either already covered by an existing item, out of scope or otherwise not appropriate for inclusion. **Annex V-7** presents a summary of the themes identified and actions taken. The newly selected items related to:

- Algorithm or hardware malfunction
- Human errors
- Unexpected errors and newly identify safety risk (two items)
- Ethics (already proposed by participants under section V.3.4)
- Workload
- Integration in pathway and workflow
- Subgroup analysis (already proposed by participants under section V.3.4)

V.3.6 Revised item list

Table V-9 shows the revised item list. The revised list was developed based on the feedback from the first round of Delphi and used during the second round of Delphi. Topic headers were added to the list to support the thematic organisation of the guideline and Delphi items reorganised into main and sub-items to align with this new subdivision. Overall, nine new items were added (see sections V.3.4 and V.3.5), two items were dropped, nine items remained unchanged, 22 items were reworded/completed, 21 items were reorganised (merged, split, or both, becoming 13 items). The revised list comprised 53 items. The two items dropped were related to health economic assessment and were the only two items with a median score below 7 (median: 6, IQR: 2–9 for both). Moreover, several comments described health economic assessment as an entirely separate aspect of evaluation. A detailed diagram of item addition, deletion, and reorganisation between the two rounds can be found in **Figure VI-2**.

Topic	Item	Recommendation
Title and abstract		
Title/abstract	1a	Identify the study as early stage or formative clinical evaluation of an artificial intelligence or machine learning based decision support system, mentioning the clinical problem addressed.
	1b	Provide a structured summary of the study, including: target clinical problem, intended use of the algorithm and integration in the clinical pathway, type of algorithm, study design, study setting, number of patients and users included, control group if applicable, primary and secondary outcomes, key safety endpoints, human factors aspects evaluated, main results, conclusions.
Introduction		
Target clinical problem and population	2	Describe the target clinical problem and medical condition, including the current state of the art practice, and the target patient population.
Intended use	3	Describe the intended use of the algorithm, its planned integration in the care pathway and the impact in terms of patient outcomes it intends to achieve.
Current stage of development	4	Describe the current stage of development of the algorithm (both from a scientific and a regulatory perspective). State if the algorithm is tested as a medical device and, if so, which regulatory approval is sought/was obtained.
Objectives	5	State the study objectives.
Methods		
Research governance	6a	Provide a reference to any study protocol, study registration number and ethics approval.
	6b	State what measures were taken to protect patient privacy and data security.
Study design	7	Describe the study design.

Participants	8a	Describe precisely how patients were recruited, stating the inclusion and exclusion criteria, and how the number of recruited patients was selected.
	8b	Describe precisely how users were recruited, stating the inclusion and exclusion criteria, and how the number of recruited users was selected. If applicable, describe the control group in sufficient detail to allow replication.
	8c	Describe any attempts to familiarise the users with the algorithm, including any training received.
Algorithm	9	Briefly describe the algorithm, including: the version number, the type of AI model used, the characteristics of the patient population on which it was trained and the expected performance from in silico study. Refer to any previous development work.
Implementation	10a	Describe precisely the environment in which the algorithm was tested, including the availability of the algorithm's input data and which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm.
	10b	Describe the clinical workflow/pathway in which the algorithm was deployed and who held the responsibility for the final clinical decision.
	10c	Describe precisely how the algorithm was used and the timing of the decision support.
	10d	Describe the technical details of the implementation, including the integration within the existing study site IT infrastructure, the software and hardware needed to run the algorithm and any algorithmic thresholds used.
	10e	Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, any pre-processing applied and how missing/low-quality data were handled.
	10f	Describe the algorithm outputs and how they were presented to the users.
Outcomes	11a	Specify the primary and secondary outcomes measured.
	11b	Describe how algorithm recommendation/output errors were defined and how they were identified.
Analysis	12	Describe the pre-specified analysis plan for the primary and secondary outcomes as well as for any prespecified additional analyses, including subgroup analyses and their rationale.
Safety	13a	Define the algorithm safety requirements, how these were established preclinically, and how compliance to these requirements was evaluated during the study.
	13b	Describe the methodology used to detect any new, unexpected risks arising from the real-life clinical use of the algorithm.
Human factors	14	Describe the human factors tools, methods or frameworks used, the use cases considered and the users involved in the human factors evaluation.
Patient engagement	15	State whether patients were involved in any aspect of the study design, conduct or in the development of the research question or outcome measures.
Ethics consideration	16	Describe any ethics methodology, consultation or involvement during the design or implementation of the study.
Results		
Participants	17a	Describe the patient study group baseline characteristics (number, number of centres, age, sex, ethnicity if relevant, comorbidities, prevalence of the target conditions, etc.).
	17b	Describe the users study group baseline characteristics (number, number of centres, specialty, seniority, previous experience with digital support, etc.).
Implementation	18a	Report on the user exposure to the algorithm (implementation reach), on the number of instances the algorithm was used (implementation dose) and on the users' adherence to the intended implementation (implementation fidelity).
	18b	Report changes caused by the algorithm to the clinical workflow, if any.

Modifications	19	Report any changes made to the algorithm or its hardware platform between the prototype used at the beginning of the study and its final version. Report the timing of these modifications and the changes in outcomes observed after each of them.
Main results	20a	Report on the prespecified outcomes for the algorithm-assisted users (both overall and at an individual user level), including any variation over time.
	20b	Report on the prespecified outcomes for the stand-alone algorithm, if applicable.
	20c	Report on the prespecified outcomes for the control group, if applicable.
Safety and errors	21a	Report on the compliance with the specified safety requirements and any severe adverse events.
	21b	Report any additional risks identified from the real-life clinical use of the algorithm.
	21c	Report any algorithm malfunction or issues with hardware or software during the study.
	21d	Report any algorithm recommendation errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual impact on patient care.
	21e	Report any human errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual implication for patient care.
Subgroup analysis	22	Report on the difference in the main outcomes according to the specified subgroups.
Human factors	23a	Report on the user agreement with the algorithm. Describe any instances of and reasons for user deviation from the algorithm's recommendations and, if applicable, user changing their mind based on the algorithm recommendations.
	23b	Report on the evolution of users' trust in the algorithm.
	23c	Report on the usability evaluation, including time to task completion, display interface evaluation and user satisfaction.
	23d	Report on the user workload and learning curves evaluation.
	23e	Report on the user perception of the algorithm outputs' interpretability and clinical value.
Discussion		
Support intended purpose	24	Discuss whether the obtained results support the intended purpose of the algorithm in real world clinical settings.
Safety and errors	25	Discuss what the results suggest about the safety profile of the algorithm. Discuss the algorithm's errors and, if appropriate, identify any underlying pattern or algorithmic bias, explain how these can be mitigated.
Human factors	26	Discuss the results of the human factors evaluation and the reasons for human deviation from the algorithm's recommendations or intended use.
Scale up	27	Discuss the scale-up feasibility and requirements, as well as the possible design of large-scale summative evaluation in light of the obtained results. Summarise the lessons learned from the study.
Strength and limitations	28	Discuss the strengths and limitations of the study, including any bias in the study design.
Statements		
Conflicts of interest	29	Disclose any relevant conflict of interest, including: the source of funding for the study, the role of funders, any other role played by commercial compagnies and authors' conflicts of interest.
Data Availability	30	Disclose if and how data and code (pre-processing and algorithm) are available.

Table V-9: revised item list. New items, either directly proposed by participants or extracted from the thematic analysis, are shaded in grey.

V.3.7 Round 2 – item scores and comments

138 set of answers were submitted during the second round of Delphi. One set of answer was excluded from the analysis due to obvious lack of consideration for the question (all but one item scores were equal to 1) and one set of answers could not be included in the analysis because they were submitted after the deadline. Round 2 yielded 7107 item scores and 101 blank scores, 882 item-related comments (for a total of 18,772 words), and 41 general comments (for a total of 2195 words). **Table V-10** shows the percentage of participants voting to include or exclude each item as well as any stakeholder group scoring the item ≥2 points above or below the overall median. **Suppl. Figure V-2a-c**. presents the detailed score medians and interquartile range (IQR) for all items in Round 2, broken down per stakeholder group categories. **Annex V-10** presents a per item summary of the Round 2 results (including statistic summary, graph of median and IQR for each stakeholder group and a narrative summary of comments related to the item). No changes occurred during the analysis of Round 2 and the results were exclusively used to inform the discussion of the Consensus Group.

Item	Median score	25th perc.	75th perc.	% vote include	% vote exclude	Stakeholder divergence	Item	Median score	25th perc.	75th perc.	% vote include	% vote exclude	Stakeholder divergence
Item 1a	8	8	9	93	0	-	Item 17a	9	8	9	98	0	-
Item 1b	8	8	9	98	0	-	Item 17b	8	8	9	96	0	-
Item 2	8	8	9	96	0	-	Item 18a	8	7	9	87	1	-
Item 3	9	8	9	96	1	-	Item 18b	8	7	9	86	1	-
Item 4	7	6	8	72	2	-	Item 19	8	8	9	89	1	-
Item 5	9	8	9	99	1	-	Item 20a	8	8	9	92	1	-
Item 6a	8	8	9	90	0	-	Item 20b	8	6	8	74	5	IMP
Item 6b	7	6	8	69	4	ETH	Item 20c	8	7	9	91	2	-
Item 7	9	8	9	99	1	-	Item 21a	9	8	9	93	0	-
Item 8a	9	8	9	93	2	-	Item 21b	8	8	9	91	2	-
Item 8b	9	8	9	90	0	-	Item 21c	8	7	9	83	1	-
Item 8c	8	7	9	85	2	-	Item 21d	9	8	9	97	0	-
Item 9	8	8	9	90	0	-	Item 21e	8	7	9	85	1	-
Item 10a	8	8	9	95	0	-	Item 22	8	7	9	78	4	-
Item 10b	8	8	9	91	1	-	Item 23a	9	8	9	93	1	-
Item 10c	8	8	9	93	2	-	Item 23b	7	5	8	55	6	PS, JE, ETH
Item 10d	8	6	8	73	5	HF, ETH, OF/REG	Item 23c	7	7	8	77	2	-
Item 10e	8	8	9	96	0	-	Item 23d	7	6	8	62	4	-
Item 10f	8.5	8	9	97	1	-	Item 23e	7	6	8	67	5	HF
Item 11a	9	8	9	96	0	-	Item 24	9	8	9	96	0	-
Item 11b	8	8	9	93	1	-	Item 25	9	8	9	97	1	-
Item 12	8	7	9	85	1	-	Item 26	8	7	8	84	2	-
Item 13a	8	7	9	90	2	-	Item 27	8	7	8	78	2	-
Item 13b	8	7	9	84	2	-	Item 28	9	8	9	96	0	-
Item 14	7.5	7	8	76	2	-	Item 29	9	9	9	98	0	-
Item 15	7	6	8	57	7	-	Item 30	8	8	9	93	1	-
Item 16	7	6	8	62	10	-	-	-	-	-	-	-	-

Table V-10: summary statistics of the second round of Delphi. A score ≥7 was defined as a recommendation to include and a score ≤3 as a recommendation to exclude. A stakeholder group divergence was defined as a median score ≥2 points above or below the overall median. ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators.

V.3.8 Round 2 – supplementary questions

Qualitative evaluation of the list evolution

97 participants answered that the revised item list was qualitatively better than the initial list, 20 thought that they were qualitatively similar, 3 answered that the revised item list was not qualitatively better, and 16 had no opinion.

Selected and desired guideline length

There was a marked discrepancy between the number of items the participants answered the guideline should contain, and the number of item they actually recommended to include. **Figure V-3** shows the distribution of values given and numbers of items scored ≥7.

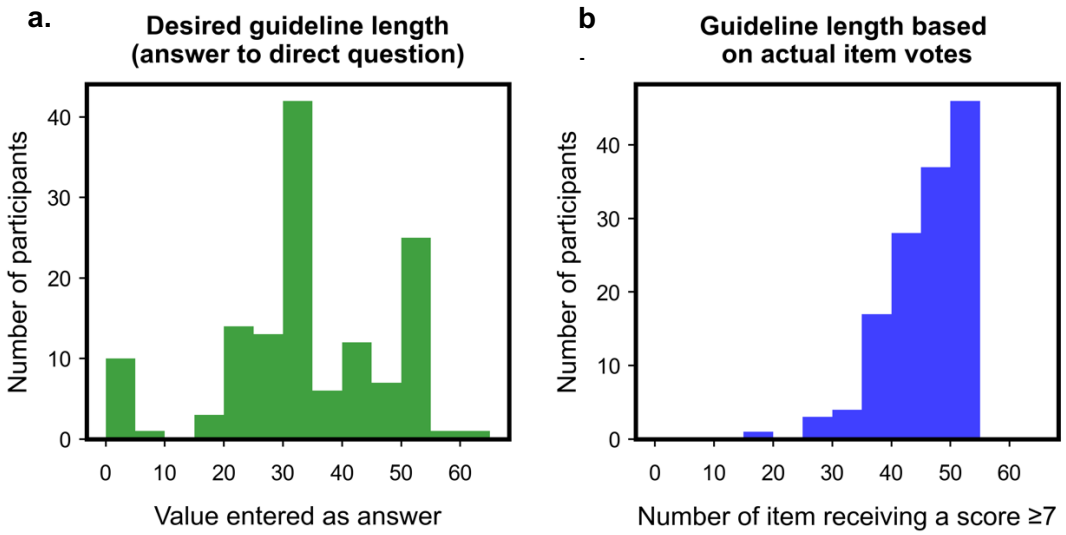


Figure V-3: desired (a.) vs actual (b.) guideline length. Total number of answers = 136

V.4 Discussion

The two first rounds of the Delphi process provided a refined list of reporting items, collected detailed feedback on the proposed items from a large group of experts, and analysed this feedback. These results were presented to the DECIDE-AI Consensus Group and informed their discussion during the consensus meeting (see Chapter VI).

The original item list was 54 items long and it was hoped that the first round of Delphi would give indications on where to shorten it, along with proposition of new items. However, most of the initial items proved popular and only two received a median score smaller than seven (the defined threshold for inclusion). These were the two items on the health economic assessment of the AI system, which many participants considered as a separate line of research. Considering the addition of nine new items proposed by the participants, it was hence necessary to reorganise the items in order to produce a revised list of a manageable length for the participants in the second round. Items were merged, split, and modified in a revision informed by the comprehensive feedback received in both open-ended answers and item-specific comments.

This rearrangement and editing might have led to a more marked difference between the initial and revised item list than in other Delphi processes. Although allowing such changes gave a stronger influence to participants' opinion and reinforced the consensus nature of the DECIDE-AI Delphi, some might question the impact of such extended modifications. This is the reason why, in order to assess if the editing done in-between rounds was adequate, a subsidiary question was added at the end of the second round, asking the participant to judge the qualitative improvement of the list between the rounds. Ninety-seven out of 120 participants (81%, excluding 16 abstention) agreed with a qualitative improvement. Another indicator that the item modifications performed were appropriate was that the level of consensus increased between the first and second rounds. Whereas in theory, the higher level of consensus

observed in subsequent rounds is attributed to participants reading each other's comment and considering diverging opinion, many participants may not in practice have read the full summary of the first round and rather based their vote in the second round uniquely on the revised items themselves.

Despite a higher level of consensus in the second round, the final results did not provide clear indications about which items should be selected for the final guidelines (all items obtained a median score of seven or more). While around half of the participants wished a guideline of 30 items or less, the vast majority of them scored over 40 items with a score recommending inclusion. In this regard, the two first Delphi rounds acted principally as a refinement of the initial item list, validation of the revised list, and a source of information for the Consensus Group, leaving this latter with extended responsibility to shortlist the final DECIDE-AI guideline.

Representativity is an important aspect of guideline development. However, it is also an ill-defined concept in this context and one that can be difficult to balance with other interests such as expertise requirements or logistics. A diverse group of experts is beneficial to contribute different perspective to a subject and to increase the chances of adoption if the guideline addresses the preoccupation of a greater number of its potential users. Little evidence exists, though, to inform how much diversity exactly is appropriate; a fact reflected in the widely different ways of reporting information relative to participant diversity, the type of diversity considered, and the number of groups described found in the literature. SPIRIT-AI and CONSORT-AI, for example, only consider expertise diversity and define 12 stakeholder groups[167,168]. TRIPOD also reports on expertise diversity, but defines only five stakeholder groups, as does the 2015 STARD update with six groups[170,172]. The IDEAL reporting guidelines describes six groups based on expertise, five based on the number of years of experience, six groups based on geographic origin and two based on sex[192]. As a comparison, DECIDE-AI reported on expertise and geographical diversity, defining 20 stakeholder groups and reporting country origins of the participants. In the case of DECIDE-AI, participation required a certain degree of expertise in either clinical machine learning or clinical evaluation

methodology. Both necessitate environments with large quantities of medical data available in digital form, funding for innovation evaluation and clinical trials, and/or performing research institutions. The pool of available experts was therefore unfortunately not equally distributed, or connected, across the world. For example, out of 30 DECIDE-AI participants who contacted the research team after the publication of an open call to contribute[105], or otherwise of their own initiative, 26 were from Europe and only one from a region which could be considered as underrepresented. The underrepresentation of certain geographical areas in the debate around AI is a serious issue reflecting disparities in the global access to high education and modern healthcare. However, correcting these inequalities falls outside the scope of guideline development, which should instead focus on minimising the impact of the disparities caused by this situation. Finally, the actual diversity in a Delphi process can be challenging to predict as participants don't necessarily see themselves belonging to the same groups as the research team would expect. A way to address this issue would be to reactively adapt the invitation list as the rounds progress, but this could in turn become a logistical issue as it would prolong these phases of the study.

Another aspect relating to the study conduct is the engagement of patient representatives. Patient and public involvement with research (PPI) is currently heavily supported by public institutions, funders, and journals. Whereas the full impact of PPI is still debated, it is increasingly accepted that PPI increases the relevance of research and can help study conduct by tailoring design to patient considerations[341]. In the context of the DECIDE-AI guideline development, adequate PPI was also considered an important factor. Patient representatives were approached and invited to participate from an early stage. However, out of the five patient representatives amongst the participants, one of them asked to end their participation and another considered doing so. Both stated that the questions asked were very technical and that they could therefore not contribute as meaningfully as they would wish. This raised the question of the most suitable PPI format for guideline development. In the case of DECIDE-AI, the PPI strategy was inspired by similar reporting guideline recently

developed[167,168]. A specific project presentation document (see **Annex V-2**) was created and individual call with patient representatives organised to explain the project and answer any of their questions. The patient representatives were then considered as any other participants and invited to complete the Delphi questionnaires. In retrospect, the value of PPI for guideline development doesn't lie in a specific number of participants from this stakeholder group, but from in depth discussions about selected items, which were eventually conducted as response to the mentioned PPI concerns.

V.4.1 Strengths and limitations

Compared to other consensus-based guideline development processes, the DECIDE-AI Delphi is based on one of the largest number of participants. SPIRIT-AI and CONSORT-AI, for example, had 103 participants[167,168]. TRIPOD was based on the feedback of 27 experts and the IDEAL reporting guideline had 54 respondents to the first round[170,192]. All the 20 predefined stakeholder groups were represented in at least one of the rounds, with a response rate of 89% and 85% in the first and second round respectively. The continuity rate of participants from the first to second round was 89%.

The DECIDE-AI Delphi was also characterised by a high level of engagement of the participants. In total, 74,606 words of comments and 64 suggestions of new items were submitted. This is an average of 291 words (approximately the equivalent of one of this thesis' page) per participant and round. The research team also strived to be as transparent as possible with the outcomes of the rounds, publishing 130 pages of summary in total. The consensus process was based on a recognised methodology, widely used for similar exercises in the field. It is also to the best of my knowledge, the first guideline development process to conduct a full and double reviewed thematic analysis on participants' answers.

However, the two first rounds of the Delphi process should be interpreted in the context of several limitations. First of all, any consensus process is susceptible to biases, such as anchoring or participant selection biases[342]. Although the research team did its best to mitigate them through adaptation of the study and questionnaire design, it is likely that not all could be totally controlled.

Geographical representation, for example, was skewed toward Europe and more specifically the United Kingdom. This geographical imbalance was the result of a combination of factors, some of which outside the research team's control. All of the major AI reporting guidelines, currently published or in development, were initiated, at least partly, from the UK. Following good practice recommended by the EQUATOR Network, main authors from these guidelines were invited to the DECIDE-AI Steering Group. The Steering Group was therefore predominantly British from the outset, which certainly had an impact on the Steering Group recommendations of experts to invite. As already explained above, the response rate of the invited participants and the primary geographical location they chose to register were moreover not always the one expected from publicly available information. Therefore, the actual geographical representation, measured at the end of the rounds, did not necessarily reflect the one expected from the invitation logbook. However, several actions were taken to mitigate this issue: i) non-UK experts were invited to join the Steering Group; ii) active efforts were made to augment the number of non-UK experts on the invitation list; iii) invitation were sent to all first authors of the publications included in the initial systematic review, regardless of their geographic origin; iv) an open call to contribute to the guideline development was published in Nature Medicine, a non-UK journal with global reach and international readership[105], even if, out of the 30 people who contacted the research team of their own initiative, 26 were from Europe (including 17 from the UK); v) an active effort was made to contact African medical AI experts through the Africa Oxford Initiative; and vi) all themes identified during the thematic analysis were considered regardless of their prevalence, in order to consider underrepresented opinions equally.

Despite the research team's effort to balance stakeholder representation, clinicians and engineers were the most represented groups, partly due to the profile of researchers who contacted us of their own initiative. To mitigate this, all scores analysis were performed overall and broken down per stakeholder group, in order to identify any marked disagreement from less represented groups.

Finally, few examples of early-stage clinical evaluation of AI tools were available at the time the initial item list was developed, so that some important reporting item might have been missed. A similar issue was noted by the authors of the SPIRIT-AI/CONSORT-AI guidelines[167,168]. Nonetheless, most invited authors and experts had an informed opinion on the reporting requirements for the evaluation of AI system and/or clinical evaluation of innovation, and the first-round questionnaire was designed to capture any missed item.

V.4.2 Conclusion

In conclusion, the DECIDE-AI first and second rounds of Delphi have produced and refined an exhaustive list of reporting item about the early-stage live clinical evaluation of AI systems. All of the items presented in the revised list were resubmitted to the expert group during the second round, and the collected scores and comments summarised in order to inform the decision of the Consensus Group.

CHAPTER VI

DECIDE-AI: Consensus meeting and final checklist

This chapter is adapted from:

Vasey, B. et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat. Med. (2021)[105].

-

Vasey B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. (2022)[322].

-

Vasey B. et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ (2022)[323].

VI.1 Introduction

The Delphi process provided quantitative feedback on the proposed DECIDE-AI items and produced numerous comments. However, the item list was still too exhaustive to be practical and it remained unclear which comments should be reflected in the guideline, and how. One solution would have been for the research team to shorten the list and make these decisions on their own. This would likely have reduced the credibility of the final checklist and its acceptance. Rather, the research team chose to organise a consensus meeting, involving a larger number of experts to address the issue of item selection and wording amendments. This approach is commonly used to conclude consensus processes, although it represents a modification of the original Delphi method[334].

Whereas it seems logical that the diversity of a consensus group is an important factor to represent the whole range of perspectives relevant to the recommendations under development, there is no widely accepted rule to determine the ideal size of such a group. SPIRIT-AI and CONSORT-AI had 31 members in their consensus group[167,168], the IDEAL Collaboration invited 14 participants to the final consensus teleconference of its 2019 guideline revision[187], TRIPOD had 24 experts attending the consensus meeting, STROBE 23 experts and the BOGUS statement[343] only one[e]. Although a direct equivalence with a consensus group could not be established in all cases, a systematic review on 37 published reporting guidelines estimated at seven (range 3-24) the core group responsible for the writing of the retrieved guidelines[344].

Another issue pertained to the implementation of the checklist. Even the best thought-through guideline would be of little impact if not used by the research community and, in the case of reporting guidelines, implemented by authors when writing up their manuscripts. Adoption by the community can of course not be guaranteed a priori, but several factors can be optimised

[e] This last reference is from the BMJ Christmas edition and should be appreciated for its ingenious reflection on academic guideline development rather than its scientific value.

to increase its chances of occurring. First and foremost, strengthening the guideline legitimacy through robust development methodology and a large base of participating experts representing diverse stakeholder groups. Then, assessing the applicability of the checklist in various scenarios. It is unrealistic to expect that a checklist would be fully applicable to all imaginable cases, but clear implementation issues should be identified and corrected prior to release. Finally making sure that the item wording, rationale for them and implementation instructions are clear for a variety of potential users, with different training backgrounds. While the first aspect was addressed through the design of the Delphi process reported in the previous chapter and the organisation of a consensus meeting, the applicability and clarity of the checklist and accompanying E&E sections remained to be evaluated. In order to address these two latter issues, a piloting phase is recommended by the EQUATOR network[331], although not always performed.

Therefore, the objectives of the consensus meeting and subsequent piloting phase were twofold. First to select the final set of reporting items for the DECIDE-AI checklist and their exact wording. Second, to conduct a rapid qualitative evaluation of the checklist before its publication, focusing on applicability and clarity.

VI.2 Methods

VI.2.1 Consensus meeting

Participant recruitment

Participants for the consensus meeting were selected to represent most of the key Delphi stakeholder groups and address geographic as well as demographic diversity. An additional constraint was to maintain a group size compatible with productive discussion considering the online format of the meeting. Invitations were sent per email and informed consent was obtained from all participants.

Preparatory activities

To reduce the online meeting time, both in order to accommodate various time zones and to minimise online meeting fatigue, preparatory activities were conducted. A first informal and optional plenary meeting was organised three weeks prior to the consensus meeting, on May 24th 2021. The aims of this meeting were: i) to introduce the members of the Consensus Group and the Chair of the consensus meeting, ii) to present the format of the preparatory exchanges and online consensus meeting, and iii) to present the results of the second round of Delphi.

Following this meeting, the 53 items from the revised list and the per item summary of the second round of Delphi were circulated to the Consensus Group to inform the preparatory exchanges. The aim of the preparatory exchanges was to obtain the initial position of the Consensus Group's members on the different items. This would allow to move directly to the discussion of contentious points during the virtual meeting. Based on the second round of Delphi's results, items were divided into three categories, with different instructions:

- High consensus items: defined as item for which ≥ 80% of the participants scored 7 or above AND no more than 2 stakeholder groups had an IQR25 <7; items will likely be included, unless a Consensus Group's member actively arguments against it.

- Low consensus items: defined as item for which <60% of the participants scored 7 or above; items will likely be excluded, unless a Consensus Group's member actively arguments for it.

- Unclear consensus items: the remaining items; these items will constitute the main focus of the consensus meeting and initial positions should be shared with the other Consensus Group members.

Online meetings

The online consensus meeting consisted of three virtual meetings of three hours each, conducted through Microsoft Teams. The meetings were hold on June 14th, 15th, and 16th 2021 between 13:00 and 16:00 British Summer Time. The chat function of Microsoft Teams was activated and participants encouraged to use it in parallel to the main discussion. The meeting was chaired by an academic with long experience in consensus discussion, independent from the Consensus Group. The meeting was attended by two additional observers. The first observer was responsible for taking meeting minutes and the second was responsible for moderating the chat function if needed. The second observer was also the contact person for the patient representative in case they had any technical question during the discussion. All 53 items of the revised list were discussed during the consensus meeting and a vote was conducted to determine inclusion or exclusion for each of them.

In order to highlight the novelty of the DECIDE-AI guideline, it was decided to create two lists of items: an AI-specific item list, comprising items specific to AI evaluation, and a generic item list, comprising general good research practice item, which obtained a high level of consensus during the Delphi process, and could be applied to any type of study.

Voting procedure and vote analysis

Voting took place through the online voting platform Vevox (www.vevox.com). For each item, the Chair decided on the final wording, which was then entered into the voting system. Participants first voted on the inclusion of the item (include, exclude, blank). For inclusion, an

item needed to reach a pre-defined threshold of 80% of the votes (excluding blank votes). Abstention or temporary absence from the meeting were considered as blank votes. If an item was included, the Consensus Group then voted to decide in which list (AI-specific or generic) the item should be included. A simple majority (excluding blank votes) was required for this second vote. Votes were anonymous and the Consensus Group members could only see the results of the vote after the closing of the procedure.

Consensus meeting follow up

Following the consensus meeting, the research team resolved any outstanding issue about item wording or list attribution. Explanation and Elaboration (E&E) paragraphs were developed by the research team for each of the included item. A glossary of terms was also created to clarify key concepts. All modifications to item wording were reviewed and validated by the Consensus Group. The E&E document was circulated to the Consensus and Steering Groups for comments and edits.

VI.2.2 Guideline piloting

Invitation

Potential experts for the piloting phase were contacted per email and invited to fill in a specially designed form within a two-week window period. They were required to have previous experience in either publication or peer-reviewing of literature describing the development and implementation of AI systems, or human factors experience with medical devices. Experts were either identified from the Delphi participants, from recommendations of the Consensus Group, and/or from relevant literature[151]. Attention was given to geographical diversity.

Questionnaire design

A piloting form was designed to collect expert feedback. Each item of the AI-specific item list was presented and, for each item, experts were asked to answer two binary (yes/no) questions on the applicability of the item and clarity of the corresponding E&E paragraph. In case of a negative answer, or if they otherwise wished so, the experts were invited to comment on their answer. Two questions regarding the generic item list were asked at the end of the form: the first regarding the division between the AI-specific and generic item list, the second inviting the expert to submit any comment they might have of these items and their corresponding E&E paragraphs. The piloting form can be found in **Annex VI-1**.

Analysis of results

Binary answers were aggregated. No pre-specified cut off was defined, although the research team considered an item potentially problematic if more than a third of the participants responded negatively to one of the binary questions. For each item, comments were collated and independently reviewed by two reviewers. Each reviewer proposed a list of modifications (either to the item itself or to the E&E paragraph) and the effective modifications were decided in consensus. All modification to item wording were reviewed and validated by the Consensus Group via email exchanges.

VI.2.3 DECIDE-AI logo

In anticipation of the guideline distribution and to create a visual signature of the guideline, a DECIDE-AI logo was created. A graphic designer (Jeanne Constantin, Freelance Graphic Designer, 87 Dalyell Road, SW9 9UR, London) was commissioned and, through three rounds of iterative design, created the DECIDE-AI logo. The key themes of the logo were the human-AI interaction and the concept of checklist. The authors of SPIRIT-AI and CONSORT-AI offered to share the AI checklist visual developed for their logo with the purpose of reinforcing the link between AI reporting guidelines.

VI.3 Results

VI.3.1 Participant characteristics

Consensus Group

Table VI-1 and **Table VI-2** show the characteristics of the Consensus Group members. The 16 members of the Consensus Group worked in five different countries and represented 16 of the Delphi stakeholder groups. One invited expert declined the invitation, due to conflicting commitment and workload, without affecting stakeholder representation. The names and affiliations of the Consensus Group participants can be found in **Suppl. Note VI-1**.

Country	Number of participants	Country	Number of participants	Country	Number of participants
United Kingdom (UK)	9 (56%)	The Netherlands (NL)	2 (13%)	Singapore (SG)	1 (6%)
United States of America (USA)	3 (19%)	Canada (CA)	1 (6%)		

Table VI-1: geographical origin of the Consensus Group members, as self-reported and based on their main working place. Total number of participants = 16.

Stakeholder group	Number of participants	Stakeholder group	Number of participants
Clinicians	8 (50%)	Policy makers/official institutions staff	2 (13%)
Engineers/Computer scientists	4 (25%)	Administrators/hospital management	2 (13%)
Methodologists	6 (38%)	Regulators	1 (6%)
Statisticians	2 (13%)	Trialists	1 (6%)
Implementation scientists	2 (13%)	Ethicists	1 (6%)
Entrepreneurs	3 (19%)	Patient representatives	1 (6%)
Epidemiologists	0 (0%)	Private sector representatives	1 (6%)
Human factors specialists	2 (13%)	Funders	2 (13%)
Journal editors	1 (6%)	Payers/Commissioners	0 (0%)
Allied health professional	0 (0%)	Psychologists	0 (0%)

Table VI-2: stakeholder group affiliation of the Consensus Group members. Each member could be affiliated to more than one stakeholder group. Total number of participants = 16.

Piloting

16 experts, independent from the Consensus Group and working in seven different countries, agreed to pilot the guideline, with at least two experts in each of the categories described under section VI.2.2. **Table VI-3** shows the geographical distribution of the experts in the piloting phase. The names and affiliations of the experts who took part in the guideline piloting can be found in **Suppl. Note VI-2**.

Country	Number of participants	Country	Number of participants	Country	Number of participants
United Kingdom (UK)	7 (44%)	Brazil (BR)	1 (6%)	The United States of America (USA)	1 (6%)
Singapore (SG)	3 (19%)	France (FR)	1 (6%)		
The Netherlands (NL)	2 (13%)	Germany (DE)	1 (6%)		

Table VI-3: geographical origin of the experts in the piloting phase. Total number of experts = 16.

VI.3.2 Consensus meeting discussion and voting results

The meeting minutes of the consensus Group discussion are presented in **Annex VI-2**. On five occasions, Consensus Group's members suggested merging between two and three items before proceeding to a vote. Such mergers were decided by consensus. No items were split during the consensus meeting. **Table VI-4** shows the results of the Consensus meeting votes for each item of the revised item list. The Consensus Group excluded nine Delphi items and included 38 Delphi items in the final guideline, of which 22 were unanimously voted in. 28 Delphi items were organised into 18 main reporting items with subitems in the AI-specific list and 10 Delphi items were kept as 10 main reporting items in the generic list. The detailed results on the list attribution votes are presented in **Suppl. Table VI-1**. The consensus item list can be found in **Suppl. Table VI-2** and **Suppl. Table VI-3**. Section VI.3.5 provides a summary of the main modifications to the item list made during the consensus meeting (including follow up work) and after the piloting phase.

Item n°	Vote include	Vote exclude	Blank vote	% vote include*	Decision	Item n°	Vote include	Vote exclude	Blank vote	% vote include*	Decision
Item 1a	15	0	1	100	include	Item 16	12	1	3	92	include
Item 1b	13	0	3	100	include	Item 17a	12	1	3	92	include
Item 2	14	0	2	100	include	Item 17b	13	0	3	100	include
Item 3	14	0	2	100	include	Item 18a	13	1	2	93	include
Item 4	9	6	1	60	exclude	Item 18b	11	2	3	85	include
Item 5	12	2	2	86	include	Item 19	13	0	3	100	include
Item 6a	12	2	2	86	include	Item 20a+c	12	0	4	100	include
Item 6b	1	14	1	7	exclude	Item 20b	3	9	4	25	exclude
Item 7	4	11	1	27	exclude	Item 21a+b	13	0	3	100	include
Item 8a	13	0	3	100	include	Item 21c+d+e	13	0	3	100	include
Item 8b	13	1	2	93	include	Item 22	12	1	3	92	include
Item 8c	13	0	3	100	include	Item 23a	13	0	3	100	include
Item 9	13	0	3	100	include	Item 23b	3	10	3	23	exclude
Item 10a	14	0	2	100	include	Item 23c	14	0	2	100	include
Item 10b+c	15	0	1	100	include	Item 23d	12	2	2	86	include
Item 10d	8	6	2	57	exclude	Item 23e	5	9	2	36	exclude
Item 10e	14	1	1	93	include	Item 24	14	0	2	100	include
Item 10f	15	0	1	100	include	Item 25	12	2	2	86	include
Item 11a	14	0	2	100	include	Item 26	10	3	3	77	exclude
Item 11b	13	1	2	93	include	Item 27	3	11	2	21	exclude
Item 12	14	0	2	100	include	Item 28	12	2	2	86	include
Item 13a+b	14	0	2	100	include	Item 29	13	1	2	93	include
Item 14	11	2	3	85	include	Item 30	14	0	2	100	include
Item 15	13	1	2	93	include						

Table VI-4: results of the Consensus Group votes. Some items were merged before proceeding to the vote. Total number of Consensus Group's members with voting right = 16. *excluding blank votes.

VI.3.3 Piloting phase results

The piloting phase yielded 14 full sets of answers (of which 5 were grouped in 2 surveys) and 2 partial sets of answers (comments on specific items without systematic answers to the questions about applicability and clarity of the E&E paragraphs). All sets were included in the analysis. **Annex VI-1** shows the full feedback collected from the participants. Following the piloting phase, two items were reorganised and the wording of 17 subitems was modified. **Annex VI-3** summarises the modifications made after the piloting phase.

VI.3.4 Overview of the item list evolution

Figure VI-1 presents an overview of the item list evolution from the initial development to the final checklist and **Figure VI-2** a detailed summary at item level (item keys available in **Figure VI-3**). Starting from 54 initial Delphi items, nine were added, 11 dropped and reorganisation of Delphi items reduced the total number by a further 14 Delphi items. Overall, this represents an attrition rate of 40%.

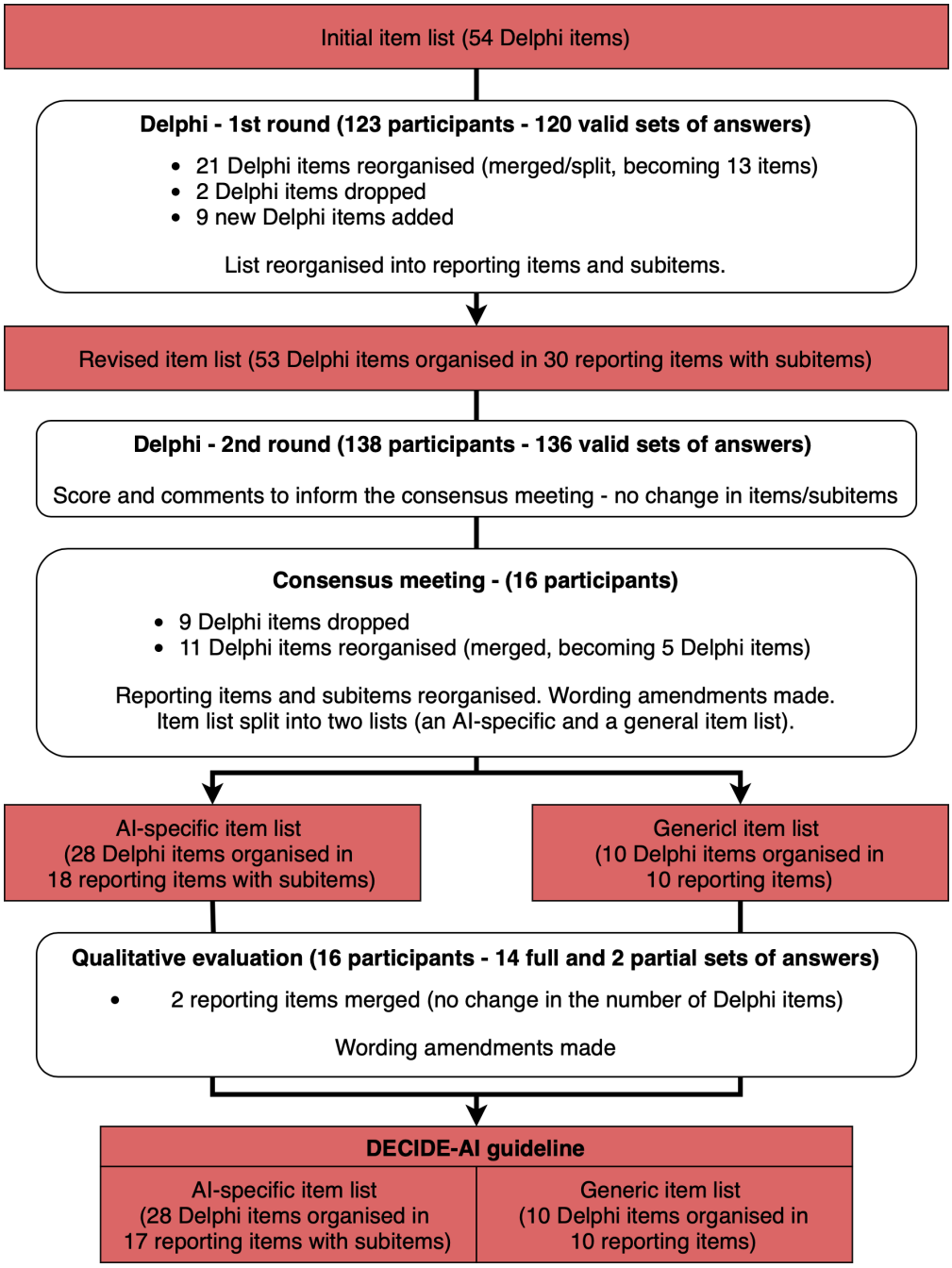


Figure VI-1: flowchart of the item list evolution. Adapted from *Vasey B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. (2022)*[322].

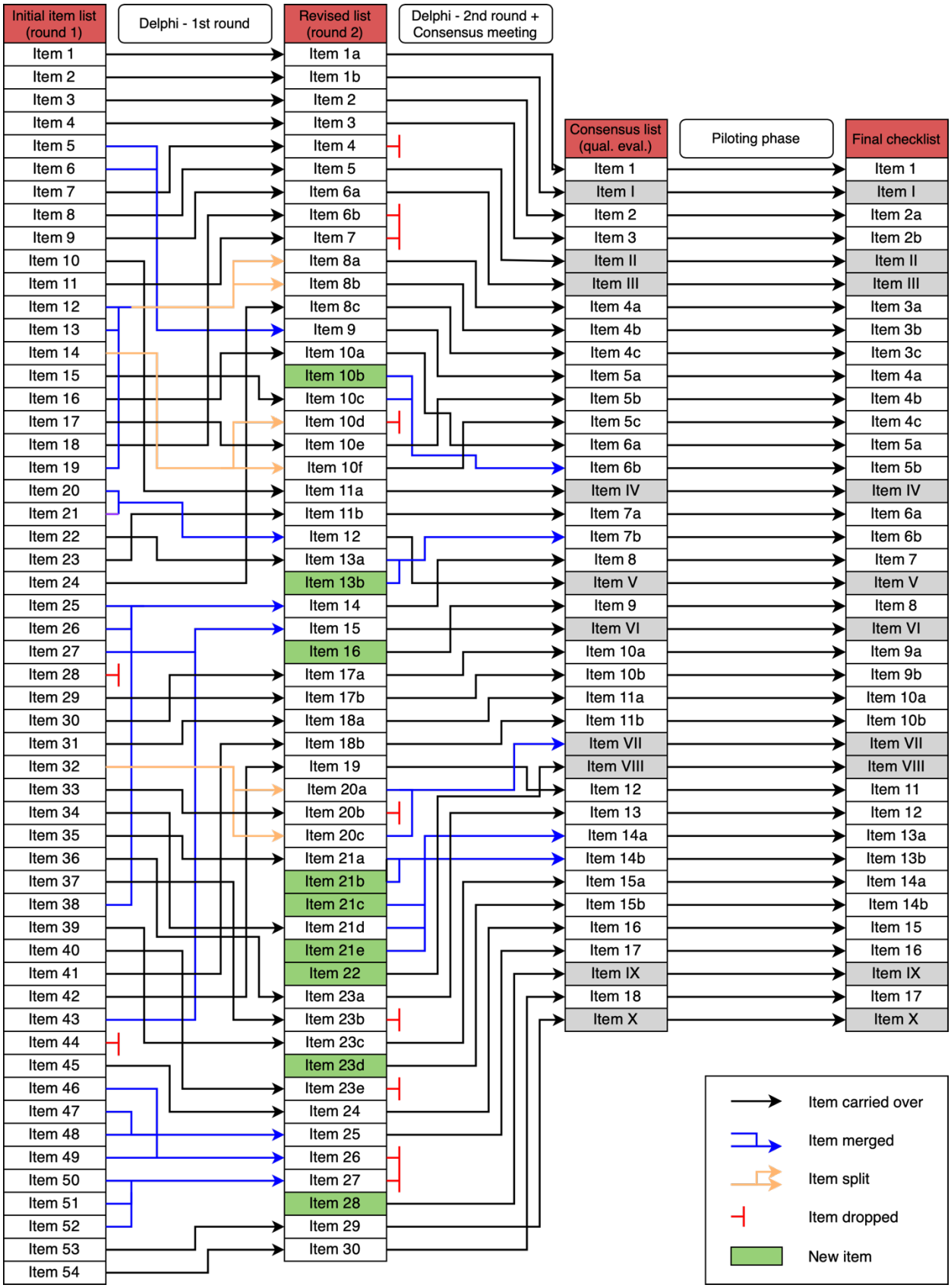


Figure VI-2: detailed graphical summary of the item list evolution. Adapted from *Vasey B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. (2022)*[322]. The item keys can be found in **Figure VI-3**. The detailed item content can be found under the following section: V.3.2 for the initial item list; V.3.6 for the revised item list; **Suppl. Table VI-2** and **Suppl. Table VI-3** for the consensus list; and section VI.3.6 for the final checklist. From the stage of the consensus meeting, the AI-specific items are numbered in Arab numerals and the generic items in Roman numerals.

	Initial list	Keys
Title & introduction	Item 1	Title
	Item 2	Abstract
	Item 3	Target condition
	Item 4	Intended use
	Item 5	Alg. dev. & validation
	Item 6	Training set
	Item 7	Current stage of dev.
	Item 8	Study objectives
Methods	Item 9	Study protocol
	Item 10	Outcomes
	Item 11	Study design
	Item 12	Participant selection
	Item 13	Sample size
	Item 14	Technical description
	Item 15	Implementation
	Item 16	Study environment
	Item 17	Data acquisition
	Item 18	Privacy & data security
	Item 19	Control group
	Item 20	Statistical analysis
	Item 21	Subgroup analysis
	Item 22	Safety requirements
	Item 23	Definition of errors
	Item 24	User training
	Item 25	Human factors
	Item 26	User acceptance
	Item 27	Patient involvement
	Item 28	Health econ. assess.
Results	Item 29	User population
	Item 30	Patient population
	Item 31	Implementation
	Item 32	User performance
	Item 33	Alg. performance
	Item 34	Errors
	Item 35	Compliance safety requ
	Item 36	User override
	Item 37	User trust
	Item 38	HF user description
	Item 39	HF usability
	Item 40	HF situa. awareness
	Item 41	HF others
	Item 42	Changes to alg.
	Item 43	Patient perspective
	Item 44	Health econ. assess.
Discussion & statements	Item 45	Performance
	Item 46	Deviation from recom.
	Item 47	Errors & biases
	Item 48	Safety profile
	Item 49	Human factors
	Item 50	User trust
	Item 51	Diff. betw. subgroups
	Item 52	Next steps
	Item 53	Funding & CoI
	Item 54	Data availability

	Revised list	Keys
Title & introduction	Item 1a	Title
	Item 1b	Abstract
	Item 2	Target clin. problem
	Item 3	Intended use
	Item 4	Current stage of dev.
	Item 5	Objectives
Methods	Item 6a	Study protocol & reg.
	Item 6b	Privacy & data security
	Item 7	Study design
	Item 8a	Patient recruitment
	Item 8b	User recruitment
	Item 8c	User training
	Item 9	Alg. description
	Item 10a	Study environment
	Item 10b	Clin. workflow/pathway
	Item 10c	Alg. use & timing
	Item 10d	Techn. implementation
	Item 10e	Input data
	Item 10f	Alg. outputs
	Item 11a	Outcomes
	Item 11b	Errors
	Item 12	Statistical analysis
	Item 13a	Safety requirements
	Item 13b	Unexpected risks
	Item 14	Human factors
	Item 15	Patient involvement
	Item 16	Ethics methodology
Results	Item 17a	Patient characteristics
	Item 17b	User characteristics
	Item 18a	Implementation
	Item 18b	Changes to workflows
	Item 19	Changes to alg.
	Item 20a	User performance
	Item 20b	Alg. performance
	Item 20c	Control group
	Item 21a	Compliance safety requ
	Item 21b	Unexpected risks
	Item 21c	Malfunctions
	Item 21d	Recommendation errors
	Item 21e	Human errors
	Item 22	Subgroup performance
	Item 23a	User-AI agreement
	Item 23b	User trust
	Item 23c	Usability
	Item 23d	Learning curves
	Item 23e	Interpretability
Discussion/statements	Item 24	Performance
	Item 25	Safety profile
	Item 26	Human factors
	Item 27	Scale up
	Item 28	Strengths & limitations
	Item 29	Funding & CoI
	Item 30	Data availability

	Consensus list	Keys	Final checklist
Title/introduction	Item 1	Title	Item 1
	Item I	Abstract	Item I
	Item 2	Target clin. problem	Item 2a
	Item 3	Intended use	Item 2b
	Item II	Objectives	Item II
Methods	Item III	Research governance	Item III
	Item 4a	Patient recruitment	Item 3a
	Item 4b	User recruitment	Item 3b
	Item 4c	User training	Item 3c
	Item 5a	AI system description	Item 4a
	Item 5b	Input data	Item 4b
	Item 5c	AI system output	Item 4c
	Item 6a	Study settings	Item 5a
	Item 6b	Clin. workflow/pathway	Item 5b
	Item IV	Outcomes	Item IV
	Item 7a	Errors identification	Item 6a
	Item 7b	Risk identification	Item 6b
	Item 8	Human factors	Item 7
	Item V	Analysis	Item V
	Item 9	Ethics methodology	Item 8
	Item VI	Patient involvement	Item VI
Results	Item 10a	Patient characteristics	Item 9a
	Item 10b	User characteristics	Item 9b
	Item 11a	Implementation	Item 10a
	Item 11b	Changes to workflows	Item 10b
	Item VII	Main results	Item VII
	Item VIII	Subgroups analysis	Item VIII
	Item 12	Modif. to AI system	Item 11
	Item 13	User-AI agreement	Item 12
	Item 14a	Errors & malfunctions	Item 13a
	Item 14b	Risks & harms	Item 13b
	Item 15a	Usability	Item 14a
	Item 15b	Learning curves	Item 14b
Discussion/stat.	Item 16	Performance	Item 15
	Item 17	Safety profile	Item 16
	Item IX	Strengths & limitations	Item IX
	Item 18	Data availability	Item 17
	Item X	CoI	Item X

Figure VI-3: item keys. Brief item description across the different stages of development for an easier interpretation of **Figure** . AI = artificial intelligence; Alg. = algorithm; assess. = assessment; betw. = between; CoI = conflicts of interest; dev. = development; diff. = difference; econ. = economic; modif. = modifications; recom. = recommendation; reg. = registration; requ. = requirements; stat. = statements.

VI.3.5 Main changes and discussions during the consensus meeting and piloting phase

The following paragraphs summarise by section the main changes made to the items and the key discussion during the consensus meeting and following the feedback received in the piloting phase. Item numbers refer to their number during the second round of Delphi and consensus meeting (revised list).

General

The word “algorithm” was replaced by “AI system” throughout the guideline to better differentiate between the underlying mathematical model responsible for the intelligent or learning aspects of a system (the algorithm) and the AI system itself, composed of the algorithm, supporting non-AI software and a hardware platform.

Title, abstract and introduction

Item 1b was completed to include a suggestion for the abstract content. Items 2 and 3 were merged and some of the wording modified to better align with the regulatory definition of a device’s intended purpose, which included a combination of: targeted medical condition and problem, intended patient population, intended users, use environment, and mode of action. Item 4 was excluded because, despite acknowledging that information about the regulatory context of a device could be useful for the readers, several Consensus Group members thought it was best to keep the focus of the guideline on scientific evaluation. No or little modifications were made to items 1a and 5.

Methods

Item 10b and 10c, as well as items 13a and 13b, were merged by the Consensus Group, in order to simplify the reporting of methodological aspects around AI system implementation and safety evaluation, respectively. Item 8a was completed with an instruction to describe

participant inclusion criteria both at patient and data level. The mention of control group in item 8b was discarded for conciseness. The Consensus Group mandated the rewriting of item 9, with a focus on the AI systems' description, including a clear version identifier, the population they were trained on, their preclinical performance, and reflecting the fact that these latter pieces of information would probably already have been described elsewhere. Item 10f was completed with a prompt to provide an image to illustrate the AI systems' output displays. Item 11b was modified to clarify that only significant errors and malfunctions should be described and reported. Item 10a was shortened after the piloting phase, based on the feedback that it was unrealistic to identify and report all additional clinical information (i.e. not provided by the AI systems) used by clinicians to orient their decision making.

Item 14 sparked an animated debate about the scope and size of the DECIDE-AI guideline (see section VI.4). It was eventually decided to include the item given the central role of human factors in the evaluation of new technology and in the rationale for the DECIDE-AI project. Items 15 and 16 were included despite a weaker Delphi participants' support in comparison to other items. The Consensus Group's rationale for item 15 was that an increasing number of journals and regulatory agencies are requiring information about patient and public involvement and that the inclusion of this item was important to recognise patients as key stakeholders in the future of clinical AI. The wording of item 15 was also clarified to ensure researchers report "how" and not only "whether" patients were involved. The group argued in favour of including item 16 stating that AI-related ethics methodology (and especially approaches to algorithmic fairness) are now well established in the literature and that a transparent reporting of these methodologies is essential to ensure a fair appraisal and comparison of the AI systems performance in an ethical context.

Item 6b was excluded as almost all Consensus Group members felt that issues around patient privacy and data security are the remits of research ethics committees and should be reported during the ethics application and approval process. Item 7 was deemed too generic in its wording and also excluded, with the justification that the most important aspects of study

design were anyway already covered in other items. Finally, item 10d was excluded after several members of the Consensus Group argued that details about the technical integration of an AI system would be better suited to an instruction manual or audit report than a scientific report. It was also argued that many research teams would probably not have integrated their AI system within the hospital infrastructure at the early stage of evaluation. No or little modifications were made to items 6a, 8c, 10e, 11a, 12.

Results

Item 20a and 20c were merged to better align outcomes reporting between the intervention group and a potential control group (if applicable). The mention of variation over time was discarded in the new merged wording, as the Consensus Group thought that this concept was already captured in item 19. Echoing the merger of item 13a and 13b, items 21a and 21b were merged to streamline the reporting of safety related results. Finally, items 21c, 21d, and 21e were also merged to unify the reporting of errors and malfunctions, while emphasising the need to understand their origins and potential impact on patient care.

Item 17a and 17b were modified to be less prescriptive on the characteristics to be reported. Furthermore, a clause on data missingness was added to item 17a. Item 18b was modified to reflect the difference between clinical workflow (i.e. the series of tasks performed by healthcare professionals in the exercise of their clinical duties) and care pathway (i.e. the series of interactions, investigations, decision-making and treatments experienced by patients in the course of their contact with a healthcare system), both being subject to modifications by the introduction of new technology. Item 23c was also reworded to accommodate for more diverse types of usability evaluation, while prescribing rigor in the process. Item 23d was shortened and the mention of user workload moved to the E&E paragraph of item 23c.

Item 20b was excluded with the rationale that collecting information about the stand-alone algorithm performance might not always be practical during clinical evaluation, depending on the modalities of the evaluated system. It was as well feared that an item focusing on stand-

alone performance could distract attention from the final, AI supported, human decisions, which are the ones actually influencing clinical outcomes. Item 23b and 23e were also excluded. The Consensus Group argued that, despite the importance of trust and interpretability to understand the reaction of users to AI systems' outputs, there are currently no commonly accepted ways to measure or quantify trust and interpretability in the context of clinical AI (see section VI.4). No or little modifications were made to items 18a, 19, 22, and 23a.

Discussion and statements

Item 25 was shortened and the part discussing underlying patterns or algorithmic bias was dropped, with the justification that the sample size of early-stage studies will most of the time not be powered to robustly detect such bias/pattern. Items 26 and 27 were excluded with the rationales that the discussion section should not be too prescriptive, that the results of the human factors evaluation should be discussed with those of the main clinical performance anyway, and that discussion about scale up might be too speculative at this stage. Moreover, scale up would not necessarily be the next logical step in many cases, such as for example, when researchers are interested in generalising the results to other settings first. No or little modifications were made to items 24 and 28, 29, and 30.

VI.3.6 The DECIDE-AI checklist

Table VI-5 and **Table VI-6** present the final AI-specific and generic item lists respectively. The DECIDE-AI final checklist can be found in **Annex VI-4**, the E&E document in **Annex VI-5**, and the glossary of terms in **Suppl. Table VI-4**.

Topic	Item nº	Recommendation
Title and abstract		
Title	1	Identify the study as early clinical evaluation of a decision support system based on artificial intelligence or machine learning, specifying the problem addressed.
Introduction		
Intended use	2	a) Describe the targeted medical condition(s) and problem(s), including the current standard practice, and the intended patient population(s).
		b) Describe the intended users of the AI system, its planned integration in the care pathway, and the potential impact, including patient outcomes, it is intended to have.
Methods		
Participants	3	a) Describe how patients were recruited, stating the inclusion and exclusion criteria at both patient and data level, and how the number of recruited patients was decided.
		b) Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided.
		c) Describe steps taken to familiarise the users with the AI system, including any training received prior to the study.
AI system	4	a) Briefly describe the AI system, specifying its version and type of underlying algorithm used. Describe, or provide a direct reference to, the characteristics of the patient population on which the algorithm was trained and its performance in preclinical development/validation studies.
		b) Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, the pre-processing applied and how missing/low-quality data were handled.
		c) Describe the AI system outputs and how they were presented to the users (an image may be useful).
Implementation	5	a) Describe the settings in which the AI system was evaluated.
		b) Describe the clinical workflow/care pathway in which the AI system was evaluated, the timing of its use, how the final supported decision was reached and by whom.
Safety and errors	6	a) Provide a description of how significant errors/malfunctions were defined and identified.
		b) Describe how any risks to patient safety or instances of harm were identified, analysed, and minimised.
Human factors	7	Describe the human factors tools, methods or frameworks used, the use cases considered, and the users involved.
Ethics	8	Describe whether specific methodologies were utilised to fulfil an ethics-related goal (such as algorithmic fairness) and their rationale.

Results		
Participants	9	a) Describe the baseline characteristics of the patients included in the study, and report on input data missingness.
		b) Describe the baseline characteristics of the users included in the study.
Implementation	10	a) Report on the user exposure to the AI system, on the number of instances the AI system was used, and on the users' adherence to the intended implementation.
		b) Report any significant changes to the clinical workflow or care pathway caused by the AI system.
Modifications	11	Report any changes made to the AI system or its hardware platform during the study. Report the timing of these modifications, the rationale for each, and any changes in outcomes observed after each of them.
Human-computer agreement	12	Report on the user agreement with the AI system. Describe any instances of and reasons for user variation from the AI system's recommendations and, if applicable, users changing their mind based on the AI system's recommendations.
Safety and errors	13	a) List any significant errors/malfunctions related to: AI system recommendations, supporting software/hardware, or users. Include details of: (i) rate of occurrence, (ii) apparent causes, (iii) whether they could be corrected, and (iv) any significant potential impacts on patient care.
		b) Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study.
Human factors	14	a) Report on the usability evaluation, according to recognised standards or frameworks.
		b) Report on the user learning curves evaluation.
Discussion		
Support for intended use	15	Discuss whether the results obtained support the intended use of the AI system in clinical settings.
Safety and errors	16	Discuss what the results indicate about the safety profile of the AI system. Discuss any observed errors/malfunctions and instances of harm, their implications for patient care and whether/how they can be mitigated.
Statements		
Data availability	17	Disclose if and how data and relevant code are available.

Table VI-5: AI-specific item list. To differentiate them from the generic items, AI-specific items are numbered in Arabic numerals. AI = Artificial Intelligence.

Topic	Item nº	Recommendation
Title and abstract		
Abstract	I	Provide a structured summary of the study. Consider including: intended use of the AI system, type of underlying algorithm, study setting, number of patients and users included, primary, secondary, safety and human factors outcomes measured, main results, conclusions.
Introduction		
Objectives	II	State the study objectives.
Methods		
Research governance	III	Provide a reference to any study protocol, study registration number, and ethics approval.
Outcomes	IV	Specify the primary and secondary outcomes measured.
Analysis	V	Describe the statistical methods by which the primary and secondary outcomes were analysed, as well as any prespecified additional analyses, including subgroup analyses and their rationale.
Patient Involvement	VI	State how patients were involved in any aspect of: the development of the research question, the study design, and the conduct of the study.
Results		
Main results	VII	Report on the prespecified outcomes, including outcomes for any comparison group if applicable.
Subgroups analysis	VIII	Report on the differences in the main outcomes according to the prespecified subgroups.
Discussion		
Strengths and limitations	IX	Discuss the strengths and limitations of the study.
Statements		
Conflicts of interest	X	Disclose any relevant conflicts of interest, including the source of funding for the study, the role of funders, any other roles played by commercial companies, and personal conflicts of interest for each author.

Table VI-6: generic item list. To differentiate them from the AI-specific items, generic items are numbered in Roman numerals. AI = Artificial Intelligence.

VI.3.7 The DECIDE-AI logo

Figure VI-4 presents the DECIDE-AI logo. Two intertwined hands, one rounded and one squared-shaped, represent the interaction between humans and AI, while the colours reinforce the hands' origins, with a blood-like, human red and a cold, mechanical blue.

Figure VI-4: the DECIDE-AI logo.

VI.4 Discussion

The DECIDE-AI checklist is the result of an international consensus process based on the Delphi methodology and following the EQUATOR Network recommendations for guideline development[331]. Emphasis was put on stakeholder diversity and giving the Delphi participants a high degree of influence on the guideline's development. The DECIDE-AI guideline was also developed to be applicable to all types of decision support modality (i.e. detection, diagnostic, prognostic and therapeutic), even though detection and diagnosis are currently predominant amongst decision support systems[63]. With its 38 Delphi items, organised in 27 main reporting items, the final version of the checklist reflects the growing acceptance that AI will, at least in the foreseeable future, not replace but augment human intelligence in clinical settings. In this context, thorough evaluation of the human-computer interaction and roles played by the human users will be key to translate AI from a promising technology to an integrant part of modern-day evidence-based medicine. In this regard, 18 items are related, directly or indirectly, to users and the implementation environment.

Out of 63 Delphi items (54 initial + nine added after the first round), 11 were dropped (25 if considering the reduction in number of Delphi items due to merging), an attrition rate of 17% (40% when considering mergers). This attrition rate is lower than, for example, CONSORT-AI (29/43, 67%), which can be partly explained by the fact that CONSORT-AI was an extension of CONSORT[193,194] and several items were excluded with the justification that they were already covered by the original guideline. When considering only the items excluded for other reasons, the attrition rate of CONSORT-AI becomes similar to DECIDE-AI at 37% (16/43). The DECIDE-AI attrition rate is also lower than the TRIPOD (39/76, 51%) and STARD 2003 (50/75, 67%) attrition rates[170,345].

To the best of my knowledge, the DECIDE-AI guideline is the first stage-specific AI reporting guideline to be developed, echoing recognised good practice in drug development and surgical innovation, as well as proposed approached for clinical AI[149–151,186–188]. From the beginning of the project, the Steering Group focused on the integration of DECIDE-AI within the broader landscape of AI guidelines, avoiding duplication of work or contradictory messages with other existing initiatives. Nonetheless, the introduction of a stage-specific guideline in a field dominated by study type-specific guidelines led to some resistance and needs for further clarifications. There are indeed more than one type of study design which would be appropriate for the early-stage of *live* clinical evaluation (e.g. prospective cohort studies, non-randomised controlled trials, …), for which reporting guidelines already exist, such as STROBE for cohort studies[346], or the CONSORT extension to randomised pilot or feasibility studies (also applicable to non-randomised pilot and feasibility studies when omitting certain items)[347,348]. However, none of the existing frameworks account for all the specificities of early-stage AI evaluation and important additional features such as modification of the intervention, prespecified subgroups analysis, or learning curve analysis, are left aside. Therefore, it was decided to develop DECIDE-AI around a backbone of generic items applicable to any study designs, enhanced by a set of AI-specific items, addressing the key issues of early *live* evaluation and building the real novelty of the guideline. This approach does not preclude the use of other study-type specific guidelines to complement the reporting when appropriate. The Template for Intervention Description and Replication (TIDieR) checklist[349] is another relevant reporting guideline, which can complement both stage and study-type specific checklists.

Several experts, members of the Consensus Group, and reviewers raised concerns about the scope of the DECIDE-AI guideline. They argued that the guideline prescribes an evaluation too exhaustive to be reported within a single manuscript. This concern was fully acknowledged by the Consensus Group and the research team. However, a majority within the Consensus Group agreed that appropriate AI evaluation is a complex endeavour necessitating the

interpretation of a wide range of data, some of which necessitating to be appraised in the context of the others, and should hence be presented together as far as possible. The use of references (the information required by several items might already be reported in previous studies or in the study protocol), online supplementary materials, and open-access repositories (e.g. Open Science Framework) was recommended to allow the sharing and connecting of all required information within one main published evaluation report. In addition, the point was made that publications reporting on AI systems might benefit from special formatting requirements in the future.

The majority of the Consensus Group also made clear that reporting guidelines aim to promote transparent reporting of studies, not to mandate the evaluation of every single aspect covered by their items. For example, if no specific ethics methodologies were used, then fulfilment of item 8 would be to simply state so, with an accompanying rationale. DECIDE-AI is indeed focused on scientific reporting, and as such do not intended to guide research conduct, although familiarity with the checklist might be useful to inform some aspects of the design and conduct of studies within the guideline's scope[350]. In this regard, it should also be clarified that adherence to the guideline is not to be interpreted in itself as an indication of methodological quality or enhanced robustness; two considerations which are the realms of methodological guidelines and risk of bias assessment tools. Reporting guidelines are designed for authors to report transparently the studies performed in order to allow readers to judge the robustness of the evaluation by themselves.

The Delphi process and subsequent consensus meeting highlighted the need for further development in specific areas of AI evaluation, namely a unified definition and nomenclature for development stages, as well as methodologies for the quantitative and qualitative assessment of user trust and AI system outputs interpretability. In the first case, the Consensus Group could not agree on a stage name from the several already proposed in the literature or inferred from other fields of medical research (e.g. formative evaluation stage, phase 2, stage 2, build/launch phase, step 2c/2d)[149–151,188]. The absence of clear guidance

about development stages' nomenclature is also detrimental to the exchange of expertise on evaluation methodology and can hinder fair comparison between studies. In the two last cases, the absence of an accepted way to measure trust and interpretability was the main argument to exclude the corresponding items from the guideline. It was agreed that, despite an overall interest in the topic and likely influence on the performance, these items could not be considered as minimum reporting standards, because researchers would not know how to evaluate these aspects in the first place.

On a related note, the Delphi process and consensus meeting corroborated the tendency of guideline development (and especially reporting guidelines) toward conservatism. Consensus decision-making inevitably prevents the recommendation of bolder, more controversial recommendations. Nonetheless, I believe that DECIDE-AI will challenge some of the status quo, by refocusing early-stage AI evaluation on the assessment of the AI systems' integration within the healthcare system, their interaction with human users, and by raising attention on the need for an evaluation tailored to each stage of development of an innovation. In doing so, DECIDE-AI will hopefully help generate new, or at least more, data on the key aspects of AI evaluation, and thereby participate to advance the field.

The final checklist should be considered as a set of minimum scientific reporting standards and do not preclude the sharing of additional information, nor are they a substitute for other regulatory reporting or approval requirements. The overlap between scientific evaluation and regulatory processes, as well as DECIDE-AI integration within the different regulatory ecosystems were taken very seriously during the guideline development process. For examples, the initial item list was aligned with information commonly required by regulatory agencies, regulatory agencies and regulatory experts were contacted early in the process and invited to participate in the Delphi process, a member of the UK's regulatory agency (MHRA) was part of the Consensus Group, and regulatory considerations are introduced in the E&E paragraphs. Early-stage scientific studies can indeed be used to inform regulatory decisions (e.g. risk classification based on the intended use stated in the study, or decision to regulate

an AI system as medical device at all, based on the degree of output interpretability[351]), and are part of the clinical evidence generation process (e.g. clinical investigations). However, given the differences between regulatory jurisdictions (e.g. FDA, MDR, MHRA, PMDA), and the sometimes different focuses of scientific evaluation and regulatory assessment[112], no reference to specific regulatory processes were made in the guideline, nor was the scope of DECIDE-AI defined within any particular regulatory framework. The primary focus of DECIDE-AI is scientific evaluation and reporting, for which regulatory documents often provide little guidance. Furthermore, when developed independently, scientific guidelines, and more specifically their discordance with regulatory practice, can be a useful influence for improving the latter.

VI.4.1 Strengths and limitations

The DECIDE-AI guideline was developed using a Delphi process, a recognised methodology, based on the feedback of a large number of participants and the in-depth discussion of a selected group of experts forming the Consensus Group. It introduces important novelties in the reporting of clinical AI. To the best of my knowledge, DECIDE-AI is the first AI or complex intervention reporting guideline to explicitly name human factors and ethical aspects (more specifically algorithmic fairness) as key reporting items. By placing the users, patients, and clinical environment at the centre of the evaluation process, it shifts the emphasis from algorithm performance to AI system and user performance. The DECIDE-AI guideline comes with an exhaustive E&E section, providing detailed justification for each item and information about their implementation. Both the checklist and E&E document were subjected to piloting in order to improve their quality and applicability. This piloting phase led to several clarifications of the E&E paragraphs and a clearer organisation of the reporting items in the introduction section.

The results of the consensus meeting and the final guideline should be appraised in the context of several limitations. First of all, the scope of the guideline is limited to decision support systems, or in other words human-in-the-loop systems, and their initial *live* evaluation. This excludes several important areas of AI systems clinical evaluation. For example, many AI systems will probably be used for mass screening purposes without necessarily depending on human supervision for referral decision[64,73]. Whereas some aspects will overlap, the evaluation of these systems will focus on stand-alone performance rather than AI-user interaction. Monitoring of AI system performance in time when routinely used in a clinical environment were also left out the DECIDE-AI scope. Despite being crucial in the evaluation process, especially due to concerns around dataset shift and performance drift[153,352], it was decided to not include items related to algorithmovigilance within the guideline. Algorithmovigilance was considered a separate phase of evaluation (see registries for pharmacovigilance and IDEAL stage 4[186–188]) and the evaluation and regulation of continuous learning algorithms is still actively debated[353,354]. *Shadow* or *silent* mode evaluations are also not covered, here again many aspects of DECIDE-AI would also be relevant, but the emphasis probably more on performance and offline user evaluation than on implementation and human factors in a context where user decisions have an actual impact on patient care.

Second, due to the COVID-19 pandemic, the Steering Group had little other options, but to organise a virtual consensus meeting. Despite being the most practicable option to finish the project within the timeframe of this thesis and having yielded the expected results, this online format had the three following effect on the consensus meeting: a reduction of the Consensus Group's size, a reduction in the number of hours allocated to the face-to-face discussion (nine in total), and a likely reduction in the overall dedication of the participants. Despite a "camera on" policy, individual absences during parts of the meeting could not be prevented. However, special attention was given to the Consensus Group selection in order to have most of the key stakeholders represented and there is no evidence that the size of consensus groups have a direct impact on the consensus quality. In the case of DECIDE-AI, a smaller group and skilled

chairing allowed for all participants who wished so to express their opinion and participate to the discussions preceding votes. Furthermore, actions were taken to counteract the effects of potentially reduced engagement opportunities. Preparatory exchanges and post-meeting communication were designed to offer participants additional platforms to debate the guideline content and item wording.

Third, consensus is no proof of correctness or optimal position. Even though the DECIDE-AI Steering Group based the content of the guideline on existing best practice in related fields and believes that DECIDE-AI is a step up on the current environment for reporting early stage live evaluation of AI-systems, the actual impact of the guideline on the quality of reporting remains an open question and will need to be further investigated. Evidence exists that consensus-based guidelines indeed improve the quality of reporting[355], and a piloting was conducted to evaluate if the selected items were theoretically applicable to real-world projects. Nonetheless, data on the implementation of the DECIDE-AI guideline, its added value, and shortcomings will need to be collected and inform future iteration of the checklist.

Fourth, given the relative low number of publications describing the early-stage *live* evaluation of AI-based decision support systems, satisfactory examples of good reporting could not be identified for all items. A similar issue was encountered by the authors of SPIRIT-AI/CONSORT-AI when developing their guidelines [personal communication, Dr Xiaoxuan Liu][167,168]. To avoid having to invent examples or prescribing reporting templates, and to maintain consistency between the E&E paragraphs, it was eventually decided not to provide examples at all. This limitation will hopefully be addressed in the next iteration of the guideline, once adoption will have generated more materials to draw the examples from.

Finally, despite the diversity of experts who took part in the piloting, no prospective application on an ongoing project took place. This would have been unrealistic in the available timescale, but would have allowed a qualitative evaluation closer to the actual implementation conditions.

It should also be noted that no protocol was made publicly available prior to the conduct of the study. Even if the study design was formalised in the research ethics committee application and the key analytical decisions (including item inclusion cut-offs) recorded in the Steering Group meeting minutes prior to data analysis, this information was not openly accessible to wider scrutiny. In a time of always increasing support to open science and given DECIDE-AI own recommendations toward more transparency, this information should have been made available either through publication of the protocol in a peer-reviewed journal, or at least on an open platform such as the Open Science Framework (OSF).

VI.4.2 Conclusion

In conclusion, the new DECIDE-AI guideline was developed to improve the reporting of early-stage *live* clinical evaluation of AI systems. Robust and comprehensive evidence generation and reporting in the early stages of clinical implementation are necessary, both to lay the foundations for larger (comparative) clinical studies and to build the trust of patients, practitioners, and purchasers in the evaluated AI systems. Only if developed and evaluated on strong grounds will this technology be accepted for an adoption at scale and enabled to realise its full potential to improve patient care.

CHAPTER VII

Future work

VII.1 Retrospective

This thesis is the results of an initial research project, later shaped and influenced by the scholars I had the privilege to meet over almost four years and external events which occurred during this period. These external circumstances provided constrains but also opportunities which have profoundly marked the final content of the thesis. Overall, and with different level of impact, I believe the presented work made several contributions to the field.

First, the chapter on clinician support needs provided, to the best of my knowledge, a first formalised analysis of the cognitively challenging steps encountered, barriers faced, and coping strategies used when managing surgical complications on the ward, produced with a recognised task analysis methodology. It also provided a list of desired computerised support modalities, which could inspire the development of future AI systems.

Second, the systematic review demonstrated the lack of robust evidence that ML-based CDSSs actually improve human performance in clinical settings. It also highlighted how little research has been conducted using AI-assisted clinician performance as outcome and how few of these studies included human factors in their investigations. It is my hope that these findings will foster a debate on the most appropriate outcome to measure clinical benefits of AI systems and influence the design of future AI studies.

Third, the algorithm development process raised several issues relevant to future work on AI systems, even if the algorithm showed mixed performance itself. Limitations at metadata level, the influence of certain drug categories in the overall prescription pattern, and discrepancies between mathematical and clinical optima offer useful considerations for future research.

Finally, the main contribution of the thesis is, in my opinion, the DECIDE-AI guideline. The guideline offers a novel perspective on clinical evaluation of AI systems, emphasising the importance of early-stage investigations, and practical guidance for the reporting of such studies. The content of the final chapters is the most likely to influence other research in the field and the area in which future work is planned, as described in the following sections.

VII.2 DECIDE-AI – Future work

VII.2.1 Short-term: guideline dissemination

The publication of a reporting guideline is only the first step towards improving research practice and does not mean in itself that the guideline will be used by the community, nor adequately championed by journals[356]. Therefore, a phase of dissemination is necessary after the initial publication. This dissemination can happen through different channels such as: publication of commentary pieces in relevant specialised journals, presentation of the guideline at conferences or seminars, and integration of the guideline in official guidance or recommendations.

Since the publication of DECIDE-AI on May 18th 2022, our group has been active in these areas, with already two commentary pieces published[107,357], two reviews published[358,359], one correspondence in writing and an additional commentary piece in project. Presentations were made at several events such as the NHS Digital Health Safety 2022 conference, the BrainX community June webinar, the Alan Turing Institute special interest group on clinical AI, and the IDEAL 2022 Meeting. We are also currently participating in two initiatives at European level where the lessons learned from DECIDE-AI could be useful to inform policy making.

Beyond the ongoing dissemination work, further efforts are still needed to increase the reach of the guideline and stimulate debate about its content. Additional presentations and publications would support these objectives, as well as developing an advisory platform to help researchers in the implementation of the guideline, hence reducing the barriers to adoption. Translation of the guideline into different languages is also being considered. Finally, discussion with journals specialising in digital health and clinical AI should be pursued in order to advocate for endorsement of the guideline at editorial level.

VII.2.2 Mid-term: guideline evaluation and evidence synthesis

As pointed out during the peer-review process, consensus is not equal to correctness and the current checklist will need to be evaluated now that in use. Several partners from both the private and academic sectors have already been approached and showed interest in the prospect of a collaboration. Such evaluation would be mainly qualitative in nature in a first stage, whereas quantitative evaluation of the guideline impact on reporting could be considered at a later stage[360]. The mid-term qualitative evaluation is planned in the form of feedback collection amongst researchers using the checklist when writing articles for submission. Narrative comments on the perceived usefulness or problems of the guideline would be collected and systematically analysed using a dedicated thematic analysis approach. Additional in-depth interviews of selected participants would also be considered after intermediary analysis of the feedback. In addition, manuscripts analysis may be conducted before and after the use of the guideline in order to identify key components initially omitted but later added thanks to the guideline as well as key components present in the manuscript but not covered in the guideline.

Another mid-term project relevant to the future of DECIDE-AI is the undertaking of knowledge synthesis on key areas of AI evaluation which were considered for inclusion but eventually not selected for the final checklist, such as human trust in AI and model interpretability. As mentioned in the publication, the non-inclusion of these topics reflects more a current lack of evidence than a lack of interest. Indeed, very few studies have rigorously investigated these topics in practice[101] and there is no overview of the evidence generated so far. We intend to start a literature review on human-AI trust evaluation in clinical practice. Given the expected heterogeneity of the target studies, the review will likely take the form of a scoping review with narrative synthesis. The main focuses of the review will include: the impact of trust on performance when used in actual clinical settings, the dynamic changes in the level of trust in AI over time (trust curve), and the identification of factors impacting trust in AI.

Following up on the evidence synthesis phase, we plan to conduct our own study on the influence of trust on AI systems performance. The exact design of such a study will depend on the most interesting questions identified during the scoping review. We will aim at collaborating with groups implementing AI system in live clinical settings and have already identified potential partners for this work. If the expected gap in knowledge is confirmed through the review of the literature, a well-designed study generating evidence on human-AI trust in the clinical context would bring valuable new evidence not only for the future development of the guideline but also to guide the development of AI-based decision support in practice.

VII.2.3 Long-term: guideline update and development of a standardised evaluation pathway

Within a five to ten-year horizon, work on the next iteration of DECIDE-AI will be started. This is not only common practice for major guidelines[172,187,194,350], but will also be necessary to remain up-to-date in the rapidly evolving field of clinical AI. Results from the qualitative evaluation and the evidence synthesis review will feed into the proposed updates and a new Delphi process will likely be organised, with invitations sent to all authors who have used the guideline in practice as well as the original collaborators. Special attention will be given to geographical diversity, with early efforts to engage with researchers from areas underrepresented during the initial DECIDE-AI Delphi. This next iteration will be the occasion to include new developments in the field into the checklist, but also to trim it from items which have become obsolete or which may limit adoption of the guideline by being too burdensome. An update of the Explanation & Elaboration (E&E) section will also be carried out. We hope that there will by this time be enough high-quality publications to add examples of good reporting to the E&E sections.

Over the same timeline, the next major initiative in terms of AI development evaluation guidance would be to propose a standardised evaluation pathway, analogous to that for drug development, synthesising scientific and regulatory requirements for a staged evaluation from *in silico* testing to bedside effectiveness. An interesting challenge for the later stages of this kind of pathway will be how to develop simultaneous continuous surveillance of the changes in the AI algorithm and interface, it's use profile (i.e. the changing patient population it is used in) and clinical outcomes associated with its use, to ensure that its value remains stable or improves over time.

Such an endeavour is of course beyond the reach of a single research group and would need a coordinated effort from key institutions and recognised experts in the field of clinical AI. Developing an integrated and standardised evaluation pathway would also require the remaining methodological and regulatory challenges of clinical AI evaluation to be identified and addressed. Filling this evidence gap could add years to the process and would need coordinated research workflows across institutions. We are currently active in two international consortium projects whose objectives are to provide policy makers with updated evidence on clinical evaluation of AI and to develop proposals for an evaluation framework. We hope to contribute to these collaborations with the experience and knowledge acquired during the present thesis and to continue demonstrating the crucial role of early-stage clinical evaluation in the safe and efficient translation of clinical AI into healthcare practice.

LIST OF REFERENCES

1. Khuri, S. F. *et al.* Determinants of long-term survival after major surgery and the adverse effect of postoperative complications. *Ann. Surg.* 242, 326–343 (2005).

2. Tevis, S. E. & Kennedy, G. D. Postoperative complications and implications on patient-centered outcomes. *The Journal of surgical research* vol. 181 106–113 (2013).

3. The international Surgical Outcomes Study Group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. *Br. J. Anaesth.* 117, 601–609 (2016).

4. Pinto, A., Faiz, O., Davis, R., Almoudaris, A. & Vincent, C. Surgical complications and their impact on patients' psychosocial well-being: a systematic review and meta-analysis. *BMJ Open* 6, e007224 (2016).

5. Tevis, S. E., Cobian, A. G., Truong, H. P., Craven, M. W. & Kennedy, G. D. Implications of Multiple Complications on the Postoperative Recovery of General Surgery Patients. *Ann. Surg.* 263, (2016).

6. Gawande, A. A., Thomas, E. J., Zinner, M. J. & Brennan, T. A. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery* 126, 66–75 (1999).

7. Haynes, A. B. *et al.* A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population. *N. Engl. J. Med.* 360, 491–499 (2009).

8. Kable, A. K., Gibberd, R. W. & Spigelman, A. D. Adverse events in surgical patients in Australia. *Int. J. Qual. Heal. care J. Int. Soc. Qual. Heal. Care* 14, 269–276 (2002).

9. Rosen, A. K., Geraci, J. M., Ash, A. S., McNiff, K. J. & Moskowitz, M. A. Postoperative adverse events of common surgical procedures in the Medicare population. *Med. Care* 30, 753–765 (1992).

10. Dencker, E. E., Bonde, A., Troelsen, A., Varadarajan, K. M. & Sillesen, M. Postoperative complications: an observational study of trends in the United States from 2012 to 2018. *BMC Surg.* 21, 393 (2021).

11. Strasberg, S. M., Linehan, D. C. & Hawkins, W. G. The accordion severity grading system of surgical complications. *Ann. Surg.* 250, 177–186 (2009).

12. Clavien, P. A., Sanabria, J. R. & Strasberg, S. M. Proposed classification of complications of surgery with examples of utility in cholecystectomy. *Surgery* 111, 518–526 (1992).

13. Dindo, D., Demartines, N. & Clavien, P.-A. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann. Surg.* 240, 205–213 (2004).

14. Clavien, P. A. *et al.* The Clavien-Dindo classification of surgical complications: five-year experience. *Ann. Surg.* 250, 187–196 (2009).

15. Ko, C. Y., Hall, B. L., Hart, A. J., Cohen, M. E. & Hoyt, D. B. The American College of Surgeons National Surgical Quality Improvement Program: achieving better and safer surgery. *Jt. Comm. J. Qual. patient Saf.* 41, 199–204 (2015).

16. Silber, J. H., Williams, S. V, Krakauer, H. & Schwartz, J. S. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. *Med. Care* 30, 615–629 (1992).

17. Ghaferi, A. A., Birkmeyer, J. D. & Dimick, J. B. Variation in Hospital Mortality Associated with Inpatient Surgery. *N. Engl. J. Med.* 361, 1368–1375 (2009).

18. Ghaferi, A. A., Birkmeyer, J. D. & Dimick, J. B. Complications, failure to rescue, and mortality with major inpatient surgery in medicare patients. *Ann. Surg.* 250, 1029–1034 (2009).

19. Ghaferi, A. A., Birkmeyer, J. D. & Dimick, J. B. Hospital volume and failure to rescue with high-risk surgery. *Med. Care* 49, 1076–1081 (2011).

20. Brooke, B. S. *et al.* Variations in surgical outcomes associated with hospital compliance with safety practices. *Surgery* 151, 651–659 (2012).

21. Kendall-Gallagher, D., Aiken, L. H., Sloane, D. M. & Cimiotti, J. P. Nurse specialty certification, inpatient mortality, and failure to rescue. *J. Nurs. Scholarsh. an Off. Publ. Sigma Theta Tau Int. Honor Soc. Nurs.* 43, 188–194 (2011).

22. Ludikhuize, J., de Jonge, E. & Goossens, A. Measuring adherence among nurses one year after training in applying the Modified Early Warning Score and Situation-Background-Assessment-Recommendation instruments. *Resuscitation* 82, 1428–1433 (2011).

23. Yasunaga, H., Hashimoto, H., Horiguchi, H., Miyata, H. & Matsuda, S. Variation in cancer surgical outcomes associated with physician and nurse staffing: a retrospective observational study using the Japanese Diagnosis Procedure Combination Database. *BMC Health Serv. Res.* 12, 129 (2012).

24. Ghaferi, A. A., Osborne, N. H., Birkmeyer, J. D. & Dimick, J. B. Hospital Characteristics Associated with Failure to Rescue from Complications after Pancreatectomy. *J. Am. Coll. Surg.* 211, 325–330 (2010).

25. Ward, S. T., Dimick, J. B., Zhang, W., Campbell, D. A. & Ghaferi, A. A. Association Between Hospital Staffing Models and Failure to Rescue. *Ann. Surg.* 270, 91–94 (2019).

26. Johnston, M. J. *et al.* A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. *Surgery* 157, 752–763 (2015).

27. Calzavacca, P. *et al.* A prospective study of factors influencing the outcome of patients after a Medical Emergency Team review. *Intensive Care Med.* 34, 2112 (2008).

28. Quach, J. L. *et al.* Characteristics and outcomes of patients receiving a medical emergency team review for respiratory distress or hypotension. *J. Crit. Care* 23, 325–331 (2008).

29. Downey, A. W. *et al.* Characteristics and outcomes of patients receiving a medical emergency team review for acute change in conscious state or arrhythmias. *Crit. Care Med.* 36, 477–481 (2008).

30. Barwise, A. *et al.* Delayed Rapid Response Team Activation Is Associated With Increased Hospital Mortality, Morbidity, and Length of Stay in a Tertiary Care Institution. *Crit. Care Med.* 44, 54–63 (2016).

31. Liu, V., Kipnis, P., Rizk, N. W. & Escobar, G. J. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. *J Hosp Med* 7, (2012).

32. van Galen, L. S. *et al.* Delayed Recognition of Deterioration of Patients in General Wards Is Mostly Caused by Human Related Monitoring Failures: A Root Cause Analysis of Unplanned ICU Admissions. *PLoS One* 11, e0161393–e0161393 (2016).

33. Gerry, S. *et al.* Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 369, m1501 (2020).

34. Royal College of Physicians. *National Early Warning Score (NEWS): standardising the assessment of acute-illness severity in the NHS*. (2012).

35. Royal College of Physicians. *National Early Warning Score (NEWS) 2 - Standardising the assessment of acute-illness severity in the NHS*. (2017).

36. Pimentel, M. A. F. *et al.* Detecting Deteriorating Patients in Hospital: Development and Validation of a Novel Scoring System. *Am. J. Respir. Crit. Care Med.* (2021).

37. Pimentel, M. A. F., Clifton, D. A., Clifton, L., Watkinson, P. J. & Tarassenko, L. Modelling physiological deterioration in post-operative patient vital-sign data. *Med. Biol. Eng. Comput.* 51, 869–877 (2013).

38. Pimentel, M. A. F. *et al.* A comparison of the ability of the National Early Warning Score and the National Early Warning Score 2 to identify patients at risk of in-hospital mortality: A multi-centre database study. *Resuscitation* 134, 147–156 (2019).

39. Bedoya, A. D. *et al.* Minimal Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration. *Crit. Care Med.* 47, 49–55 (2019).

40. Cioffi, J., Salter, C., Wilkes, L., Vonu-Boriceanu, O. & Scott, J. Clinicians' responses to abnormal vital signs in an emergency department. *Aust. Crit. Care* 19, 66–72 (2006).

41. Peebles, E., Subbe, C. P., Hughes, P. & Gemmell, L. Timing and teamwork—An observational pilot study of patients referred to a Rapid Response Team with the aim of identifying factors amenable to re-design of a Rapid Response System. *Resuscitation* 83, 782–787 (2012).

42. Donohue, L. A. & Endacott, R. Track, trigger and teamwork: Communication of deterioration in acute medical and surgical wards. *Intensive Crit. Care Nurs.* 26, 10–17 (2010).

43. O'Neill, S. M. *et al.* Why do healthcare professionals fail to escalate as per the early warning system (EWS) protocol? A qualitative evidence synthesis of the barriers and facilitators of escalation. *BMC Emerg. Med.* 21, 15 (2021).

44. Ede, J. *et al.* Human factors in escalating acute ward care: a qualitative evidence synthesis. *BMJ open Qual.* 10, (2021).

45. Gawronski, O. *et al.* Qualitative study exploring factors influencing escalation of care of deteriorating children in a children's hospital. *BMJ Paediatr. open* 2, e000241 (2018).

46. Buist, M. D. *et al.* Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. *BMJ* 324, 387–390 (2002).

47. Hillman, K. Introduction of the medical emergency team (MET) system: A cluster-randomised controlled trial. *Lancet* 365, 2091–2097 (2005).

48. Robertson, E. R., Morgan, L., Bird, S., Catchpole, K. & McCulloch, P. Interventions employed to improve intrahospital handover: a systematic review. *BMJ Qual. Saf.* 23, 600–607 (2014).

49. McCulloch, P., Rathbone, J. & Catchpole, K. Interventions to improve teamwork and communications among healthcare staff. *Br. J. Surg.* 98, 469–479 (2011).

50. NASH, F. A. Differential diagnosis, an apparatus to assist the logical faculties. *Lancet (London, England)* 266, 874–875 (1954).

51. Lipkin, M. & Hardy, J. D. Mechanical Correlation of data in differential diagnosis of hematological diseases. *J. Am. Med. Assoc.* 166, 113–125 (1958).

52. Ledley, R. S. & Lusted, L. B. Reasoning Foundations of Medical Diagnosis. *Science (80-.).* 130, 9 LP – 21 (1959).

53. Jaspers, M. W. M., Smeulers, M., Vermeulen, H. & Peute, L. W. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J. Am. Med. Inform. Assoc.* 18, 327–334 (2011).

54. Riches, N. *et al.* The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: A Systematic Review and Meta-Analysis. *PLoS One* 11, 1–26 (2016).

55. Bright, T. J. *et al.* Effect of clinical decision-support systems: A systematic review. *Ann. Intern. Med.* 157, 29–43 (2012).

56. Garg, A. X. *et al.* Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 293, 1223–1238 (2005).

57. Kwan, J. L. *et al.* Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ* 370, m3216 (2020).

58. Poole, D., Mackworth, A. & Goebel, R. *Computational Intelligence: A Logical Approach*. (1998).

59. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* 3, 210–229 (1959).

60. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Heal.* 1, e271–e297 (2019).

61. Handelman, G. S. *et al.* eDoctor: machine learning and the future of medicine. *J. Intern. Med.* 284, 603–619 (2018).

62. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* 25, 30–36 (2019).

63. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digit. Med.* 3, 118 (2020).

64. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digit. Med.* 1, 39 (2018).

65. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* 28, 31–38 (2022).

66. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56 (2019).

67. Shortliffe, E. H. & Sepúlveda, M. J. Clinical Decision Support in the Era of Artificial IntelligenceClinical Decision Support in the Era of Artificial IntelligenceClinical Decision Support in the Era of Artificial Intelligence. *JAMA* 320, 2199–2200 (2018).

68. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 160035 (2016).

69. Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* 5, 180178 (2018).

70. Willemink, M. J. *et al.* Preparing Medical Imaging Data for Machine Learning. *Radiology* 295, 4–15 (2020).

71. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510 (2018).

72. Wagner, M. *et al.* Surgomics: personalized prediction of morbidity, mortality and long-term outcome in surgery using machine learning on multimodal data. *Surg. Endosc.* 36, 8568–8591 (2022).

73. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115 (2017).

74. Ting, D. S. W. *et al.* Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 318, 2211–2223 (2017).

75. Ravindran, T. K. S., Teerawattananon, Y., Tannenbaum, C. & Vijayasingham, L. Making pharmaceutical research and regulation work for women. *BMJ* 371, m3808 (2020).

76. Steinberg, J. R. *et al.* Analysis of Female Enrollment and Participant Sex by Burden of Disease in US Clinical Trials Between 2000 and 2020. *JAMA Netw. Open* 4, e2113749–e2113749 (2021).

77. Avery, E. & Clark, J. Sex-related reporting in randomised controlled trials in medical journals. *Lancet* 388, 2839–2840 (2016).

78. Geller, S. E., Koch, A., Pellettieri, B. & Carnes, M. Inclusion, Analysis, and Reporting of Sex and Race/Ethnicity in Clinical Trials: Have We Made Progress? *J. Women's Heal.* 20, 315–320 (2011).

79. Ingelman-Sundberg, M., Oscarson, M. & McLellan, R. A. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *Trends Pharmacol. Sci.* 20, 342–349 (1999).

80. Roden, D. M., Wilke, R. A., Kroemer, H. K. & Stein, C. M. Pharmacogenomics: the genetics of variable drug responses. *Circulation* 123, 1661–1670 (2011).

81. Haga, S. B. & Burke, W. Using Pharmacogenetics to Improve Drug Safety and Efficacy. *JAMA* 291, 2869–2871 (2004).

82. Titano, J. J. *et al.* Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* 24, 1337–1341 (2018).

83. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* 110, 12–22 (2019).

84. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368, m689 (2020).

85. Vasey, B. *et al.* Association of Clinician Diagnostic Performance with Machine Learning-Based Decision Support Systems: A Systematic Review. *JAMA Netw. Open* 4, (2021).

86. Aggarwal, R. *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit. Med.* 4, 65 (2021).

87. Maddox, T. M., Rumsfeld, J. S. & Payne, P. R. O. Questions for Artificial Intelligence in Health CareQuestions for Artifical Intelligence in Health CareQuestions for Artifical Intelligence in Health Care. *JAMA* 321, 31–32 (2019).

88. el-Bouri, R., Zhu, T. & Clifton, D. *Towards Scheduling Federated Deep Learning using Meta-Gradients for Inter-Hospital Learning*. (2021).

89. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A. & Shah, N. H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Informatics Assoc.* 27, 2011–2015 (2020).

90. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv e-prints* arXiv:1706.07269 (2017).

91. Kim, B., Khanna, R. & Koyejo, O. Examples are not enough, learn to criticize! Criticism for interpretability. in *Advances in Neural Information Processing Systems* 2288–2296 (2016).

92. High-Level Independent Group on Artificial Intelligence (AI HLEG). *Ethics Guidelines for Trustworthy AI*. *European Commission* vol. 32 https://ec.europa.eu/digital (2019).

93. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv e-prints* arXiv:1606.03490 (2016).

94. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Heal.* 3, e745–e750 (2021).

95. Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Med.* 3, 70 (2020).

96. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-.).* 366, 447–453 (2019).

97. Adamson, A. S. & Smith, A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology* 154, 1247–1248 (2018).

98. Groh, M. *et al. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset*. (2021).

99. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17, 195 (2019).

100. Henry, K. E. *et al.* Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat. Med.* 28, 1447–1454 (2022).

101. McIntosh, C. *et al.* Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat. Med.* 27, 999–1005 (2021).

102. Sendak, M. P. *et al.* Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Med Inf.* 8, e15182 (2020).

103. Connell, A. *et al.* Implementation of a Digitally Enabled Care Pathway (Part 2): Qualitative Analysis of Experiences of Health Care Professionals. *J. Med. Internet Res.* 21, e13143 (2019).

104. A. Keane, P. & J. Topol, E. *With an eye to AI and autonomous diagnosis*. *npj Digital Medicine* vol. 1 (2018).

105. Vasey, B. *et al.* DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* 27, (2021).

106. Cahan, A. & Cimino, J. J. A Learning Health Care System Using Computer-Aided Diagnosis. *J. Med. Internet Res.* 19, e54 (2017).

107. Park, S. H., Choi, J. Il, Fournier, L. & Vasey, B. Randomized Clinical Trials of Artificial Intelligence in Medicine: Why, When, and How? *Korean J. Radiol.* 23, 1119–1125 (2022).

108. Brocklehurst, P. *et al.* Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. *Lancet* 389, 1719–1729 (2017).

109. Connell, A. *et al.* Evaluation of a digitally-enabled care pathway for acute kidney injury management in hospital emergency admissions. *NPJ Digit. Med.* 2, 67 (2019).

110. Connell, A. *et al.* Implementation of a Digitally Enabled Care Pathway (Part 1): Impact on Clinical Outcomes and Associated Health Care Costs. *J. Med. Internet Res.* 21, e13147 (2019).

111. van de Sande, D., van Genderen, M. E., Huiskens, J., Gommers, D. & van Bommel, J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive care medicine* vol. 47 750–760 (2021).

112. Park, S. H. Regulatory Approval versus Clinical Validation of Artificial Intelligence Diagnostic Tools. *Radiology* 288, 910–911 (2018).

113. van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B. & de Rooij, M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur. Radiol.* 31, 3797–3804 (2021).

114. Price, W. N. 2nd, Gerke, S. & Cohen, I. G. Potential Liability for Physicians Using Artificial Intelligence. *JAMA* (2019) doi:10.1001/jama.2019.15064.

115. Reverberi, C. *et al.* Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* 12, 14952 (2022).

116. Steiner, D. F. *et al.* Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* 42, 1636–1646 (2018).

117. Sayres, R. *et al.* Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* 126, 552–564 (2019).

118. International Ergonomics Association. What is Ergonomics? https://iea.cc/what-is-ergonomics/.

119. World Health Organization. *Technical series on Safer Primary Care - Human factors*. (2016).

120. Bainbridge, L. Ironies of automation. in (eds. JOHANNSEN, G. & RIJNSDORP Design and Evaluation of Man–Machine Systems, J. E. B. T.-A.) 129–135 (Pergamon, 1983). doi:https://doi.org/10.1016/B978-0-08-029348-6.50026-9.

121. Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C. & Jenkins, D. P. *Human Factors Methods*. (Ashgate Publishing, 2005).

122. Wronikowska, M. W. *et al.* Systematic review of applied usability metrics within usability evaluation methods for hospital electronic healthcare record systems. *J. Eval. Clin. Pract.* 27, 1403–1416 (2021).

123. Carayon, P., Xie, A. & Kianfar, S. Human factors and ergonomics as a patient safety practice. *BMJ Qual. & Saf.* 23, 196 LP – 205 (2014).

124. Sujan, M. *et al.* The contribution of human factors and ergonomics to the design and delivery of safe future healthcare. *Futur. Healthc. J.* 8, e574 LP-e579 (2021).

125. Catchpole, K. Spreading human factors expertise in healthcare: untangling the knots in people and systems. *BMJ Qual. & Saf.* 22, 793 LP – 797 (2013).

126. International Ergonomics Association. Safety and Health. https://iea.cc/member/safety-health/.

127. Chartered Institute of Ergonomics & Human Factors. Healthcare. https://ergonomics.org.uk/connect/sector-groups/healthcare.html.

128. Regulation (EU) 2017/745 of the European Parliament and of the Council (EU Medical Device Regulation - MDR). (2017).

129. Medicines & Healthcare products Regulatory Agency (MHRA). *Guidance on applying human factors and usability engineering to medical devices including drug-device combination products in Great Britain*. (2021).

130. US Food and Drug Administration (FDA). *Applying Human Factors and Usability Engineering to Medical Devices - Guidance for Industry and Food and Drug Administration Staff*. (2016).

131. International Organization for Standardization. *Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (ISO 9241-11:2018)*. (2018).

132. International Organization for Standardization. *Ergonomics of human-system interaction - Usability methods supporting human-centred design (ISO/TR 16982:2002)*. (2002).

133. International Organization for Standardization. *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems (ISO 9241-210:2019)*. (2019).

134. Asan, O. & Choudhury, A. Research Trends in Artificial Intelligence Applications in Human Factors Health Care: Mapping Review. *JMIR Hum. factors* 8, e28236 (2021).

135. Sujan, M. *et al.* Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Heal. & Care Informatics* 26, e100081 (2019).

136. Chartered Institute of Ergonomics & Human Factors. *Human Factors and Ergonomics in Healthcare AI*. https://ergonomics.org.uk/Public/News_Events/News_Items/AI-in-Healthcare.aspx (2021).

137. Fraser, A. G. *et al.* Improved clinical investigation and evaluation of high-risk medical devices: the rationale and objectives of CORE-MD (Coordinating Research and Evidence for Medical Devices). *Eur. Hear. journal. Qual. care Clin. outcomes* 8, 249–258 (2022).

138. US Food and Drug Administration (FDA). *Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI/ ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback*. (2019).

139. US Food and Drug Administration (FDA). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. (2021).

140. Medicines & Healthcare products Regulatory Agency (MHRA). Software and AI as a Medical Device Change Programme. (2021) https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme.

141. Medicines & Healthcare products Regulatory Agency (MHRA). *Government response to consultation on the future regulation of medical devices in the United Kingdom*. (2022).

142. IMDRF Software as Medical Device (SaMD) Working Group. *'Software as a Medical Device': Possible Framework for Risk Categorization and Corresponding Considerations*. (2014).

143. IMDRF Software as Medical Device (SaMD) Working Group. *Software as a Medical Device (SAMD): Clinical Evaluation*. (2017).

144. IMDRF Medical Device Clinical Evaluation Working Group. *Clinical Evaluation*. (2019).

145. IMDRF Medical Device Clinical Evaluation Working Group. *Clinical Investigation*. (2019).

146. IMDRF AIMD Working Group. *Machine Learning-enabled Medical Devices-A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions*. (2021).

147. World Health Organization*. Ethics and governance of artificial intelligence for health: WHO guidance.* 2021. Licence: CC BY-NC-SA 3.0 IGO.

148. McCradden, M. D., Stephenson, E. A. & Anderson, J. A. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat. Med.* 26, 1325–1326 (2020).

149. Sendak, M. P. *et al.* A path for translation of machine learning products into healthcare delivery. *EMJ Innov.* (2020) doi:DOI/10.33590/emjinnov/19-00172.

150. Higgins, D. & Madai, V. I. From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Adv. Intell. Syst.* 2, 2000052 (2020).

151. Park, Y. *et al.* Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 3, 326–331 (2020).

152. Liu, X. *et al.* The medical algorithmic audit. *Lancet Digit. Heal.* 4, e384–e397 (2022).

153. Embi, P. J. Algorithmovigilance—Advancing Methods to Analyze and Monitor Artificial Intelligence–Driven Health Care for Effectiveness and Equity. *JAMA Netw. Open* 4, e214622–e214622 (2021).

154. Lekadir, K. *et al.* FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Future Medical Imaging. (2021).

155. P., S. P. *et al.* Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist. *JACC Cardiovasc. Imaging* 13, 2017–2035 (2020).

156. Schwendicke, F. *et al.* Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J. Dent.* 107, 103610 (2021).

157. Hawkins, R. *et al. Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)*. (2021).

158. International Organization for Standardization; & International Electrotechnical Commission. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (ISO/IEC 22989:2022)*. (2022).

159. International Organization for Standardization & International Electrotechnical Commission. *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) - ISO/IEC 23053:2022*. (2022).

160. International Organization for Standardization. *Medical devices - Application of risk management to medical devices (ISO 14971:2019)*. (2019).

161. International Organization for Standardization. Medical devices - Guidance on the application of ISO 14971 (ISO/TR 24971:2020). (2020).

162. International Electrotechnical Commission. Medical device software - Software life cycle processes (IEC 62304:2006). (2006).

163. International Electrotechnical Commission. *Medical devices - Part 1: Application of usability engineering to medical devices (IEC 62366-1:2015)*. (2015).

164. International Electrotechnical Commission. Medical devices - Part 2: Guidance on the application of usability engineering to medical devices (IEC/TR 62366-2:2016). (2016).

165. British Standards Institution. *Medical devices. Application of usability engineering to medical devices (BS EN 62366-1:2015+A1:2020)*. (2020).

166. International Organization for Standardization. *Clinical investigation of medical devices for human subjects - Good clinical practice (ISO 14155:2020)*. (2020).

167. Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* 26, 1364–1374 (2020).

168. Cruz Rivera, S. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* 26, 1351–1363 (2020).

169. Collins, G. S. *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 11, e048008 (2021).

170. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* 162, 55–63 (2015).

171. Sounderajah, V. *et al.* Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 11, e047709 (2021).

172. Bossuyt, P. M. *et al.* STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ Br. Med. J.* 351, h5527 (2015).

173. Kitaguchi, D. *et al.* Artificial Intelligence for Computer Vision in Surgery: A Call for Developing Reporting Guidelines. *Ann. Surg.* 275, (2022).

174. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* 26, 1320–1324 (2020).

175. Olczak, J. *et al.* Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop.* 92, 513–525 (2021).

176. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* 2, e200029 (2020).

177. Sendak, M. P., Gao, M., Brajer, N. & Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit. Med.* 3, 41 (2020).

178. Wolff, R. F. *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* 170, 51–58 (2019).

179. Daneshjou, R. *et al.* Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology: CLEAR Derm Consensus Guidelines From the International Skin Imaging Collaboration Artificial Intelligence Working Group. *JAMA Dermatology* 158, 90–96 (2022).

180. Petzold, A. *et al.* Artificial intelligence extension of the OSCAR-IB criteria. *Ann. Clin. Transl. Neurol.* 8, 1528—1542 (2021).

181. Vollmer, S. *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 368, l6927 (2020).

182. Scott, I., Carter, S. & Coiera, E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Heal. & Care Informatics* 28, e100251 (2021).

183. Craig, P. *et al.* Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 337, a1655 (2008).

184. Campbell, M. *et al.* Framework for design and evaluation of complex interventions to improve health. *BMJ* 321, 694–696 (2000).

185. Skivington, K. *et al.* A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 374, n2061 (2021).

186. McCulloch, P. *et al.* No surgical innovation without evaluation: the IDEAL recommendations. *Lancet* 374, 1105–1112 (2009).

187. Hirst, A. *et al.* No Surgical Innovation Without Evaluation: Evolution and Further Development of the IDEAL Framework and Recommendations. *Ann. Surg.* 269, 211–220 (2019).

188. Sedrakyan, A. *et al.* IDEAL-D: A rational framework for evaluating and regulating the use of medical devices. *BMJ* 353, i2372 (2016).

189. Marcus, H. J. *et al.* IDEAL-D Framework for Device Innovation: A Consensus Statement on the Preclinical Stage. *Ann. Surg.* (2021).

190. Beard, D. *et al.* Evidence-Based Evaluation of Practice and Innovation in Physical Therapy Using the IDEAL-Physio Framework. *Phys. Ther.* 98, 108–121 (2018).

191. Verkooijen, H. M. *et al.* R-IDEAL: A Framework for Systematic Clinical Evaluation of Technical Innovations in Radiation Oncology. *Front. Oncol.* 7, 59 (2017).

192. Bilbro, N. A. *et al.* The IDEAL reporting guidelines: A delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. *Ann. Surg.* 273, 82–85 (2021).

193. Begg, C. *et al.* Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 276, 637–639 (1996).

194. Schulz, K. F., Altman, D. G. & Moher, D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, c332 (2010).

195. Barwise, A., Caples, S., Jensen, J., Pickering, B. & Herasevich, V. Information needs for the rapid response team electronic clinical tool. *BMC Med. Inform. Decis. Mak.* 17, 142 (2017).

196. Aakre, C. A., Chaudhry, R., Pickering, B. W. & Herasevich, V. Information Needs Assessment for a Medicine Ward-Focused Rounding Dashboard. *J. Med. Syst.* 40, 183 (2016).

197. Herasevich, V., Ellsworth, M. A., Hebl, J. R., Brown, M. J. & Pickering, B. W. Information needs for the OR and PACU electronic medical record. *Appl. Clin. Inform.* 5, 630–641 (2014).

198. Ellsworth, M. A., Lang, T. R., Pickering, B. W. & Herasevich, V. Clinical data needs in the neonatal intensive care unit electronic medical record. *BMC Med. Inform. Decis. Mak.* 14, 92 (2014).

199. Sullivan, M. E. *et al.* The use of cognitive task analysis to improve the learning of percutaneous tracheostomy placement. *Am. J. Surg.* 193, 96–99 (2007).

200. Sullivan, M. E. *et al.* Assessing the teaching of procedural skills: can cognitive task analysis add to our traditional teaching methods? *Am. J. Surg.* 195, 20–23 (2008).

201. Joeres, F. *et al.* How well do software assistants for minimally invasive partial nephrectomy meet surgeon information needs? A cognitive task analysis and literature review study. *PLoS One* 14, e0219920 (2019).

202. Craig, C. *et al.* Using cognitive task analysis to identify critical decisions in the laparoscopic environment. *Hum. Factors* 54, 1025–1039 (2012).

203. Militello, L. G. & Hutton, R. J. Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *Ergonomics* 41, 1618–1641 (1998).

204. Klein, G. A., Calderwood, R. & MacGregor, D. Critical decision method for eliciting knowledge. *IEEE Trans. Syst. Man. Cybern.* 19, 462–472 (1989).

205. Polson, P. G., Lewis, C., Rieman, J. & Wharton, C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *Int. J. Man. Mach. Stud.* 36, 741–773 (1992).

206. Tong, A., Sainsbury, P. & Craig, J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int. J. Qual. Heal. Care* 19, 349–357 (2007).

207. Morgan, M. G., Fischhoff, B., Bostrom, A. & Atman, C. J. *Risk Communication: A Mental Models Approach*. (Cambridge University Press, 2001).

208. Guest, G., Bunce, A. & Johnson, L. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field methods* 18, 59–82 (2006).

209. Kiger, M. E. & Varpio, L. Thematic analysis of qualitative data: AMEE Guide No. 131. *Med. Teach.* 42, 846–854 (2020).

210. TP Transcription. Transcription Editing Levels. https://www.tptranscription.co.uk/verbatim-transcription-levels/.

211. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101 (2006).

212. Patton, M. Q. *Qualitative evaluation and research methods (2nd edition)*. (Sage, 1990).

213. Guest, G., Namey, E. & Chen, M. A simple method to assess and report thematic saturation in qualitative research. *PLoS One* 15, e0232076 (2020).

214. Glaser, B. G. & Strauss, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research (1st ed.)*. (Routledge, 1999).

215. Shimizu, T., Nemoto, T. & Tokuda, Y. Effectiveness of a clinical knowledge support system for reducing diagnostic errors in outpatient care in Japan: A retrospective study. *Int. J. Med. Inform.* 109, 1–4 (2018).

216. Phua, J., See, K. C., Khalizah, H. J., Low, S. P. & Lim, T. K. Utility of the electronic information resource UpToDate for clinical decision-making at bedside rounds. *Singapore Med. J.* 53, 116–120 (2012).

217. Universitätspital Basel. medStandards. http://www.medstandards.ch.

218. Zamberg, I. *et al.* A Mobile Medical Knowledge Dissemination Platform (HeadToToe): Mixed Methods Study. *JMIR Med Educ* 6, e17729 (2020).

219. Connell, A. *et al.* Service evaluation of the implementation of a digitally-enabled care pathway for the recognition and management of acute kidney injury. *F1000Research* 6, 1033 (2017).

220. Wijnberge, M. *et al.* Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *JAMA* 323, 1052–1060 (2020).

221. Wong, D. *et al.* SEND: a system for electronic notification and documentation of vital sign observations. *BMC Med. Inform. Decis. Mak.* 15, 68 (2015).

222. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* 24, 1716–1720 (2018).

223. Pehrson, L. M., Nielsen, M. B. & Ammitzbøl Lauridsen, C. Automatic Pulmonary Nodule Detection Applying Deep Learning or Machine Learning Algorithms to the LIDC-IDRI Database: A Systematic Review. *Diagnostics (Basel, Switzerland)* 9, (2019).

224. Hassan, C. *et al.* Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointest. Endosc.* 93, 77-85.e6 (2021).

225. Fenn, A. *et al.* Development and Validation of Machine Learning Models to Predict Admission From Emergency Department to Inpatient and Intensive Care Units. *Ann. Emerg. Med.* 78, 290–302 (2021).

226. Kassahun, Y. *et al.* Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int. J. Comput. Assist. Radiol. Surg.* 11, 553–568 (2016).

227. Ma, R., Vanstrum, E. B., Lee, R., Chen, J. & Hung, A. J. Machine learning in the optimization of robotics in the operative field. *Curr. Opin. Urol.* 30, 808–816 (2020).

228. Panesar, S. *et al.* Artificial Intelligence and the Future of Surgical Robotics. *Ann. Surg.* 270, 223–226 (2019).

229. Zhou, X.-Y., Guo, Y., Shen, M. & Yang, G.-Z. Application of artificial intelligence in surgery. *Front. Med.* 14, 417–430 (2020).

230. Bryman, A. *Quantity and Quality in Social Research*. (Routledge, 1988).

231. Haselton, M. G., Nettle, D. & Murray, D. R. The Evolution of Cognitive Bias. *The Handbook of Evolutionary Psychology* 1–20 (2015).

232. Varghese, J., Kleine, M., Gessner, S. I., Sandmann, S. & Dugas, M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. *J. Am. Med. Inform. Assoc.* 25, 593–602 (2018).

233. Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339, b2535 (2009).

234. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.

235. @EricTopol on Twitter. Recent FDA approvals/clearances. https://twitter.com/EricTopol/status/1119683505603006469.

236. Whiting, P. F. *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* 155, 529–536 (2011).

237. Sterne, J. A. C. *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355, i4919 (2016).

238. Aissa, J. *et al.* Performance and clinical impact of machine learning based lung nodule detection using vessel suppression in melanoma patients. *Clin. Imaging* 52, 328–333 (2018).

239. Aslantas, A., Dandil, E., Sağlam, S. & Çakiroğlu, M. CADBOSS: A computer-aided diagnosis system for whole-body bone scintigraphy scans. *J. Cancer Res. Ther.* 12, 787–792 (2016).

240. Bargalló, X. *et al.* Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. *Eur. J. Radiol.* 83, 2019–2023 (2014).

241. Barinov, L. *et al.* Impact of Data Presentation on Physician Performance Utilizing Artificial Intelligence-Based Computer-Aided Diagnosis and Decision Support Systems. *J. Digit. Imaging* 32, 408–416 (2019).

242. Bartolotta, T. V. *et al.* Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. *Radiol. Med.* 123, 498–506 (2018).

243. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* 15, e1002699 (2018).

244. Van Den Biggelaar, F. J. H. M., Kessels, A. G. H., Van Engelshoven, J. M. A., Boetes, C. & Flobbe, K. Computer-aided detection in full-field digital mammography in a clinical population: Performance of radiologist and technologists. *Breast Cancer Res. Treat.* 120, 499–506 (2010).

245. Blackmon, K. N. *et al.* Computer-aided detection of pulmonary embolism at CT pulmonary angiography: Can it improve performance of inexperienced readers? *Eur. Radiol.* 21, 1214–1223 (2011).

246. Cha, K. H. *et al.* Diagnostic Accuracy of CT for Prediction of Bladder Cancer Treatment Response with and without Computerized Decision Support. *Acad. Radiol.* 26, 1137–1145 (2019).

247. Chabi, M. L. *et al.* Evaluation of the Accuracy of a Computer-aided Diagnosis (CAD) System in Breast Ultrasound according to the Radiologist's Experience. *Acad. Radiol.* 19, 311–319 (2012).

248. Cho, E., Kim, E.-K., Song, M. K. & Yoon, J. H. Application of Computer-Aided Diagnosis on Breast Ultrasonography: Evaluation of Diagnostic Performances and Agreement of Radiologists According to Different Levels of Experience. *J. ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* 37, 209–216 (2018).

249. Choi, J.-H., Kang, B. J., Baek, J. E., Lee, H. S. & Kim, S. H. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. *Ultrason. (Seoul, Korea)* 37, 217–225 (2018).

250. Choi, J. S. *et al.* Effect of a Deep Learning Framework-Based Computer-Aided Diagnosis System on the Diagnostic Performance of Radiologists in Differentiating between Malignant and Benign Masses on Breast Ultrasonography. *Korean J. Radiol.* 20, 749–758 (2019).

251. Cole, E. B. *et al.* Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR. Am. J. Roentgenol.* 203, 909–916 (2014).

252. Endo, M. *et al.* Content-based image-retrieval system in chest computed tomography for a solitary pulmonary nodule: method and preliminary experiments. *Int. J. Comput. Assist. Radiol. Surg.* 7, 331–338 (2012).

253. Engelke, C., Schmidt, S., Auer, F., Rummeny, E. J. & Marten, K. Does computer-assisted detection of pulmonary emboli enhance severity assessment and risk stratification in acute pulmonary embolism? *Clin. Radiol.* 65, 137–144 (2010).

254. Giannini, V. *et al.* Multiparametric magnetic resonance imaging of the prostate with computer-aided detection: experienced observer performance study. *Eur. Radiol.* 27, 4200–4208 (2017).

255. Hwang, E. J. *et al.* Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw. Open* 2, e191095–e191095 (2019).

256. Lindsey, R. *et al.* Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11591–11596 (2018).

257. Park, H. J. *et al.* A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: Added value for the inexperienced breast radiologist. *Medicine (Baltimore).* 98, e14146–e14146 (2019).

258. Rodríguez-Ruiz, A. *et al.* Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* 290, 305–314 (2019).

259. Romero, C. *et al.* Impact on breast cancer diagnosis in a multidisciplinary unit after the incorporation of mammography digitalization and computer-aided detection systems. *AJR. Am. J. Roentgenol.* 197, 1492–1497 (2011).

260. Samulski, M. *et al.* Using computer-aided detection in mammography as a decision support. *Eur. Radiol.* 20, 2323–2330 (2010).

261. Sanchez Gómez, S. *et al.* Impact of a CAD system in a screen-film mammography screening program: a prospective study. *Eur. J. Radiol.* 80, e317-21 (2011).

262. Shimauchi, A. *et al.* Evaluation of Clinical Breast MR Imaging Performed with Prototype Computer-aided Diagnosis Breast MR Imaging Workstation: Reader Study. *Radiology* 258, 696–704 (2011).

263. Sohns, C., Angic, B. C., Sossalla, S., Konietschke, F. & Obenauer, S. CAD in full-field digital mammography-influence of reader experience and application of CAD on interpretation of time. *Clin. Imaging* 34, 418–424 (2010).

264. Stoffel, E. *et al.* Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis. *Eur. J. Radiol. open* 5, 165–170 (2018).

265. Sun, L. *et al.* A computer-aided diagnostic algorithm improves the accuracy of transesophageal echocardiography for left atrial thrombi: a single-center prospective study. *J. ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* 33, 83–91 (2014).

266. Sunwoo, L. *et al.* Computer-aided detection of brain metastasis on 3D MR imaging: Observer performance study. *PLoS One* 12, e0178265 (2017).

267. Tang, F.-H., Ng, D. K. S. & Chow, D. H. K. An image feature approach for computer-aided detection of ischemic stroke. *Comput. Biol. Med.* 41, 529–536 (2011).

268. Taylor, J. C. *et al.* Computer-aided diagnosis for (123I)FP-CIT imaging: impact on clinical reporting. *EJNMMI Res.* 8, 36 (2018).

269. Vassallo, L. *et al.* A cloud-based computer-aided detection system improves identification of lung nodules on computed tomography scans of patients with extra-thoracic malignancies. *Eur. Radiol.* 29, 144–152 (2019).

270. Watanabe, A. T. *et al.* Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. *J. Digit. Imaging* (2019).

271. Way, T. *et al.* Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance. *Acad. Radiol.* 17, 323–332 (2010).

272. Zhang, J., Wang, Y., Yu, B., Shi, X. & Zhang, Y. Application of Computer-Aided Diagnosis to the Sonographic Evaluation of Cervical Lymph Nodes. *Ultrason. Imaging* 38, 159–171 (2016).

273. Han, S. S. *et al.* Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders. *J. Invest. Dermatol.* 140, 1753–1761 (2020).

274. Rodger, M., Ramsay, T. & Fergusson, D. Diagnostic randomized controlled trials: the final frontier. *Trials* 13, 137 (2012).

275. Lehman, C. D. *et al.* Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern. Med.* 175, 1828–1837 (2015).

276. Fenton, J. J. *et al.* Effectiveness of computer-aided detection in community mammography practice. *J. Natl. Cancer Inst.* 103, 1152–1161 (2011).

277. Fenton, J. J. *et al.* Short-term outcomes of screening mammography using computer-aided detection: a population-based study of medicare enrollees. *Ann. Intern. Med.* 158, 580–587 (2013).

278. Price, K. Anything you can do, I can do better (No you can't)... *Comput. Vision, Graph. Image Process.* 36, 387–391 (1986).

279. Radboud University Medical Center. Grand Challenge - Challenges. https://grand-challenge.org/challenges/.

280. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med.* 25, 1467–1468 (2019).

281. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* 393, 1577–1579 (2019).

282. Sounderajah, V. *et al.* Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat. Med.* 26, 807–808 (2020).

283. Hopper, A. N., Jamison, M. H. & Lewis, W. G. Learning curves in surgical practice. *Postgrad. Med. J.* 83, 777 LP – 779 (2007).

284. Brown, S.-A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Front. Physiol.* 7, 561 (2016).

285. Radcliffe, K., Lyson, H. C., Barr-Walker, J. & Sarkar, U. Collective intelligence in medical decision-making: a systematic scoping review. *BMC Med. Inform. Decis. Mak.* 19, 158 (2019).

286. Weissman, G. E., Ungar, L. H. & Halpern, S. D. Chess Lessons: Harnessing Collective Human Intelligence and Imitation Learning to Support Clinical Decisions. *Ann. Intern. Med.* (2023).

287. Van Spall, H. G. C., Toren, A., Kiss, A. & Fowler, R. A. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 297, 1233–1240 (2007).

288. Ioannidis, J. P. & Lau, J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA* 285, 437–443 (2001).

289. Gallego, B. *et al.* Bringing cohort studies to the bedside: framework for a 'green button' to support clinical decision-making. *J. Comp. Eff. Res.* 4, 191–197 (2015).

290. Cahan, A. & Cimino, J. J. Visual assessment of the similarity between a patient and trial population: Is This Clinical Trial Applicable to My Patient? *Appl. Clin. Inform.* 7, 477–488 (2016).

291. Vranas, K. C. *et al.* Identifying Distinct Subgroups of ICU Patients: A Machine Learning Approach. *Crit. Care Med.* 45, 1607–1615 (2017).

292. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* 7, 311ra174 LP-311ra174 (2015).

293. Parimbelli, E., Marini, S., Sacchi, L. & Bellazzi, R. Patient similarity for precision medicine: A systematic review. *J. Biomed. Inform.* 83, 87–96 (2018).

294. Sharafoddini, A., Dubin, J. A. & Lee, J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Med. informatics* 5, e7–e7 (2017).

295. Ng, K., Sun, J., Hu, J. & Wang, F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* vol. 2015 132–136 (2015).

296. Lee, J., Maslove, D. M. & Dubin, J. A. Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric. *PLoS One* 10, e0127428 (2015).

297. Song, Y., Dixon, S. & Pearce, M. *A Survey of Music Recommendation Systems and Future Perspectives*. (2012).

298. Arsan, T., Koksal, E. & Bozkus, Z. *Comparison of Collaborative Filtering Algorithms with Various Similarity Measures for Movie Recommendation*. *International Journal of Computer Science, Engineering and Applications* vol. 6 (2016).

299. Tarassenko, L. & EJ, T. Monitoring jet engines and the health of people. *JAMA* 320, 2309–2310 (2018).

300. Bruynseels, K., Santoni de Sio, F. & van den Hoven, J. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Front. Genet.* 9, 31 (2018).

301. Maheshwari, K. *et al.* Identify and monitor clinical variation using machine intelligence: a pilot in colorectal surgery. *J. Clin. Monit. Comput.* 33, 725–731 (2019).

302. Zhou, Z., Guo, B. & Zhang, C. DoseGuide: A Graph-based Dynamic Time-aware Prediction System for Postoperative Pain. in *2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS)* 474–481 (2021).

303. Stam, W. T. *et al.* The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: A systematic review. *Surgery* 171, 1014–1021 (2022).

304. Ren, Y. *et al.* Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Predict Postoperative Complications and Report on a Mobile Platform. *JAMA Netw. Open* 5, e2211973–e2211973 (2022).

305. Soguero-Ruiz, C. *et al.* Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *J. Biomed. Inform.* 61, 87–96 (2016).

306. Muralitharan, S. *et al.* Machine Learning-Based Early Warning Systems for Clinical Deterioration: Systematic Scoping Review. *J. Med. Internet Res.* 23, e25187 (2021).

307. Shamout, F. *et al.* Early warning score adjusted for age to predict the composite outcome of mortality, cardiac arrest or unplanned intensive care unit admission using observational vital-sign data: a multicentre development and validation. *BMJ Open* 9, e033301 (2019).

308. Scott, J. W. *et al.* Use of National Burden to Define Operative Emergency General Surgery. *JAMA Surg.* 151, e160480 (2016).

309. Youssef, A. *et al.* Development and validation of early warning score systems for COVID-19 patients. *Healthc. Technol. Lett.* 8, (2021).

310. O'Reilly Nugent, A. *et al.* Measurement of oxygen concentration delivered via nasal cannulae by tracheal sampling. *Respirology* 19, 538–543 (2014).

311. Malycha, J. *et al.* The effect of fractional inspired oxygen concentration on early warning score performance: A database analysis. *Resuscitation* 139, 192–199 (2019).

312. Wagstaff, T. A. J. & Soni, N. Performance of six types of oxygen delivery devices at varying respiratory rates. *Anaesthesia* 62, 492–503 (2007).

313. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* 40, 373–383 (1987).

314. Chen, J., Sun, L., Guo, C., Wei, W. & Xie, Y. A data-driven framework of typical treatment process extraction and evaluation. *J. Biomed. Inform.* 83, 178–195 (2018).

315. Li, J. *et al.* Imputation of missing values for electronic health record laboratory data. *npj Digit. Med.* 4, 147 (2021).

316. Shamout, F. E., Zhu, T., Sharma, P., Watkinson, P. J. & Clifton, D. A. Deep Interpretable Early Warning System for the Detection of Clinical Deterioration. *IEEE J. Biomed. Heal. Informatics* 24, 437–446 (2020).

317. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).

318. Chollet, F. keras. *GitHub Repos.* (2015).

319. Wang, Y., Tian, Y., Tian, L.-L., Qian, Y.-M. & Li, J.-S. An Electronic Medical Record System with Treatment Recommendations Based on Patient Similarity. *J. Med. Syst.* 39, 55 (2015).

320. Panahiazar, M., Taslimitehrani, V., Pereira, N. L. & Pathak, J. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. *Stud. Health Technol. Inform.* 210, 369—373 (2015).

321. Nijman, S. W. J. *et al.* Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J. Clin. Epidemiol.* 142, 218–229 (2022).

322. Vasey, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* 28, 924–933 (2022).

323. Vasey, B. *et al.* Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 377, e070904 (2022).

324. Kapur, N., Parand, A., Soukup, T., Reader, T. & Sevdalis, N. Aviation and healthcare: a comparative review with implications for patient safety. *JRSM open* 7, 2054270415616548–2054270415616548 (2015).

325. Corbridge, C., Anthony, M., McNeish, D. & Shaw, G. A New UK Defence Standard For Human Factors Integration (HFI). *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 60, 1736–1740 (2016).

326. Stanton, N. A., Salmon, P., Jenkins, D. & Walker, G. *Human factors in the design and evaluation of central control room operations*. (CRC Press, 2009).

327. Felmingham, C. M. *et al.* The Importance of Incorporating Human Factors in the Design and Implementation of Artificial Intelligence for Skin Cancer Diagnosis in the Real World. *Am. J. Clin. Dermatol.* 22, 233–242 (2021).

328. Scheder-Bieschin, J. *et al.* Improving Emergency Department Patient-Physician Conversation Through an Artificial Intelligence Symptom-Taking Tool: Mixed Methods Pilot Observational Study. *JMIR Form Res* 6, e28199 (2022).

329. Bilbro, N. A. *et al.* The IDEAL reporting guidelines: A delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. *Ann. Surg.* 273, 82–85 (2021).

330. EQUATOR Network. How to develop a reporting guideline. https://www.equator-network.org/toolkits/developing-a-reporting-guideline/ (2018).

331. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for Developers of Health Research Reporting Guidelines. *PLOS Med.* 7, e1000217 (2010).

332. Powell, C. The Delphi technique: myths and realities. *J. Adv. Nurs.* 41, 376–382 (2003).

333. Linstone, H. & Turoff, M. *The Delphi Method: Techniques and Applications*. *Technometrics* vol. 18 (1975).

334. Dalkey, N. & Helmer, O. An Experimental Application of the DELPHI Method to the Use of Experts. *Manage. Sci.* 9, 458–467 (1963).

335. Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci. Eng. Ethics* (2019).

336. Xie, Y. *et al.* Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. *Transl. Vis. Sci. Technol.* 9, 22 (2020).

337. National Institute for Health and Care Excellence (NICE). *Evidence standards framework for digital health technologies*. (2019).

338. Accelerated Access Collaborative & NHSx. *AI-Award Evaluation Playbook - Version 1*. (2020).

339. Harris, P. A. *et al.* Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* 42, 377–381 (2009).

340. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* 95, 103208 (2019).

341. Russell, J., Greenhalgh, T. & Taylor, M. *Patient and public involvement in NIHR research 2006-2019: policy intentions, progress and themes*. (2019).

342. Winkler, J. & Moser, R. Biases in future-oriented Delphi studies: A cognitive perspective. *Technol. Forecast. Soc. Change* 105, 63–76 (2016).

343. Bauer, G. R. Biased Outcome reporting Guidelines for Underwhelming Studies (BOGUS) statement and checklist. *BMJ* 375, (2021).

344. Simera, I., Altman, D. G., Moher, D., Schulz, K. F. & Hoey, J. Guidelines for Reporting Health Research: The EQUATOR Network's Survey of Guideline Authors. *PLOS Med.* 5, e139 (2008).

345. Bossuyt, P. M. *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. Intern. Med.* 138, W1-12 (2003).

346. von Elm, E. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 335, 806–808 (2007).

347. Eldridge, S. M. *et al.* CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ* 355, i5239 (2016).

348. Lancaster, G. A. & Thabane, L. Guidelines for reporting non-randomised pilot and feasibility studies. *Pilot Feasibility Stud.* 5, 114 (2019).

349. Hoffmann, T. C. *et al.* Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ Br. Med. J.* 348, g1687 (2014).

350. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71 (2021).

351. US Food and Drug Administration (FDA). *Clinical Decision Support Software - Draft Guidance for Industry and Food and Drug Administration Staff*. (2019).

352. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *N. Engl. J. Med.* 385, 283–286 (2021).

353. Hwang, T. J., Kesselheim, A. S. & Vokinger, K. N. Lifecycle Regulation of Artificial Intelligence– and Machine Learning–Based Software Devices in Medicine. *JAMA* 322, 2285–2286 (2019).

354. U. S. Food & Drug Administration. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. (2021).

355. Moher, D., Jones, A., Lepage, L. & Group, for the C. Use of the CONSORT Statement and Quality of Reports of Randomized TrialsA Comparative Before-and-After Evaluation. *JAMA* 285, 1992–1995 (2001).

356. Hirst, A. & Altman, D. G. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS One* 7, e35621 (2012).

357. Vasey, B. & Collins, G. S. Invited Commentary: Transparent reporting of artificial intelligence models development and evaluation in surgery: The TRIPOD and DECIDE-AI checklists. *Surgery* (2023).

358. Vasey, B., Novak, A., Ather, S., Ibrahim, M. & McCulloch, P. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. *Clin. Radiol.* 78, 130–136 (2023).

359. Fraser, A. G. *et al.* Artificial intelligence in medical device software and high-risk medical devices – a review of definitions, expert recommendations and regulatory initiatives. *Expert Rev. Med. Devices* 20, 467–491 (2023).

360. Turner, L., Shamseer, L., Altman, D. G., Schulz, K. F. & Moher, D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst. Rev.* 1, 60 (2012).