The use of propensity score methods to examine comparative effectiveness of treatments provided to children in Kenyan hospitals

Lucas Owor Malla

Somerville College

University of Oxford

Supervisors: Prof. Mike English, Prof. Rafael Perera-Salazar and Dr. Emily McFadden

A thesis submitted for the degree of Doctor of Philosophy

Hilary Term, 2018

Acknowledgements

I must first thank my supervisors – Prof. Mike English, Prof. Rafael Perera and Dr. Emily McFadden for their encouragement and guidance throughout the DPhil degree. Undertaking a DPhil degree would not have been possible without an Oxford studentship fully funded by the Nuffield Department of Medicine and Clarendon. Further, I am grateful to Kenya Medical Research Institute/Wellcome Trust (through Prof. Mike English) for allowing me to use their data for analyses included in this thesis. Finally, I would like to thank my family for their encouragement and support.

Abstract

Background: Although randomised controlled trials are the best means of measuring efficacy of interventions and inferring causality challenges of limited generalisability to real life populations, cost and applicability for some types of intervention have prompted efforts to develop alternative approaches using observational data. Such approaches are very rarely used in low income settings. This research utilised a large dataset of well characterised admission events to paediatric wards in Kenya's county referral hospitals to develop approaches to, and execute, comparative effectiveness analyses for treatment of diarrhoea and pneumonia as the top causes of morbidity and mortality in children. Specifically, the two clinical questions addressed were: (i) comparative effectiveness of penicillin monotherapy versus penicillin plus gentamicin in children with (indrawing) pneumonia; (ii) effectiveness of Zinc in treatment of children admitted with diarrhoea.

Methods: To answer these questions, a strategy for multiple imputation of missing data was developed. Propensity score based methods were then developed and evaluated to identify balanced groups receiving alternative treatment allocations (for example different antibiotic regimens for childhood pneumonia) using key covariates. In the next step the effectiveness of alternative treatment outcomes were compared. Finally, sensitivity analyses were then conducted to verify missing data assumptions and also to examine the influence of unmeasured confounders. Further, methods were tested on how to estimate propensity scores after multiple imputation.

Results: The antibiotic analyses indicated that there was no statistical difference between penicillin monotherapy and penicillin plus gentamicin in the treatment of (indrawing) pneumonia. By extension, treatment using penicillin plus gentamicin would likely not offer any advantage over using amoxicillin, the recently recommended treatment of indrawing pneumonia, as equivalence had been demonstrated between penicillin monotherapy and amoxicillin in previous trials, although non-trivial mortality was observed in this group suggesting other measures need to be investigated to prevent deaths. In the second case, Zinc was demonstrated to be effective in shortening the length of stay for children admitted with diarrhoea including those who were well-nourished (WAZ >-2) and those aged 1-5 months.

Conclusion: The analyses in this thesis provided potentially useful information to support development of clinical treatment policies. However, there remains some concern that comparative effectiveness evaluation using routine observational data may not fully mitigate risks of bias due to unobserved confounders. Rigorous approaches to such analyses using propensity scores, with detailed accounts of strategies to reduce bias and appropriate sensitivity analyses may be useful when randomised controlled trials are not feasible. However, it is probably better to pursue pragmatic trials (where randomisation is feasible) where possible to generate high quality, generalisable evidence of treatment effectiveness.

Publications

Three articles on the thesis results chapters have been published (and the fourth article on Zinc effectiveness is currently under review). The corresponding references are as follows:

- **Malla, L.,** Perera-Salazar, R., McFadden, E., Ogero, M., Stepniewska, K., English, M. (2017, ahead of print). "Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review". *Journal of Comparative Effectiveness Research*.

- **Malla L**, Perera-Salazar R, McFadden E, et al. Comparative effectiveness of injectable penicillin versus a combination of penicillin and gentamicin in children with pneumonia characterised by indrawing in Kenya: protocol for an observational study. *BMJ Open* 2017;7:e016784. doi: 10.1136/bmjopen-2017-016784.

- **Malla L,** Perera-Salazar R, McFadden E, et al. Comparative effectiveness of injectable penicillin versus a combination of penicillin and gentamicin in children with pneumonia characterised by indrawing in Kenya: a retrospective observational study. *BMJ Open* 2017;7:e019478. doi: 10.1136/bmjopen-2017-019478.

- **Malla L,** Perera-Salazar R, Akech S, et al. Examining the effectiveness of Zinc treatment in children admitted with diarrhoea in Kenya's public hospitals: An observational comparative effectiveness study. 2018 (*Submitted to International Journal of Epidemiology*).

Acronyms and abbreviations

RCTs	Randomised Controlled Trials (RCTs)
UNICEF	The United Nations Children's Fund
MDG	United Nations Millennium Development Goal
GHO	Global Health Observatory
WHO	World Health Organisation
PS	Propensity score
LMIC	Lower and Middle Income Countries
CIN	Clinical Information Network
MOH	Ministry of Health
SARAM	Service Availability and Readiness Assessment Mapping
KHSRIG	Kenya Health Sector Referral Implementation Guidelines
KEMRI	Kenya Medical Research Institute
KPA	Kenya Pediatric Association
PAR	Paediatric Admission Record
REDCap	Research Electronic Data Capture
URTI	Upper Respiratory Tract Infection
LRTI	Lower Respiratory Tract Infection
AVPU	Alert, Verbal, Pain, Unresponsive
IPTW	Inverse Probability Treatment Weighting
SMD	Standardised Mean Difference
ASMD	Absolute Standardised Mean Difference
NRCM	Neyman Rubin Causal Model
ATE	Average Treatment Effect
ATT	Average Treatment Effect for the Treated
MCAR	Missing Completely at Random
MAR	Missing at Random
MNAR	Missing Not at Random
MI	Multiple Imputation
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
LOCF	Last Observation Carried Forward
ITT	Intention to Treat
PP	Per Protocol
SHR	Sub-distribution Hazard Ratio
CER	Comparative Effectiveness Research
CPRD	Clinical Practice Research Data Link
KP	Kaiser Permanente
ALPHA	Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA) network
EMRs	Electronic Medical Record Systems

Definitions

1. **Key variables** – are variables that should determine illness severity and thus treatment assignment and are typically referred to directly in clinical treatment protocols.
2. **Auxiliary variables** – are variables that may be examined for an explicit influence on treatment assignment although according to the formal rules (the guidelines) they are not considered reasons to alter treatment assignment.
3. **Surrogate variables** – are variables which may be used as proxies for (or in place of) *key variables*.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter is subdivided into four sections which I summarise (and thereafter discuss) as follows:

1.1. Background – provides childhood mortality statistics for sub – Saharan African countries (narrowing to Kenya) compared to developed countries; and also specifically examines the top causes of child morbidity and mortality (which are pneumonia and dehydration) in Kenya. This section then proceeds to discuss the strengths and limitations for using randomised controlled trials (RCTs) to compare treatment effects, summarises the current evidence from RCTs with respect to antibiotic use in pneumonia treatment and Zinc co – treatment for children with diarrhoea, and provides the rationale and guidelines for conducting comparative effectiveness evaluation using routinely collected observational data.

1.2. Data description – Here I provide a brief introduction to the health system in Kenya before introducing the Clinical Information Network (a platform for routinely collected data in Kenya). In this second section, I also provide a descriptive analysis of the admission and mortality patterns within the Clinical Information Network hospitals to give the reader an understanding of this low-income context.

1.3. Thesis objectives – outlines the thesis objectives.

1.4. Overview of thesis chapters – In this final section I provide a brief explanation of the content of each of the chapters that form the rest of this thesis and how they match the thesis objectives.

1.1. Background

1.1.1. Childhood mortality in Sub – Saharan Africa compared to developed countries

The need to scale up interventions with proven efficacy in reducing high rates of childhood mortality is particularly urgent in developing countries. Many organisations, the United Nations in particular, have promoted intensified campaigns and programmes that have seen worldwide child mortality decline by approximately 50% in the last quarter of a century (1). According to the UNICEF 2014 report, however, developing countries still contribute up to 90% of childhood deaths despite the overall decline (1). Sub – Saharan African countries have remained the largest contributors to overall global under five mortality (2). The fourth United Nations Millennium Development Goal (MDG 4) targeted reduction of child mortality by two-thirds before 2015 (3). Most developed and middle income countries made notable progress towards achievement of this goal while countries in sub – Saharan Africa made much less progress. This variation in progress is demonstrated using the all-cause mortality data derived from the Global Health Observatory (GHO) Data – published in the WHO website (4) and presented in **Figure 1.1.1**. By 2015, under five mortality rate was about 83 per 1 000 children in Sub – Saharan African countries – which was 12 times that of the overall rate for developed countries (which reported 6.8 deaths per 1000 children). Kenya's child mortality rate was about 49.2 per 1000 children – which was lower than that of the overall mortality rate for Sub – Saharan African countries though 7.2 times that of developed countries.

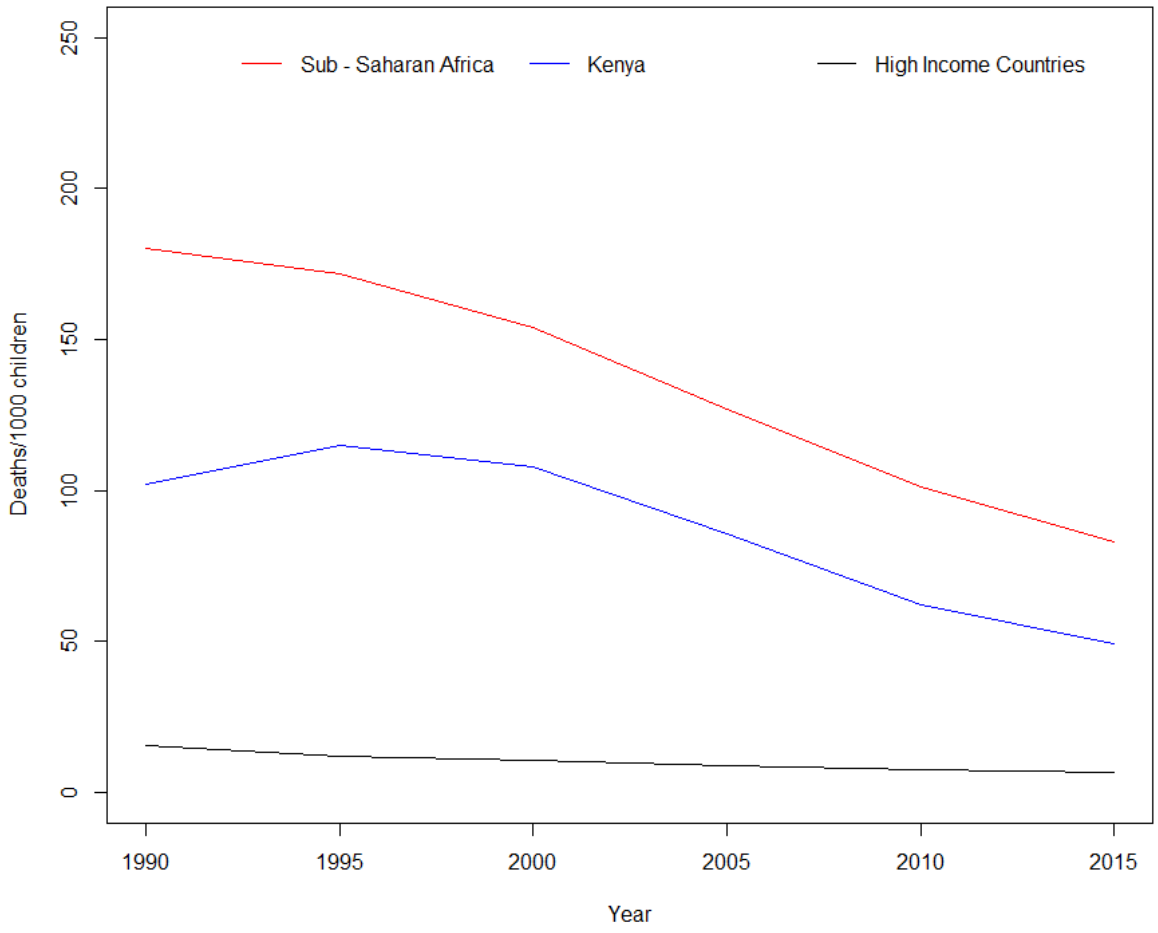


Figure 1.1.1: Comparison of under five deaths, rate per 1000 between Kenya, sub - Saharan and high income countries

1.1.2. Top causes of morbidity and mortality in Kenya

According to the 2014 Kenya Demographic and Health Survey report (5) (also UNICEF (2010) website (6)), the top two causes of both morbidity and mortality in Kenya were pneumonia and diarrhoea/dehydration. Indeed, based on the 2015 GHO data, percentages of child deaths caused by pneumonia and diarrhoea were 20.3% and 13.1% respectively (7). This is despite improved access to recommended treatments and deployment of childhood vaccines at high

coverage, including those against *H. influenzae* Type B and pneumococcus (for pneumonia, with at least 79% coverage in 2014 (5)) and rotavirus (for diarrhoea, with 66% coverage in 2015 (8)). The proportion of child deaths contributed by these two top causes have remained somewhat constant between 2000 and 2015 (**Figure 1.1.2**) despite the decline in the absolute under five mortality rates in Kenya (9) (also see **Figure 1.1.1**).

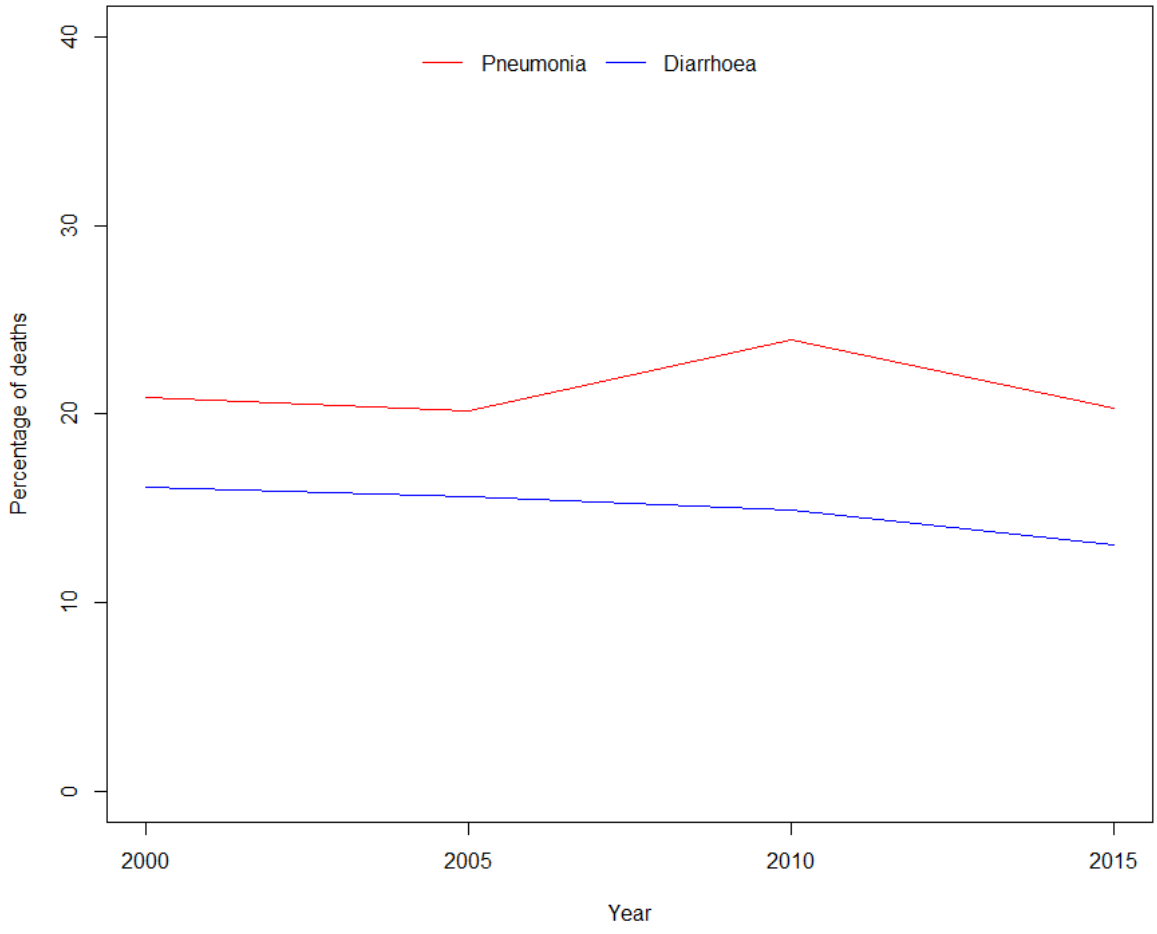


Figure 1.1.2: Percentages of child deaths in Kenya contributed by pneumonia and diarrhoea from 2000 – 2015 (GHO data).

Further reductions in under five mortality would, in theory, be accelerated if additional efficacious interventions were implemented on a large scale and in

accordance with the latest evidence, often summarised in WHO or national guidelines (10). When prioritising which interventions to deliver at scale, of the many that are possible, we typically consider an intervention's magnitude of effect, its costs and the size of the population who might benefit. Ideally, we also consider effects on equity, and user and patient preferences. Critical to much of this decision making is our estimate of an intervention's magnitude of effect. These estimates are ideally derived from randomised, controlled, clinical trials (RCTs) which are generally considered the gold standard for estimating causal effects of treatments (11).

1.1.3. Strengths and limitations of using randomised controlled trials

If a RCT is well-designed and executed, randomisation ensures that both measured and unmeasured factors are comparable between treatment groups, meaning that any effects seen are likely to be due to the treatment and not to other factors (12). The randomisation process may be concealed such that either the investigators or patients or both are not aware of the treatments that are to be assigned and they and the investigators may remain blinded to the nature of the treatment being received to further minimise any bias that might arise (13). Randomised trials, if well conducted, have high internal validity providing an intervention efficacy estimate typically within a defined, trial population. Such efficacy estimates can be examined for consistency and be made more precise if there are multiple trials and meta-analysis is possible. Most RCTs are, however, conducted at single sites, or at a limited number of selected sites. They are often also conducted under highly controlled conditions with strict patient inclusion/exclusion criteria. For example, Silverman (2009) stated that more

than half of osteoporosis patients would not be eligible for a trial due to the presence of comorbidities and or previous use of bone active treatment agents (14). There may also be very limited unbiased (or representative) efficacy information on a number of patient groups, particularly where there is greatest need of care (e.g. the disabled and those with complicated medical situations). Additionally, trial populations may not always include the heterogeneous populations presenting for care, many of whom at hospital level may have co-morbidity (15).

This lack of information creates uncertainty about possible outcomes in the population that was not enrolled or did not meet RCT inclusion criteria (and this is important if it forms a sizeable fraction of the population). Therefore, true benefits and harms of a treatment at a population level may not be known and effective treatment may be underutilised with only a small population benefitting from it or a treatment may be given to a larger population in which its effects are unknown (16). Effects at population level may then be beneficial, potentially harmful or there may be no effect, the latter two possibilities also resulting in waste. To overcome this challenge in a RCT including different patient groups, sub – group analysis may be possible though often these have only limited power. As a result, RCTs often have limited external validity (generalizability) to the populations of patients who may receive such treatments, even after meta-analysis. Lastly, RCTs are expensive and time consuming and they will likely get more expensive as we move from interventions with a large effect size to ones with smaller but nonetheless

valuable effects, while we may need multiple trials to obtain convincing evidence of effects. Also, it takes a long time to translate RCT results into practice (17).

RCT findings are used by the World Health Organisation (WHO) to inform treatment policies for millions of children with illnesses every year across low and middle income countries (18). However, trials supporting recommendations for hospitalized children (for example those with pneumonia) have included fewer participants from Africa than other settings (19) and it is suggested that African children have higher mortality (7). In sub – section 1.1.4, I summarise the evidence from RCTs some of which have informed treatment guidelines for pneumonia and diarrhoea.

1.1.4. Current evidence for pneumonia antibiotic treatment and Zinc co – treatment for diarrhoea.

1.1.4.1. Pneumonia antibiotic treatment

The most current systematic reviews by Lassi (2014) (20) and Lodha (2013) (21) were examined to determine the antibiotic treatments for pneumonia that have been investigated in RCTs, and also locations and number of participants enrolled in these RCTs. The specific pneumonia trials (forty-one in number), which were published between 1981 and 2011, have been summarised in **Appendix A.1**. The participant enrolment was as low as 34 and as high as 2188 with the majority conducted in single sites in non – African countries. In the systematic reviews, more than one study compared amoxicillin with co-trimoxazole, injectable penicillin or ampicillin with oral amoxicillin, and ampicillin or penicillin with chloramphenicol. Classifications of pneumonia severity in these studies were done according to the 2005 WHO guidelines (22).

In those classified as having very severe community acquired pneumonia, death rates were higher in those treated with chloramphenicol alone than in those treated with penicillin or ampicillin plus gentamicin. In those with severe pneumonia, approximately similar treatment failure rates were reported for those treated with either oral amoxicillin/co – trimoxazole or injectable penicillin. And for those classified with non – severe pneumonia, amoxicillin and co – trimozazole had similar failure/cure rates.

Of relevance to this thesis, there were no comparisons of narrow-spectrum monotherapy with the broader spectrum penicillin plus gentamicin combination therapy regimen for treating any form of pneumonia. Penicillin plus gentamicin combination therapy is the standard broad spectrum antibiotic regimen for serious infection in children in Kenya. As indicated in the systematic reviews, the trials conducted are not informed by knowledge of aetiological agents (either viral or bacterial). The basis for selecting narrow spectrum or broad spectrum regimens for different types of pneumonia is therefore largely based on extrapolation from specific aetiological studies (23) and on observational studies examining the risk of poor outcomes, with use of broader spectrum regimens typically used when risk is higher. In practice, therefore deployment of treatments identified as effective from RCTs are given to children with different aetiological agents. This may increase uncertainties in outcomes if the aetiology of disease varies across countries. In African children even within hospitals there is no thorough diagnostic workup to guide treatment and clinicians typically rely on clinical signs and symptoms to direct treatment and knowledge. Thus, it is

important that the actual effectiveness of regimens is examined in contexts relevant to Kenya.

1.1.4.2. Zinc co – treatment for diarrhoea

According to the WHO guidelines (24), Zinc should be used as a co – treatment for children with diarrhoea (first line treatment is detailed in sub – section 1.2.5.1). Lazzerrini (2016) in a systematic review, examined the use of Zinc in reducing the duration of diarrhoea for children aged 1 – 5 and 6 – 59 months respectively (25). This review included 33 trials (with 10 841 children) most of which were conducted in the Asian countries. Only two studies from Africa (Ethiopia (n = 177) and Nigeria (n = 60)) were included and these examined the use of Zinc treatment in children less than six months. No trial was conducted in Africa to provide evidence for effectiveness of Zinc in children aged 6 – 59 months (though it is indicated that Zinc effectiveness is likely to vary from place to place (26)). Based on the review, the use of Zinc was found to shorten the duration of diarrhoea by about half a day for those aged 6 – 59 months. The effects appeared greater for those who were malnourished as the duration of diarrhoea shortened by a day. However, no benefits were reported for children aged 1 – 5 months. The review also concluded there was still inadequate evidence for well – nourished children.

Kenyan guidelines recommended Zinc for all children aged 1 – 59 months with diarrhoea (27) despite the relatively limited evidence. Thus, there is also uncertainty around the types of outcomes expected with Zinc treatment in the Kenyan childhood population with diarrhoea.

Having discussed the available evidence for pneumonia and diarrhoea treatment, I discuss in sub – section 1.1.5 the guideline recommendations on how children with these illnesses should be treated in Kenya.

1.1.5. Kenyan Guideline recommendations

1.1.5.1. Recommendations for pneumonia (and other respiratory tract infections) diagnosis and treatment

A child who is older than 60 days and has either history of cough or difficulty in breathing should be examined for either upper or lower respiratory tract infection (URTI or LRTI) (28). Majority of URTIs are viral infections of the upper respiratory airways consisting of the nose, pharynx and larynx (these include the common cold and pharyngitis among others). While the lower respiratory tract consists of the trachea, primary bronchi and lungs. LRTIs are conditions caused by either bacteria (examples *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Staphylococcus aureus*) or viruses (examples *influenza A and B, respiratory syncytial virus*) (29). Examples of LRTIs include bronchitis, bronchiolitis and pneumonia. The most serious LRTI among children is pneumonia which results in inflammation of parts of the lung (lobes) in one or both of the lungs. In this condition, the lung space normally filled with air may fill and or solidify with pus or fluids, a situation referred to as consolidation that is sometimes seen on chest X ray (30).

In clinical practice it is typically not possible for clinicians with limited skills to distinguish reliably between pneumonia caused by bacteria – the target of treatment with antibiotics – and pneumonias or other LRTI caused by viruses

even using chest X ray (30). In practice therefore, especially in developing countries where pneumonia causes considerable mortality, it is recommended that antibiotic treatment is provided based on characteristics of the clinical presentation alone focused particularly on the severity of illness. Until March 2016, the Kenyan paediatric protocol recommended classification of pneumonia severity into very severe, severe and non – severe categories (see **Box 1.1.1** for clinical signs that were integrated to form these respective severity categories). Children with very severe pneumonia were then treated with a combination of penicillin and gentamicin, those with severe pneumonia (specifically characterised by the clinical sign lower chest wall indrawing) were treated with penicillin alone and those with non – severe pneumonia were treated with amoxicillin. The planned comparative effectiveness for antibiotics for pneumonia will be based on data until March 2016. For the purposes of this thesis I will use the term 'indrawing pneumonia' to indicate those children meeting criteria for the pre-2016 category of severe pneumonia. Some children presenting with what appears to be pneumonia may have wheeze and possible asthma. Kenyan guidelines for managing possible pneumonia in the context of wheeze are also discussed in **Box 1.1.1**.

The treatment guidelines for pneumonia were revised in March 2016 after a Kenyan trial demonstrated equivalence between oral amoxicillin and injectable penicillin for the treatment of indrawing pneumonia (31), a finding compatible with trials conducted in Asia. The pneumonia severity classifications were then revised into only two categories, severe (equivalent to the earlier very severe category) and just pneumonia (combining the earlier severe (or indrawing

pneumonia) and non – severe categories). As indicated in **Box 1.1.1**, treatment for the former category is still penicillin plus gentamicin while recommended treatment for the latter is oral amoxicillin in outpatient settings (children with indrawing pneumonia were treated in inpatient settings in the pre-March 2016 guidelines).

However, in African settings there is concern that there is still high mortality in children with indrawing pneumonia (32). Moreover, while amoxicillin and penicillin (injection) were equivalent in trials, no study had compared penicillin plus gentamicin versus monotherapy with either oral amoxicillin or penicillin injection for this group of indrawing pneumonia patients – (see subsection 1.1.4). The analysis in this thesis (as stated in sub – section 1.1.7) will aim to examine if there is any beneficial effect associated with using gentamicin plus penicillin as opposed to penicillin monotherapy which is regarded as a narrow spectrum antibiotic treatment. If the combination therapy is not advantageous this might allay fears of a need for broader spectrum antibiotics for indrawing pneumonia and clinicians would be encouraged to continue using amoxicillin (as equivalence with penicillin had been demonstrated) in outpatient settings. Conversely if a higher mortality is observed in non-trial populations of children with indrawing pneumonia and combination antibiotic therapy shows some benefit then the 2016 revision of pneumonia guidelines promoting use of antibiotic monotherapy and outpatient treatment might need to be reviewed.

Box 1.1.1: Pneumonia treatment algorithm

The pneumonia severity classification that was recommended by Kenyan guidelines up to March 2016 (28) (and previously by WHO guidelines (22)) defined the following three severity classes:

1. **Very severe pneumonia:** If a child had either oxygen saturation less than 90% or central cyanosis or was grunting or unable to drink or not alert, then s/he was classified as having very severe pneumonia, put on oxygen and treated with a combination of gentamicin and penicillin.

{The new WHO (33) and Kenyan guidelines (28) renamed this class as "severe pneumonia" – and currently recommend treatment with a combination of ampicillin (or penicillin) with gentamicin plus oxygen}.

2. **Severe pneumonia:** If a child had lower chest wall indrawing (but did not have any of qualifying signs for very severe pneumonia above) and was alert then s/he was to be classified as having severe pneumonia and be treated with benzyl penicillin only.

Note: The term indrawing pneumonia is hereafter used in this thesis to define this category of children to avoid confusion.

3. **(Non – severe) Pneumonia:** If a child had none of the mentioned signs but had cough or difficulty breathing and a respiratory rate greater than or equal to 50 breaths/minute (for age between 2 and 11 months) or respiratory rate greater than or equal to 40 breaths/minute (for age above 12 months) then s/he was classified as having non severe pneumonia and treated with cotrimoxazole or amoxicillin if previously treated with cotrimoxazole.

If a child has wheeze in addition to qualifying signs for any pneumonia severity, then this is treated as a possible case of asthma. Asthma is treated immediately with bronchodilators according to algorithms documented in the basic paediatric protocols and the child clinically reassessed after this treatment. If signs suggesting pneumonia persists, then treatment for wheeze is continued but this is in addition to antibiotic treatment – as pneumonia cannot be reliably ruled out if wheeze is present. Treatment of URTIs is not defined in the basic paediatric protocols but (non-antibiotic; non-drug)

symptomatic treatment only is recommended in training of health workers in Kenya (34). Because most initial treatment, even in Kenyan hospitals, is provided by clinicians with very little training in paediatric care and as there are no diagnostic tools to distinguish bacterial infections from other LRTIs (bronchiolitis or bronchitis), guidelines make no attempt to provide a clinical distinction between the possible causes of pneumonia / other LRTIs (bronchiolitis or bronchitis). Treatment is therefore based on severity of illness, assumed to be pneumonia if meeting the criteria above, and not on aetiology. Similarly, although wheeze is also found in viral LRTI, especially bronchiolitis, this is not distinguished from pneumonia/asthma in Kenyan guidelines.

1.1.5.2. Recommendations for diarrhoea/dehydration diagnosis and treatment

Dehydration in children with diarrhoea is classified on the basis of a set of key clinical signs as 'no', 'some', 'severe' dehydration, or 'shock' (see the diarrhoea/dehydration treatment algorithms (27) for specific signs used in classification) and these are treated with oral or intravenous fluids as defined in the protocols. Correct fluid prescription is defined as: oral rehydration (plan A) (without intravenous fluids or bolus) for children classified as having no dehydration; plan B oral rehydration (75ml/kg of oral fluid over 4 hours) for those with some dehydration and; prescription of plan C for rehydration for children with severe dehydration. Plan C treatment consists of step 1 (30ml/kg of fluid) and 2 (70ml/kg of fluid) administered over 30min or 1 hour and 2½ hours or 5 hours for infants and children aged over 1 year respectively. Those with shock are treated with bolus fluid followed by step 2 of plan C. In addition to fluid prescriptions, for all diarrhoea cases, children are supposed to be treated with oral Zinc supplement as a dispersible tablet.

1.1.6. The use of observational/routine datasets in comparative effectiveness evaluations

In order to minimise uncertainties about potential outcomes, harms and benefits of treatments derived from RCTs, it is also important to conduct effectiveness studies in the diverse patient populations that are potential recipients of the treatment. These may complement (or rarely replace) (35, 36) the results of RCTs aiming to assess efficacy, especially if such studies are in narrowly defined groups, by observing the impact of interventions in the real world (observational studies) across a range of patients, sites and settings.

Routine settings can provide an array of patients that may offer rich observational data on multiple end points, both long and short term, which could be important in evaluating benefits and adverse effects of a treatment. Such studies may be especially valuable if existing RCT data only used a surrogate end point for a more substantive health outcome. RCTs have two major phases; (a) design and (b) outcome analysis (where outcome analysis only happens after the design is fully set). Much effort and time is spent on the design of randomised experiments while corresponding design of observational studies has received less attention (37). The design of observational studies here entails everything that happens before examining or analysing outcome data (38). Compared to the highly controlled RCTs, observational studies require even more careful attention in their design since they are fraught with problems that could compromise their internal validity (39). In an effort to replicate the strengths of RCTs highlighted in sub – section 1.1.3, Rubin (2008) proposed guidelines for conducting credible comparative effectiveness analysis (to optimise

internal validity) using observational data (39). These include: i) clear conceptualisation of study question(s), ii) identification of decision makers or treatment guidelines which helps in explaining possible rules that were used in treatment assignment, iii) examination of the quality of key variables (for example in terms of completeness) that may have influenced treatment assignment, iv) verification of sample size, and v) use of statistical approaches to overcome non – random allocation of treatments and missing data (38, 39). These guidelines are now briefly explained.

1.1.6.1. Conceptualisation of research questions

Routine datasets may be re-imagined to come from complex trials whose treatment assignment rules have been lost and need to be reconstructed (39). At the start therefore, clear research questions with a specific design should be formulated with the help of experts to ensure analyses address a potential research gap.

1.1.6.2. Identification of treatment decisions and or decision makers

It is important to identify treatment decisions and or decision makers in order to understand how treatment assignment is conducted in routine settings[1]. Clinicians are supposed to use guidelines in diagnosis and treatment of a number of childhood illnesses in Kenya. These guidelines, largely adapted from WHO guidance, have been documented in the basic paediatric protocols booklet (28) based on available evidence and developed by consensus by national guideline panels (see (40-42)). In standard practice, the process of treatment

[1] This also helps in defining the target population through refined inclusion and exclusion criteria.

assignment happens in three steps; first, there is assessment and documentation of each clinical sign. Step two involves integration of clinical information into a severity classification for each disease (syndrome), and in step three severity classification is translated into a treatment assignment.

The ability to use routine data to compare treatment effects requires that patients with similar problems receive different treatments. Previous studies, conducted in Kenya and elsewhere, have indicated that clinicians often do not follow guideline recommendations in treating pneumonia and other illnesses (43). Variation from the guideline recommended approach can occur at the point of severity assignment (clinicians do not follow a nationally approved protocol linking clinical signs and severity category) and at the point of treatment assignment (clinicians do not follow this protocol that links treatment and severity). If the clinician records the clinical signs on which severity assignment is based, it is possible to construct a severity level which is guideline compliant using these observed signs in the dataset. This assigned / data derived severity level may be different to the clinician's choice of severity level. This variability in adherence to protocols and the ability to derive a protocol compliant severity level provide an opportunity for comparative effectiveness evaluation where patients with similar characteristics may receive different treatments. See sub – section 1.2.5 for detailed treatment assignment methods.

1.1.6.3. Recording and examining quality of key variables

As defined earlier, key variables are those pre-specified in Kenyan protocols that should determine illness severity and thus treatment assignment. Having all the key variables in routine datasets might be a challenge and if a key variable is missing then surrogate indicators (variables which may be used as proxies for or in place of *key variables*) should be used. Also, surrogate indicators should be used for key variables that are not measured adequately. A major focus of the Clinical Information Network on which this work draws has been to improve the documentation of important key, auxiliary and surrogate clinical variables by clinicians.

1.1.6.4. Sample Size Verification

Usually in RCTs, sample size calculation is done before commencement of the study, while in routine settings data already exist in databases but it is important to verify whether the existing data will be sufficient for any meaningful analyses. This includes examining minimum detectable effect size alongside risk of type 1 and 2 errors. It is important to note that sometimes if the sample size is not sufficient and the dataset to be analysed is the only one to answer an important question – then one should proceed though efforts should be made to generate better data (39). Also in these circumstances, one needs to estimate the size of effect that the small sample is able to detect.

1.1.6.5. Overcoming non – random treatment allocation and missing data

As observational/routine datasets are often not primarily meant for research, they face two analytic challenges: non – random allocation of treatments and missing data. In routine clinical practice, patient allocation to alternative treatments for the same illness may be conditioned on varying characteristics. So, clinicians may override the use of key variables that should determine severity and treatment allocation and may be influenced by observed auxiliary variables, perceived to be serious, in assigning treatments (see sub- section 1.2.5 for detailed discussion on non – guideline treatment assignment). This may result in differences in characteristics between any two or more treatment groups. To examine differences in treatment outcome(s) there is therefore a need to make groups comparable or 'enforce' covariate balance prior to comparative analysis (44). The ideal approach is to compare individuals with exactly similar values on observed covariates. However, this approach may be inefficient when dealing with a larger number of covariates as it may be difficult to find sufficient individuals with exactly similar covariate values in treatment groups (37). An alternative that is increasingly being used in the last decade (across many disciplines) is the propensity score (PS) approach (45, 46) developed by Rosenbaum and Rubin (1983) (47). Here, the PS describes the probability of a patient's assignment into a treatment group given the observed covariates. Outcomes for patients with similar propensity scores (between treatment groups) are then comparable as it is assumed they have a similar distribution of observed covariates. However, as propensity scores are derived from prediction models, they may be inefficient when covariate data are missing, as the effective sample

size is likely to be reduced. See Chapter 2 (sub – sections 2.1) for brief discussions of PS and missing data methods.

1.1.7. Complementing RCTs with observational comparative effectiveness evaluations.

More generally, RCTs and observational studies should be seen as complementary rather than competing approaches (35, 36). We may benefit from both as effectiveness evaluation should be perceived as a continuous process right from RCT experiments in homogenous to more heterogeneous real life populations (**Figure 1.1.3**). Once treatments are found to be efficacious in RCTs, they are then studied (to further examine their efficacy and toxicity) using larger populations for a slightly longer period as part of initial effectiveness surveillance. Beyond this then treatments are allowed to be used in routine care for the general populations. These populations constitute everybody (including those that may have formed part of exclusion criteria in RCTs). As data from routine care are consolidated in healthcare databases, it is therefore important to examine the effectiveness for everybody taking these treatments using these observational data.

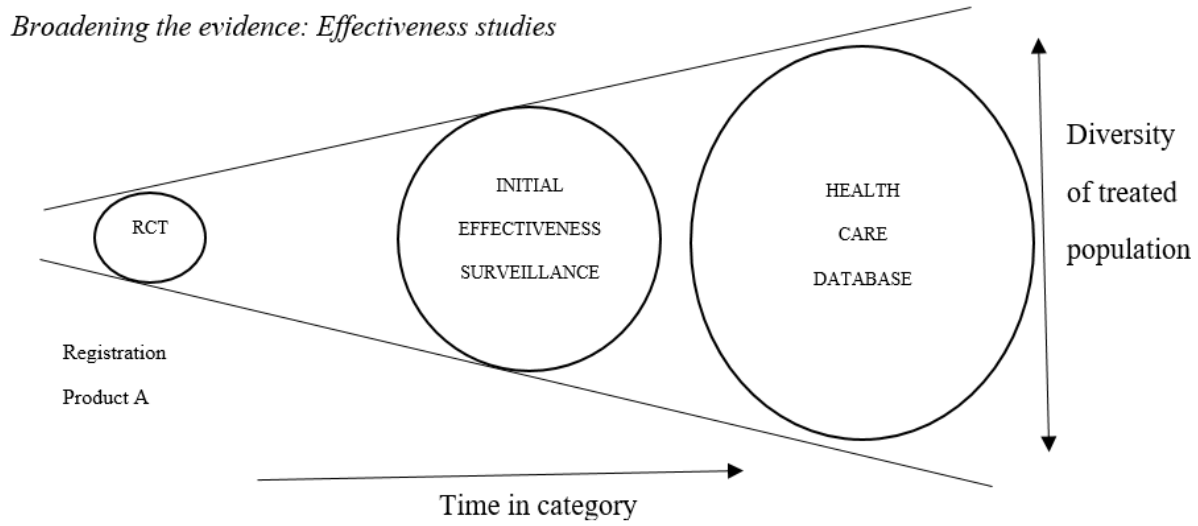


Figure 1.1.3: Treatment effectiveness evolution, adapted from Lindsay (2007) (36).

In summary (see **Table 1.1.1**), RCTs are designed with internal validity in mind and often use carefully defined populations which are randomly assigned into treatment groups – in which randomisation (depending on the quality of process, sample size and the number of important covariables) balances on both measured and unmeasured variables. Such a focus on internal validity makes them the best method to establish biological efficacy. However, in optimising internal validity in RCTs, generalisability of findings can be restricted to carefully selected populations, often those that maximise the likelihood of establishing biological efficacy. Additional RCT limitations may include limited sample size (often due to costs of trials) and the use of short term end points. These limitations in RCTs may be less apparent in observational studies. The latter can use real world populations, may be cheaper to conduct and thus may have larger sample sizes, and may provide opportunities to test a range of important outcome measures. However, there are two main limitations

associated with the use of observational datasets: (a) non – random allocation (and missing data) which have been discussed in sub-section 1.1.6 and (b) possible influence on results from unobserved variables. Sensitivity methods used to examine the latter are discussed in Chapter 2 (sub-section 2.1.1).

Table 1.1.1: Summary of strengths and limitations of RCTs and observational comparative effectiveness studies.

Randomised controlled trials	
Strengths	*Limitations*
Designed with internal validity in mind	Can results in poor generalizability.
Use carefully defined study population.	Excludes a number of patients requiring treatment.
Balance both measured and unmeasured covariates through randomisation.	Can use limited sample size (and sub-group analysis problematic)
	May use short term end points thus may not allow for long term adverse events to be observed.
Observational comparative effectiveness studies	
Strengths	*Limitations*
Uses real world patients.	Non – randomised in nature thus highly vulnerable to selection / allocation bias.
Appropriate for assessing long term adverse events or rare outcomes.	Data limited to variables in the dataset.

As the analysis will be based on data from the Clinical Information Network, the next section (section 1.2) of this chapter provides a description of the Kenyan Health Care System, introduces the CIN, explains treatment assignment methods, explores causes of admission and mortality in the CIN hospitals, and examines the completeness of clinicians' documentation of key variables used in pneumonia and dehydration diagnoses.

1.2. Data description

1.2.1. Kenyan Health Care System

The Kenyan Health Care System comprises private and public systems. The private sector includes non-governmental, faith based and private for profit organisations, while the public sector includes parastatal organisations and the Ministry of Health (MOH). There are over 5000 facilities through which health services are offered in Kenya and about 51% of these are managed by the public sector (48). Further, the Kenya Service Availability and Readiness Assessment Mapping (SARAM) 2013 report noted that most inpatient units were government owned especially outside major cities (49).

According to the Kenya Health Sector Referral Implementation Guidelines (KHSRIG) 2014 (50), the MOH and stakeholders adopt a tiered system for health facilities (**Figure 1.2.1**) in order to make health services accessible to as many people as possible. This structure consists of community health services (level 1), primary healthcare services (levels 2 and 3), county referral hospitals (levels 4 and 5) and national health referral services (level 6). Hospitals in upper levels act as referral facilities for those in lower levels. Also, Kenya is subdivided into 47 administrative counties and each county has a referral hospital which is either a level 4 or 5 facility. This decentralised system enables access of health services at the county level.

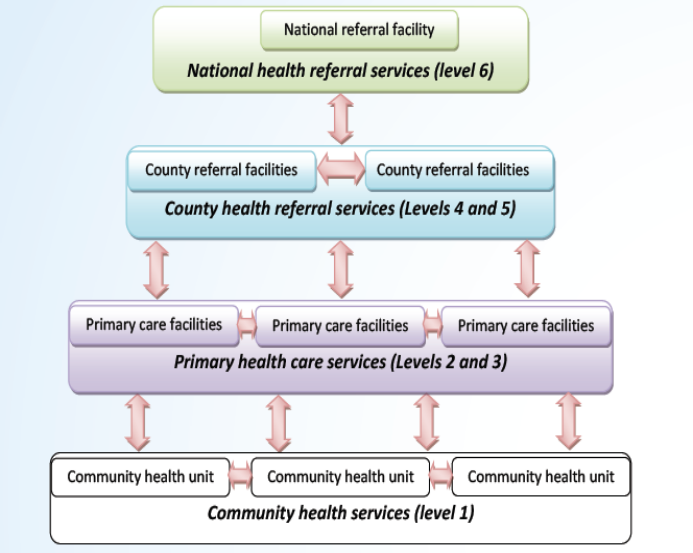


Figure 1.2.1: Referral linkages, adapted from the Kenya Health Sector Referral Implementation Guidelines (2014)

1.2.2. Background to the Clinical Information Network

Thirteen county referral hospitals (at Level 4 and 5 in the health system) plus one sub – county hospital were purposively selected with direction from MOH and recruited into the Clinical Information Network (CIN) (**Figure 1.2.2**), which represents collaborative work between the MOH, the Kenya Medical Research Institute (KEMRI)/Wellcome Trust Research Programme, the Kenya Pediatric Association (KPA) and participating hospitals. These hospitals were recruited into the study at different times; Nairobi sites in September 2013, Western sites in October 2013 and Central sites in February 2014 (Nairobi, Western and Central are three blocks in which the hospitals have been categorized).

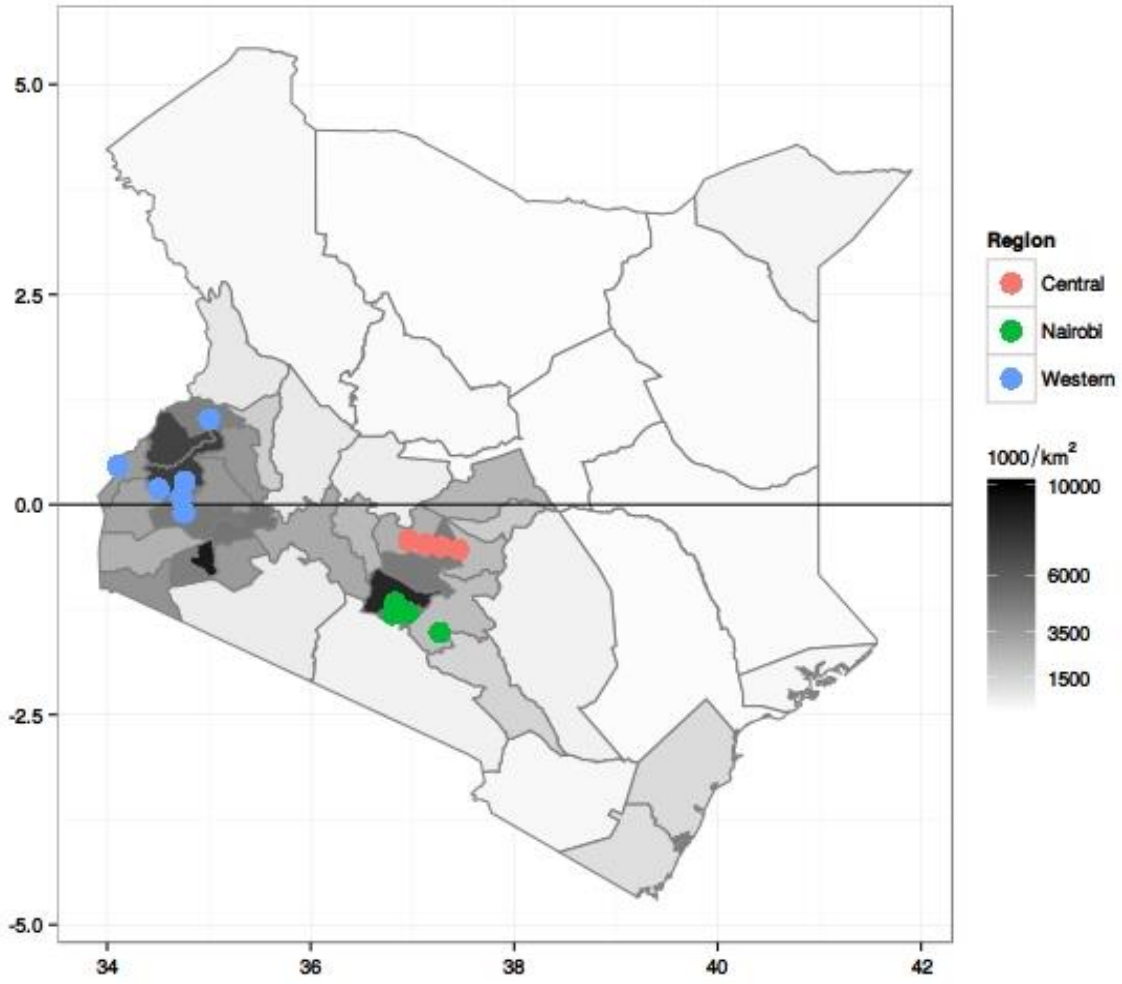


Figure 1.2.2: Distribution of CIN hospitals across Kenya (the shaded regions represent population density)

The Kenyan Ministry of Health has made efforts to foster improved medical record keeping in clinical areas by providing and promoting paediatric and newborn admission record forms. These efforts are linked to recent research that has shown that Kenyan hospitals have challenges in providing quality data on inpatient workload, mortality and morbidity (51). Additionally, hospital data are not often suitable or not used within hospitals to monitor quality of care provided to patients (51, 52). Thus, the CIN was initiated to improve data availability

from secondary care in paediatrics and as a model for demonstrating the value of routine data in improving quality of care in the hospitals.

1.2.3. Data Collection and Tools

Data are collected retrospectively post discharge by trained data clerks, guided by well-defined standard operating procedures, under close supervision by the hospital medical records department and Kenya Medical Research Institute/Wellcome Trust Team. It is worth noting that the research team has no personnel checking quality of clinical process and whether clinicians correctly document what they do. However, the patient record is the formal (and legal) document describing the clinical condition and management. These documents are used for data abstraction and they include patient files with Paediatric Admission Record (PAR) forms (see **Appendix A.2**), treatment sheets, nursing cardex, discharge summary forms, laboratory reports and clinician notes. CIN study uses the Research Electronic Data Capture (REDcap) platform, which is an open source technology, to collect data. The collected data are used to assess documentation of history, physical examination, diagnosis, laboratory investigations, treatment and discharge plans.

The total possible variables collected per record are 382 and are grouped into; biodata, history, examination, investigations, admission diagnosis, treatment, supportive care, monitoring and discharge information tools. However, many variables are collected for only specific sub-groups (e.g. linked to diagnoses) or only if a child receives specific treatments. The total number of variables for which data are collected per patient is therefore well below this maximal figure of 382. The biodata tool contains basic patient's demographic information;

history and examination has information on clinical signs and symptoms; investigations has laboratory test orders and results; and the discharge information tool contains discharge diagnoses, outcomes and follow up information.

To prevent any backlog in data entry in very busy hospitals a more limited set of data (referred to as minimum data) are collected for children aged < 1 month and for non – medical conditions such as burns or surgical cases to whom common guidelines do not apply. The research team implemented a module in REDCap for 'minimum data' cases such that 82 variables (on biodata, investigations, diagnoses and discharge information tools) are captured for a proportion of patients. This minimum dataset was also captured on a proportion of patients – usually between 30% and 65% and randomly selected - depending on the workload in two CIN hospitals with particularly high rates of admission.

1.2.4. Data Quality Assurance

Quality is defined in terms of data completeness and consistency within valid ranges. The KEMRI/Wellcome Trust team have adopted three strategies in ensuring high quality data (summarised in **Figure 1.2.3**). First, all data entry fields have pre-specified valid entries / valid ranges that result in error alerts to the clerk should they enter an unacceptable / out of range value. As some out of range values may nonetheless reflect what is documented some data entry checks use 'soft-validation only'. Second, cleaning algorithms written in R statistical programming language are used by every clerk daily to check for potential entry errors (eg. missing values) and if data are within correct ranges. The 'cleaning script' produces a list of any identified inconsistencies such that

the clerk can correct entries where possible. The third procedure is performed weekly for consolidated data that has been submitted to the central data management team in the main server with feedback on errors by telephone by a data manager. After the clerk re-reviews case records to make corrections where possible the amended data are resubmitted. In addition, there is periodic use of external quality control and audit where the KEMRI team visits facilities approximately 3 monthly and independently enters a sample of files already entered by the clerks. Thereafter, concordance between these same independently entered patient records is examined. The overall concordance correlation for all examined records so far is above 95%.

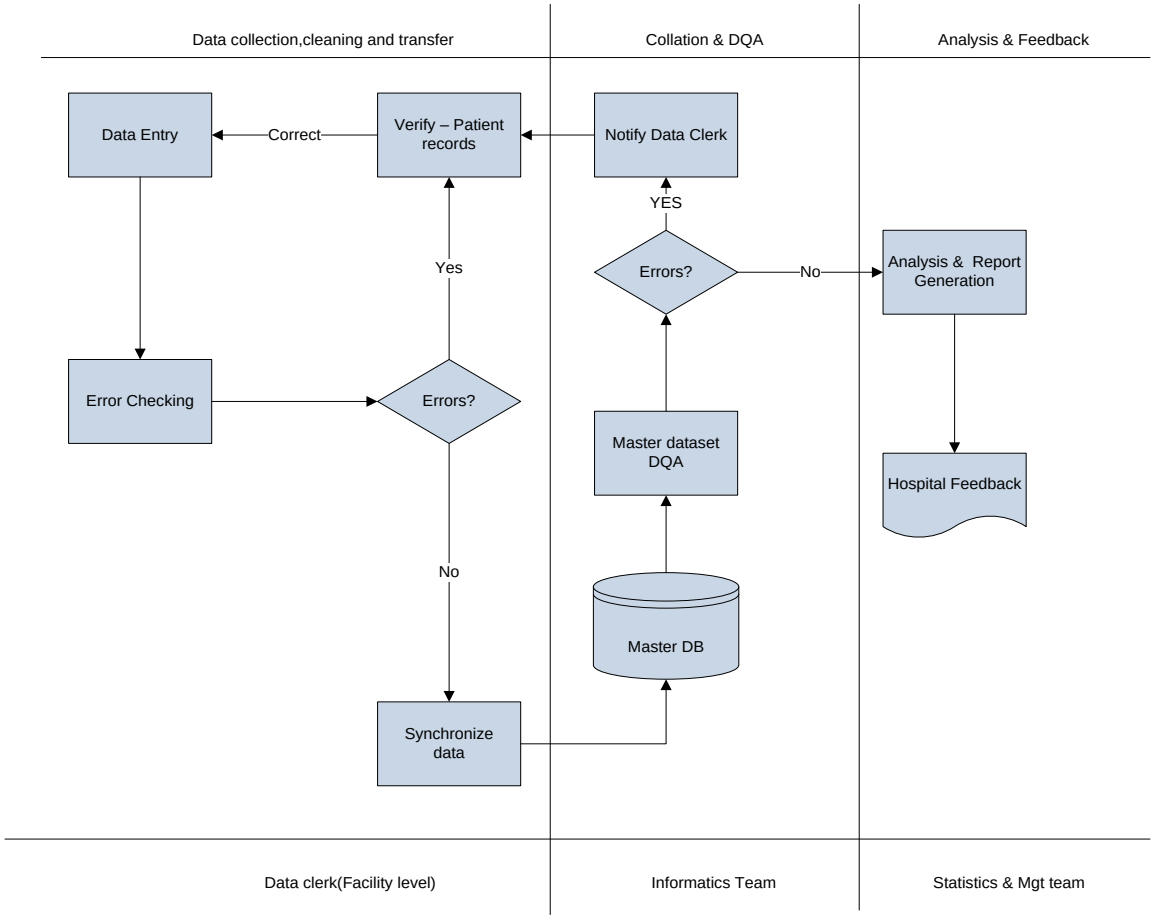


Figure 1.2.3: CIN data management process.

1.2.5. Treatment assignment methods

Kenyan clinicians working in county referral hospitals are provided with clear, nationally agreed guidelines/protocols for clinical management of common illnesses (27) – also see sub – section 1.1.5 above (for guideline recommendations for pneumonia and diarrhoea/dehydration treatment). The guideline/protocol treatment approach assumes clinicians will follow the rules – as guidelines are based on what is considered the best available evidence; thus the preferred clinical management approach can be regarded as a rational, rule based treatment assignment strategy. This strategy recognises that the clinicians attending to patients at admission are typically very junior (usually they are pre-registration interns) having had only 8 – 10 weeks total training in paediatric care. However, in routine settings, there may be guideline non – compliance due to a number of reasons. These may include; 1) clinician judgement/gut feeling based on aspects of a patient's history or appearance that are observed but not recorded (or codified into any rules), 2) explicit clinical findings that may influence clinical decisions that are recorded as part of the clinical evaluation process but are not captured in standard treatment rules, and 3) clinical mistakes. These three factors may all result in what appear to be errors in any of assessment, severity classification or treatment assignment when compared with the rule-based standard guidelines. Such influences on treatment assignment are important to consider carefully as Agweyu et. al (2014) showed that there was substantial non-compliance with guidelines (pneumonia severity misclassifications and incorrect treatment) for antibiotic treatment assignment

in a small sample of pneumonia patients at Kenyatta National Hospital in Kenya (43). And this has been seen several times over the years (53, 54).

As the two clinical comparative effectiveness questions tackled in this thesis are based on variations in practice linked to adherence / non-adherence to specific clinical guidelines, the three aspects of guideline non – compliance are introduced and discussed in detail.

1.2.5.1. Gut feeling

Gut feeling is perceived to be an intuitive judgement based on a patient's appearance or elements of the history of reported illness or context (55). Thus clinicians take note of subtle clues that influence their judgement either of the severity of the illness or their treatment selection. Quite often these clues or influences are unmeasurable. Stopler (2009) identified two categories of gut feeling that may result in alarm or reassurance. In the case of alarm, the clinician may not be sure of prognosis and therapy – there is uncertainty – and they may suspect that signs of risk are yet to develop. This may guide them, without explicit justification, to classify a patient as having more severe illness and step up treatment. Second, is a sense of reassurance where a clinician may, for example, have seen a similar case previously and is sure of a benign prognosis and therapy (56). This may result in a step down in treatment. An observational study by Van den Bruel (2010) asked clinicians to record their gut feeling (plus signs/symptoms) in a summary risk category of absent, present or unsure – which was to indicate clinical impression of something serious. In subsequent analyses this categorical description of gut feeling (which also presumably represented a summary of formally recorded clinical signs) had

greater diagnostic value than signs/symptoms alone (57). A typical example in which gut feeling may influence clinician diagnostic thinking is a situation where a patient may be put on a different treatment (or higher dose) at admission if standard treatment had been administered elsewhere (if known to the clinician) before admission. Here, one may understand the pre – admission treatment as having 'failed' and may think administering alternative treatment (or a higher dose) would be effective. If such pre – admission treatment is known to the clinician but not recorded anywhere, then it may influence treatment (and contribute to the broad notion of gut feeling). It is clearly difficult to examine these hidden influences on admission treatment assignment. However, if the pre-treatment had been recorded, then in theory its influence can be examined, in this case it would no longer be a hidden influence captured in the concept of gut feeling.

Gut feeling resulting in alarm may naturally contribute to systematic errors in classifications if consistently experienced by clinicians in practice. In this situation, patients may coincidentally benefit from treatment step up due to a clinical impression of something serious. Otherwise, if a clinician experiences a reassuring "gut feeling" in serious situations then in the worst case this may consistently risk lives of patients by systematically assigning lower doses or inferior treatments.

1.2.5.2. Explicit clinical findings

According to guidelines, *key variables* should determine illness severity and thus treatment assignment – and in the case of pneumonia and diarrhoea/dehydration, these are defined by the protocols (see sub – section

1.1.5). However, there may be additional (measurable or visible) symptoms and signs that are not formally recognised as key variables in treatment guidelines that may influence clinician diagnostic thinking. For instance, if a patient has had a convulsion or has been ill for longer than is felt typical then s/he may be felt to have a more severe illness than is suggested on the basis of the signs of the pneumonia alone. This may result in a step up in treatment with a child put on higher doses or 'stronger' treatment at admission. These factors may be considered *auxiliary* if they are recorded (measured). Their influence on treatment assignment can be examined although according to the formal rules (the guidelines) they are not considered reasons to alter treatment assignment. A number of studies (for example (53, 58)) have shown significant associations between inpatient mortality and a wide variety of clinical variables – it is reasonable to hypothesise therefore that these may influence treatment assignment in addition to key pneumonia variables. Many of these additional or auxiliary variables are or should be documented in routine practice (at admission) within the clinical information network. It is therefore possible to explore whether they have any link to treatment assignment. Variables that are related to both treatment assignment and patient outcome should be taken account of in any examination of whether treatment itself influences outcome. Consistent observation of explicit clinical signs (not part of the guidelines) may lead a clinician to always assign higher severity levels – resulting in systematic errors in classification thus always biasing treatment choice(s).

1.2.5.3. Unintentional clinical mistakes

Lack of expertise may cause random or systematic errors in clinical assessment, disease severity classification or treatment allocation. Patient history and examination are critical in clinical assessment, and they may be consistently but inexpertly evaluated by a clinician resulting in systematic severity misclassification (always making the same mistake resulting in step up/ step down in treatment). Also, a clinician may have assessed signs/symptoms correctly but ends up recording the wrong classification, something equivalent to a random error – which later influences the choice of treatment. Further, errors of understanding may lead to errors of transcription especially in the process of integrating clinical findings to follow the rules – and may be habitual to a clinician resulting in frequent undetectable errors.

I have now described the setting, the data collection process and the protocols that are intended to guide care. I have also suggested why clinicians may not follow these protocols introducing variation into treatment assignment. In Chapter 3 I present analysis on antibiotic effectiveness in treating pneumonia based on data up to March 2016 before national guidelines for pneumonia changed. In Chapter 5 I present analysis on Zinc effectiveness in treating diarrhoea based on collected data up to February 2017. However, in order to describe the admission and mortality patterns in CIN hospitals to illustrate the context in which these analyses take place I use data up to March 2016 (from October 2013) in sub – section 1.2.6.

1.2.6. Exploratory analysis

1.2.6.1. Major causes of admission

Focusing on common conditions in this description of paediatric admissions to the Clinical Information Network, Pareto and Venn plots are used to describe admission diagnoses (likely more than number of patients due to comorbidity) and patient level events (mortality) respectively. Pneumonia, malaria, dehydration/diarrhoea, meningitis, anaemia, malnutrition, febrile convulsions, URTI, asthma and convulsive disorders account for about 83% of total admission diagnoses (**Figure 1.2.4**). A significant proportion of these admission diagnoses (57%) is accounted for by the top three illnesses i.e. pneumonia, malaria and dehydration/diarrhoea; with pneumonia being most prevalent overall although this varies widely across places: Nairobi (38%), Western (19%) and Central (41%).

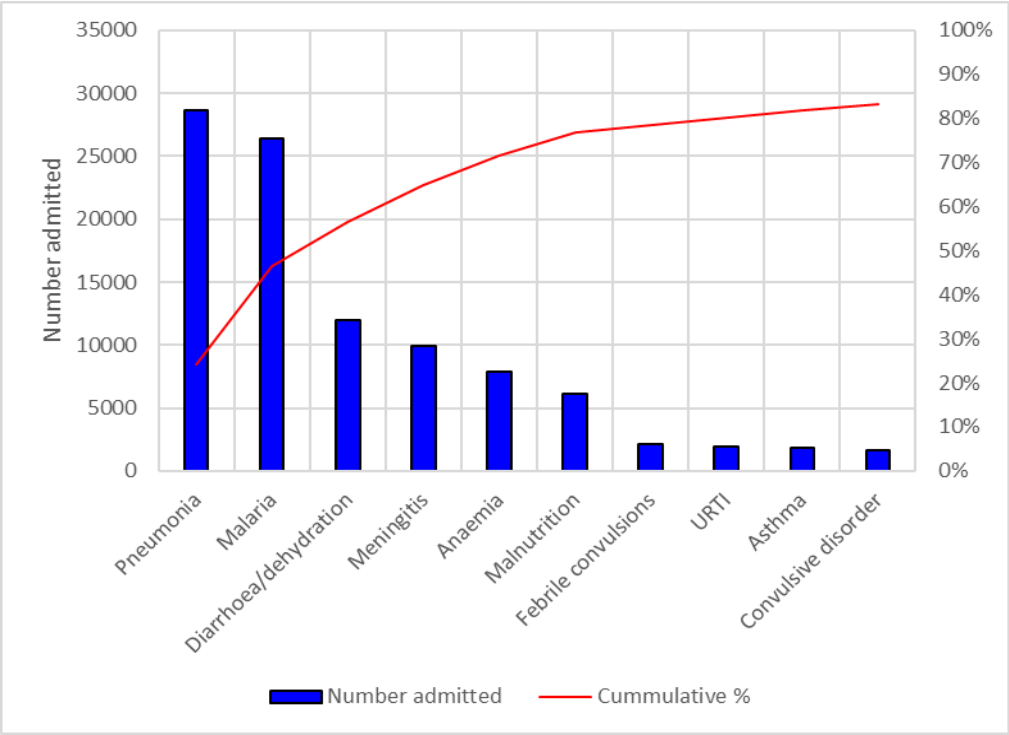


Figure 1.2.4: Overall top causes of admissions.

1.2.6.2. Major causes of mortality

Figure 1.2.5 presents patient level mortality statistics and it shows that there were 4983 (6.4%) deaths in a population of 78 168 admissions in the full dataset (considered up to March 2016). The top four causes of mortality were determined based on the number of episodes (in which a patient may have died from multiple causes). Of the 4983 deaths, there were 2169 pneumonia episodes, followed by diarrhoea/dehydration episodes (n = 1571), then malaria episodes (n = 1310) and lastly malnutrition episodes (n = 884). Together with their comorbidities they were associated with about 76% of child deaths. The other 24% of deaths were associated with diseases other than pneumonia, malaria, dehydration/diarrhoea and malnutrition.

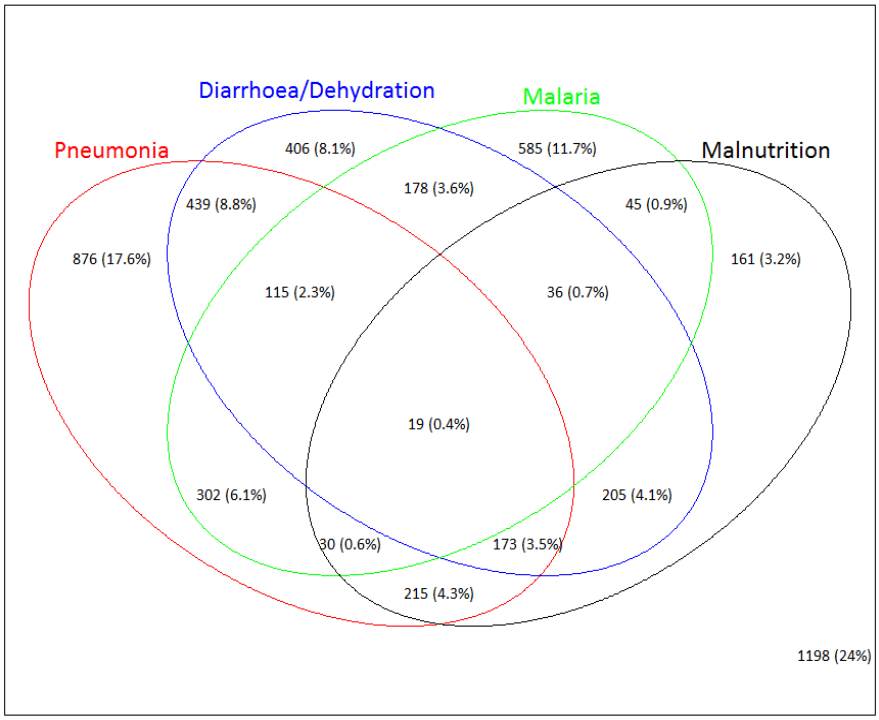


Figure 1.2.5: Mortality patterns: intersections represent the number of mortality causes – for instance 19 (0.4%) of the children who died had pneumonia, diarrhoea/dehydration, malaria and malnutrition.

1.2.6.3. Diagnosis and treatment

1.2.6.3.1. Treatment of Pneumonia

The pre March 2016 national guidelines (and as presented in **Box 1.1.1** above, sub – section 1.1.5.1) only applied to sick children without co-morbidities of either severe malnutrition, HIV, meningitis or indication of cough for more than 14 days. Antibiotic treatment recommendations for these diagnoses take precedence over and vary from those from pneumonia. In cases where pneumonia is the primary rationale for antibiotic treatment and if guidelines were correctly followed, children with very severe pneumonia were to be treated with gentamicin plus penicillin, those with severe (indrawing) pneumonia treated with penicillin alone, and those with non – severe pneumonia treated with amoxicillin. I explored how children with various severity levels of pneumonia (as assigned by clinicians) were treated in the CIN hospitals and present the results in **Figure 1.2.6**. These results demonstrate that clinicians do not adhere to guideline recommendations in treating children even if their own classifications overlap with those of the guidelines. This demonstrates variation in treatment allocation across children with similar recorded clinical symptoms and signs. Therefore, this provides the opportunity for examining if differences in outcomes are associated with differences in treatment allocation provided treatment groups can be made as similar as possible in terms of clinical characteristics. See Chapter 3 (sub – section 3.2.3) for exploratory analyses of treatment changes during admission.

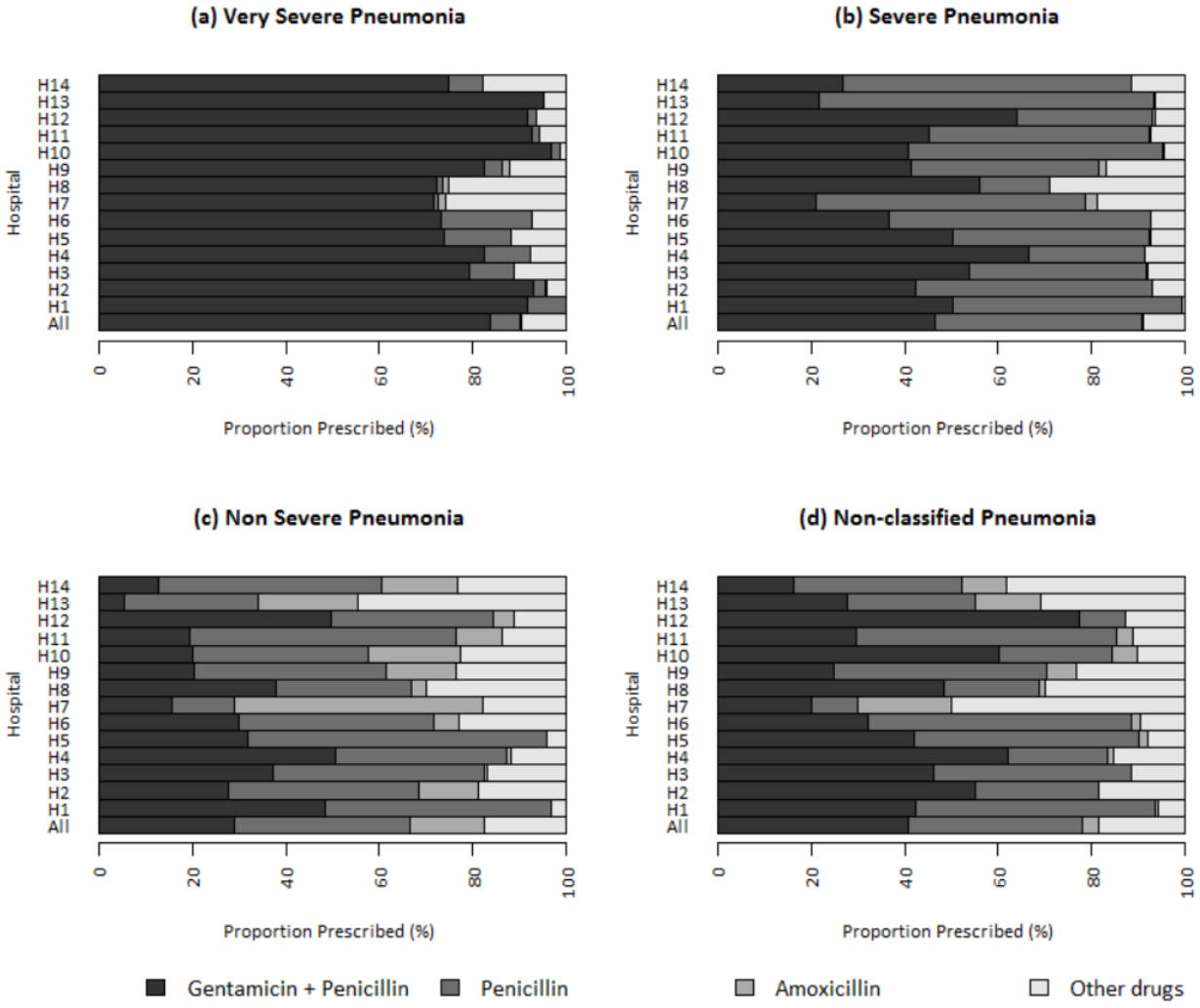


Figure 1.2.6: Prescription of common antibiotic treatment for pneumonia according to severity levels assigned by clinicians.

1.2.6.3.2. Treatment of diarrhoea/dehydration

Adherence to guidelines in oral/iv fluid prescription was examined only for children who also had dehydration as a diagnosis (diarrhoea may be a symptom of other febrile illnesses e.g. malaria and not associated with dehydration). Overall, 8667 out of 12003 children with diarrhoea/dehydration admissions had accompanying dehydration (no dehydration (n = 1730), some dehydration (n = 4320), severe dehydration (n = 2229), and shock (n = 388)). All of these children

should have received Zinc therapy. The number of children who were given a correct prescription of oral/iv fluid regimens as outlined in the guidelines was 5322/8667 (61%).

1.2.6.4. Completeness of key variables

Key variable data are needed to assign a correct severity level in accordance with guidelines and have been shown to be associated with outcome. Thus, these data are needed both to assign severity level as an analyst and also to check group balance before comparative analysis. Thus this sub – section explores the completeness of key variables used in diagnosis of both pneumonia and diarrhoea/dehydration. In case of incompleteness, there will be need to fill in plausible values using multiple imputation. Multiple imputation methods fill in missing values multiple times (59). See a full introduction to these methods in Chapter 2 sub – section 2.1.

1.2.6.4.1. Pneumonia

Respiratory rate, indrawing, central cyanosis, grunting (children < 1 year), pulse oximetry values, ability to drink and AVPU are some of the essential indicators for pneumonia diagnosis (**Box 1.1.1**). Pulse oximetry values are however poorly documented in CIN hospitals; this is largely attributable to lack of pulse oximeters which the CIN project is unable to provide. Central cyanosis is a clinical sign whose overall completeness is 95% it is therefore, used as a surrogate indicator of hypoxemia (**Table 1.2.1**). The overall documentation of respiratory rate, indrawing, AVPU and ability to drink is above 85%.

Table 1.2.1: Completeness % (key pneumonia variables)

	All	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14
N	**10670**	**965**	**835**	**907**	**1177**	**735**	**571**	**643**	**397**	**966**	**378**	**677**	**1183**	**200**	**1038**
Respiratory Rate	87	86	80	84	85	93	49	98	74	64	80	89	90	71	92
Indrawing	96	99	100	95	99	94	91	96	65	82	78	97	97	87	97
Central Cyanosis	95	98	99	95	99	94	77	98	98	94	86	96	99	91	95
Pulse oximetry values	21	0	0	73	30	1	0	58	26	0	0	1	31	0	34
AVPU	96	95	99	94	98	93	94	96	90	96	92	96	96	99	95
Ability to Drink	91	93	98	92	95	90	88	85	58	80	73	99	99	83	98
Grunting (<1 year)	94	92	95	95	97	92	83	91	59	73	69	96	92	73	90

1.2.6.4.2. Diarrhoea/dehydration

Skin pinch, sunken eyes, capillary refill, AVPU and ability to drink are considered key variables for measuring dehydration severity. Overall, completeness for each of them is above 80% (**Table 1.2.2**).

Table 1.2.2: Completeness % (key diarrhoea/dehydration variables)

	All	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14
Skin pinch	90	95	84	96	93	90	82	91	47	64	53	93	97	42	96
Sunken eyes	92	90	76	95	92	96	37	95	69	77	61	94	94	46	91
Capillary refill	80	92	81	95	93	96	61	83	26	43	39	77	95	33	91
AVPU	89	93	87	94	91	96	83	91	74	88	59	92	98	31	91
Ability to drink	87	95	91	92	91	93	82	89	46	60	45	92	91	45	89

The completeness for auxiliary variables (potentially important in understanding patient risk although not influence severity classification according to guidelines) is presented in Chapters 3 and 5.

In section 1.1, I have explained the persistent high burden of mortality and major contribution of pneumonia and diarrhoea to this in Sub – Saharan Africa. Also, I have discussed the limitations associated with RCTs, provided the current evidence for pneumonia antibiotic treatment and Zinc co-treatment for diarrhoea, and the potential complementary role of observational analyses if studies are well designed and conducted. In section 1.2, I have described data used for analysis in this thesis. I now summarise the thesis objectives in section 1.3, work that addresses some important gaps in the child health evidence base linked to treatment of pneumonia and diarrhoea.

1.3. Thesis Objective

The objective of this thesis is to examine the emerging use of propensity score (PS) methods to support observational comparative effectiveness analyses and then apply the most appropriate methods to address topical and clinically important pneumonia and diarrhoea treatment questions. This is based on the literature gaps explained in sub – sections 1.1.4.1 and 1.1.4.2. To do this I will address challenging methodological issues of non – random allocation of treatments and missing data (discussed in sub – section 1.1.6) using a large routine Kenyan dataset as an example of those that might become available from other Lower and Middle Income Countries (LMIC) in the future. In doing this, I explore the potential value of and challenges with such data for examining the many unanswered questions about treatment of common childhood illnesses in African children.

In particular, PS methods are explored and then applied to address the comparative effectiveness of alternative antibiotic treatments for pneumonia and

the use of Zinc as co-treatment for childhood diarrhoea, focusing on questions that have been relatively neglected in formal RCTs.

Data from the Clinical Information Network (CIN, introduced below) are used for these analyses (a data collection platform I helped to initiate).

The specific objectives of the work encompassed in this thesis include to:

I. Review literature and describe the use of PS and missing data methods to minimise bias in observational comparative effectiveness evaluations.

II. Explore how to take guiding principles for credible comparative effectiveness evaluation using routine datasets and propensity score methods into practice using the CIN dataset to limit bias in patient allocation and to overcome the challenge of missing covariate data that may be important in achieving balance in allocation.

III. Use the methodological framework in II to tackle candidate comparative effectiveness questions for treatments used in pneumonia and diarrhoea illnesses for hospitalised children – specifically to:

    a) Compare the use of penicillin monotherapy vs. the use of penicillin plus gentamicin in treatment of (severe) pneumonia.

    b) Conduct sensitivity analyses and examine persistence of bias due to unmeasured confounders using alternative PS methods

and instrumental variables (see Chapter 2 sub – section 2.1.1 for an introduction to this method) on antibiotic comparative effectiveness question in (a).

c) Examine the effectiveness of Zinc administration to children with diarrhoea/dehydration using the optimised methods developed in parts IIIa and IIIb.

In addressing the specific clinical questions that are of current importance for clinical practice both analyses raise issues linked to potential bias in treatment allocation. The pneumonia analysis focusses on overcoming possible bias that could occur if patients who are felt to be more ill are prescribed broader spectrum antibiotic treatment (often felt to be more powerful). This may result in imbalanced covariate distribution for groups allocated different treatments. In this example, factors that are undocumented such as "gut feeling" (described in more detail in sub – section 1.2.5.1) may also play a role in skewing treatment allocation such that children who appear to be more severely ill are more commonly allocated to one of the treatment arms (potentially resulting in a bias suggesting this treatment is less effective). In the second clinical question tackled, Zinc for diarrhoea, covariate distribution may also be associated with treatment allocation. However, in this case, as Zinc is an oral medication, children in receipt of this intervention may be perceived to be less ill (more capable of tolerating an oral medication). In both cases I will explore the balancing effect of employing propensity scores, the comparative effectiveness of interventions, and the plausible persistence of bias by conducting sensitivity

analyses using instrumental variables. Finally, I will discuss the implication of using observational data for comparative effectiveness in low income countries.

Having now provided a background to this work, and described the hospital population on which these analyses are conducted and the availability of key data together with the thesis objectives, I provide a brief explanation for each of the subsequent chapters in section 1.4.

1.4. Overview of thesis chapters

This thesis is structured to have six chapters. Chapter 1 highlights the background to this work – which leads to the thesis objectives. Chapters 2, 3, 4 and 5 describe the main results of the thesis. And Chapter 6 discusses the main findings. The contents of Chapters 2 – 6 are now briefly described to provide an outline of the methods used and findings of the thesis:

Chapter 2: The use of propensity score and missing data methods in observational comparative effectiveness evaluations: a systematic review of methods

This review was needed because prior reviews had focussed on reporting elements of PS methods and ignored how missing data are addressed with these PS methods. Therefore, objectives for this systematic review were to:

i) Describe the use of Propensity Scores (PS) focusing on the methodological aspects of their application in the general clinical literature. This review was pertinent as the latest reviews on PS had focused only on acute care and cancer literatures.

ii) Assess the methods for handling missing data in estimation of PS and if the methods used are in line with the STROBE guidelines (60). The STROBE guidelines make recommendations for how missing data should be reported in observational studies and in particular, they require researchers to report on the proportion of missing data, reasons why data are missing and how missing data are addressed.

Chapter 3: Using observational data to compare the effectiveness of antibiotic treatments for children hospitalised with pneumonia in Kenya

This chapter is sub-divided into two parts:

Part 1: Analysis protocol

- Provides the background to the design for a comparative analysis of penicillin vs. penicillin plus gentamicin (based on guiding principles explained in **Chapter 1**) in the treatment of inpatient pneumonia.
- Explains analysis steps to be undertaken after estimating propensity scores on multiply imputed datasets.

Part 2: Results

- Describes the results of the analyses using the clinical information network data based on the protocol in part 1.

Chapter 4: Sensitivity analyses on multiple imputation strategies and unmeasured confounders in comparing effectiveness of penicillin monotherapy and penicillin plus gentamicin for pneumonia treatment.

I explore the influence of alternative analytic approaches by extending work comparing the effectiveness of penicillin monotherapy and penicillin plus

gentamicin on outcomes of treatment for pneumonia characterised with indrawing.

Comparative effectiveness studies based on routine datasets (using PS methods) assume that the analysis is limited only to observed variables. However, patient populations may differ in terms of unobserved variables. Therefore, this chapter explores sensitivity analysis methods such as the use of instrumental variables, and PS trimming to investigate how sensitive the derived treatment effect estimates in Chapter 3 are to unobserved variables.

Also, the underlying assumption for multiple imputation is that data are missing at random –validity of this assumption is investigated through the use of pattern mixture models (61). A further sensitivity analysis approach will be to include outcome data in the multiple imputation models. This is because comparative effectiveness evaluations are perceived to have two phases: design and outcome analysis – and it is advisable to hide outcome data in the design phase (39). This by extension implies that outcome data are not used for imputation in the design phase and therefore the sensitivity approach examines the effect of including outcome data in the multiple imputation models.

Chapter 5: Examining the effectiveness of Zinc treatment in children admitted with diarrhoea in Kenya's public hospitals: An observational study

This chapter provides background to the design for the comparative analysis of Zinc use vs. no Zinc use in children hospitalised with diarrhoea (based on guiding principles explained in Chapter 1). It explains the analysis steps to be undertaken after estimating propensity scores on multiply imputed datasets (using the methods that perform best in the pneumonia analysis from Chapter 3

and 4), and presents the effect on outcomes of Zinc treatment for diarrhoea amongst inpatient children. The outcome analysed is a time to event measure (time to discharge from hospital) using competing risk models. Importantly mortality is treated as a competing risk as it would preclude the chance of being discharged alive.

Chapter 6: Discussion

This chapter summarises the key findings of the results presented in Chapters 2, 3, 4 and 5 on:

- Systematic review on the use of PS and missing data methods in observational comparative effectiveness studies.
- Comparative effectiveness analysis of penicillin plus gentamicin versus penicillin alone in treatment of pneumonia.
- The sensitivity analyses conducted
- The effectiveness of Zinc co – treatment in diarrhoea admissions.

Further, it discusses more broadly the potential value and limitations of using observational data to explore treatment effectiveness – drawing on the work presented in this thesis and the literature.

Chapter 2

A systematic review of the use of propensity score and missing data methods in observational comparative effectiveness evaluations

---

This chapter first presents an overview of propensity score and missing data methods (section 2.1), then presents a descriptive systematic review of how these methods are used in observational comparative effectiveness evaluations (section 2.2)[2].

[2] Findings of this Chapter on how propensity scores are estimated in the presence of missing data were published in the Journal of Comparative Effectiveness Research: **Malla, L.,** Perera-Salazar, R., McFadden, E., Ogero, M., Stepniewska, K., English, M. (2017, ahead of print). "Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review". *Journal of Comparative Effectiveness Research*.

2.1. Description of propensity score and missing data methods

2.1.1. Propensity score methods

As introduced and defined in Chapter 1 sub – section 1.1.6, a propensity score (PS) is the probability of a patient's assignment into a treatment group given the observed covariates. PS may be estimated using either logistic or tree based regression models (37) with the outcome variable defined as treatment group. Variables included in the PS models should be those that influence treatment assignment as well as outcome (62) and can be drawn from published literature (and potentially treatment guidelines). Some authors have suggested significance testing for variable selection in PS models (63). PS are used (or implemented) in four ways: matching, sub-classification, weighting, and as a covariate in a regression model for the outcome. Each of these methods is described below:

2.1.1.1. PS matching

PS matching aims to obtain treatment and (active) control patients who have approximately equivalent propensity score values. Clinical literature mostly uses two algorithms to select the nearest neighbours – either greedy or optimal. In the greedy algorithm, treated individuals are randomly picked one at a time and each matched to an individual with the closest PS value in the control group (62). Alternatively, treated individuals may be reordered based on PS values then one at a time matched to an individual with the closest PS value in the control arm (62). The performance of reordering and random picking of treated individuals have been compared elsewhere and both were shown to give relatively similar matches (64). On the other hand, the optimal algorithm matches patients with

the least global distance between them. Distance, here, is defined as the absolute difference in the propensity scores between a treated and control patient with global distance the sum of all distances between matched treated and control patients. Optimal matching is different from greedy matching in the sense that it tries to minimise the global distance while greedy matching does not; otherwise they can both result in well balanced treatment groups (65).

If the proportion of patients in control to treatment groups is k in a dataset (where k is two and above), then each of the treated individuals may be matched with k patients in the control group. This may still be implemented with greedy and optimal algorithms –in both it is referred to as 1 to k matching. Also if there are fewer controls than treated patients then one control may be shared by more than one treated participant – and this process is referred to as matching with replacement, while matching without replacement happens when an already matched control is not available for matching with other treated patients (62). Nearest neighbour matching (1:1 or 1:k) implemented either through greedy or optimal matching most of the time reduces the effective sample size. Therefore, an alternative PS matching method is optimal full matching which retains all the patients in the analysis (while forming subsets[3] of treated and control patients using the optimal algorithm). This is the only form of matching that happens without replacement.

[3] See *Treatment Estimands* sub - section 2.1.1.5 on how the matched subsets are used.

2.1.1.2. PS Weighting

There are two types of weights[4] that may be estimated using PS. The first is inverse probability of treatment weights (IPTW) such that treated individuals are assigned weights of 1/PS while those in the (active) control group are assigned weights of 1/(1 – PS). The second is weighting by odds such that those treated are assigned a weight of 1 and those in the (active) control are assigned weights of PS/(1 – PS) (37).

2.1.1.3. PS sub – classification

Sub-classification divides patients into mutually exclusive groups based on their propensity scores. A standard practice, though not supported by specific recommendations, has been subdividing patients into five subclasses (66). One approach for creating patient subclasses would be to first conduct one on one nearest neighbour matching and then split the population into subclasses (67), alternatively one may use PS quintiles (62). The number of subclasses[5] will usually depend on the sample size, and for large datasets, more classes with reasonable sample sizes would be desirable.

2.1.1.4. Regression adjustment using PS

In this method, the estimated PS is used as a covariate (alongside other covariates) in an outcome regression model.

Before implementing the PS methods and proceeding with the analysis, one is supposed to examine the distribution of PS densities in treated and control

[4] These weights are used to estimate different treatment quantities as discussed in the *Treatment Estimands* sub - section 2.1.1.5.
[5] See *Treatment Estimands* sub - section 2.1.1.5 on how the subclasses are used to estimate treatment effects of interest.

groups to examine if they have substantial overlaps (common support). If there is no substantial overlap, then the PS analysis needs to use only the patients within the area of common support (68). This is to avoid including patients in the analysis with unreasonable distances between them. A dataset may not be usable if the PS densities of the groups being compared do not overlap sufficiently. Thus, covariate balance between groups needs to be examined – and the use of standardised mean differences (SMD), statistical testing (t tests, chi-square, McNemar's among others) and graphs (boxplots, SMD plots) have been suggested in the literature (37, 62).

2.1.1.5. Treatment estimands

Neyman and Rubin proposed a causal framework called the "Neyman Rubin Causal Model (NRCM)" that justifies the need to create a logical comparator (69). The NRCM indicates that in an experiment, only one outcome (Y_{i1}) can be observed per patient i conditioned on a given treatment group. However, the outcome (Y_{io}) that would be observed if a patient received alternative treatment (that was not given) is always missing. For instance, if treatment A and (active) control B are to be evaluated between two time intervals t1 and t2, patients would be enrolled in either but not both of the two groups initiated at time t1. If a patient receives treatment A then the observed outcome at time t2 would be due to treatment A, and here the missing outcome is that which would be observed supposing the patient was enrolled in control group B at time t1. Based on this framework, there are two possible effects that could be estimated[6]: (a)

[6] The conditioning "| Treatment = X" indicates the actual treatment received, while the sub – index Y_{iz} is the counterfactual.

average treatment effect (ATE) which would be estimated as $E(Y_{i1}|Treatment = A) - E(Y_{i0}|Treatment = B)$ and; (b) average treatment effect for the treated (ATT) estimated as $E(Y_{i1}|Treatment = A) - E(Y_{i0}|Treatment = A)$. Pirracchio (2013) applied these two treatment estimands in simulated data and found that both resulted in different effect estimates (70). Thus, in their paper, they advised the use of ATT when estimating treatment effectiveness mainly to answer the counterfactual question – what the effect would be if the treated are denied the treatment. ATE may be examined particularly in prevention interventions that may benefit anybody in a larger population. Further, these estimands have been linked to particular PS methods – such that ATT would be estimated using nearest neighbour matching methods and weighting by odds; and ATE estimated using IPTW. When using PS optimal full matching, different weights are used to estimate either ATE or ATT: (a) ATE weights are generated by assigning patients weights proportional to the total number of patients per matched subset (71) and; (b) ATT weights are generated by assigning treated patients a weight of one, while patients in control subsets are assigned weights derived by dividing the number of treated patients by those in control per subset (72). These weights generated with optimal full matching may then be used to conduct a weighted regression analysis. For PS sub-classification, ATE and ATT are obtained by pooling treatment effect estimates across the subclasses: (a) ATE is obtained by weighting of subclass treatment effects using total number of patients per subclass and; (b) ATT is obtained by weighting effect estimates

using the total number of patients considered to be in the treatment group[7] per subclass (37).

2.1.1.6. Sensitivity analyses to unmeasured confounders

Propensity score methods generate matched treated and (active) control patients whose distribution of measured covariates are as similar as possible. However, two patients with similar covariate distribution may still differ in terms of unmeasured variables – and this may introduce bias in estimated treatment effects (73). It is therefore important to examine if the unmeasured variables would potentially influence the obtained results. Two approaches (among others) have been suggested: either the use of instrumental variables or Rosenbaum bounds (74). An instrumental variable method aims to find a proxy randomised experiment in a routine or observational dataset (75), while Rosenbaum's sensitivity analysis method seeks to quantify the amount of bias, due to unmeasured confounders, that would be needed to alter the observed results (76). See (74, 77) for further information on these methods.

2.1.2. Missing data methods

Estimation of propensity scores requires that all covariate data are fully observed. However, in case of missing data then patients with missing data are excluded from analysis particularly when using non-Bayesian models that are most commonly applied in practice. Three methods may be used in the face of missing data, these include: complete case analysis, multiple imputation and

[7] In comparative analysis, there are individuals considered to be in the treatment group of interest and those in the (active) control group. Therefore, ATT estimate is obtained by weighting treatment effects by those in the former group per subclass.

estimation of PS within various patterns of missing data (78). These methods are briefly discussed.

2.1.2.1. Complete case analysis

This method excludes all patients if they have missing data in at least one of the covariates (or waves of follow up in a cohort study) from the analysis leading to loss of power. This approach only results in unbiased estimates when it is demonstrable that missing data are unrelated to the study treatments or design. For instance, if a cancer patient is enrolled in a follow up study and relocates to another city such that subsequent measurements may not be obtained – then such data may be understood to be missing completely at random (MCAR) and estimation of PS and outcome analysis excluding such patients may be valid.

2.1.2.2. Multiple imputation (MI)

MI involves filling in plausible values for missing data more than once, which accounts for uncertainty in filling in unknown values (79). MI is commonly practiced under the assumption that data are missing at random (MAR) – and this means that the missing data may potentially be related to observed co-variables but not those that are unobserved (80). For instance, in a follow up study, if a patient's condition was consistently improving over previous visits and all of a sudden s/he drops out – then it might be plausible to assume the patient is doing better (based on previous information).

As MI generates multiple datasets, two approaches have been proposed to estimate PS – within and across these (59, 81). In the within approach, propensity scores (per patient) are estimated for each of the multiple imputed

datasets. Then the appropriate PS method is used on each of these datasets to create balanced patient groups with treatment effects also estimated within each dataset. While, in the across approach, the estimated multiple propensity scores per patient are averaged across the imputed datasets – then used to estimate treatment effects in each imputed dataset. Both the within and across approaches result in multiple treatment effects which are pooled using Rubin rules (59) to account for within and between imputation variability.

2.1.2.3. PS estimation by patterns of missing data

This method aims to retain all the patients in an analysis by grouping them based on observed covariates and using models to estimate PS in each group. If there are clearly defined groupings in the study related to the design – and some data may not be obtained due to this, then clearly such data are missing not at random (MNAR). And this may be a promising method under the MNAR assumption.

Other less intuitive methods have historically been used to address missing data and these include the use of the: previous value to replace missing data (last observation carried forward - LOCF), substitution with population mean value (simple imputation) and missing indicator method – where those with missing data are assigned to a group. These methods do not account for the uncertainty involved in filling in plausible values (82), and almost always bias overall estimates particularly if they are not aligned with any missingness assumption.

The three missing data assumptions defined (MCAR, MAR and MNAR) are generally not directly testable but are instead justifiable through understanding and describing the process that generated the data.

2.1.3. Software use

A comprehensive list of software packages for implementing PS methods is documented in this link: http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html; while a list of packages for conducting multiple imputation is presented in: http://www.stefvanbuuren.nl/mi/Software.html.

2.2. Systematic review of the use of propensity score and missing data methods in observational comparative effectiveness evaluations

Section summary

Introduction: There is emerging interest in utilising routine healthcare data for causal inference derived from comparative analyses of interventions not allocated at random. Such studies must overcome the challenge of non-random allocation and, frequently, that of data missing from routine datasets to help overcome the inherent limitations in their design. To overcome non-random treatment allocation, researchers have increasingly used propensity score (PS) methods to create samples of comparable treatment and control populations. Even though systematic reviews have examined how aspects of PS methods are used in literature, none has reviewed how the challenge of missing data is addressed with these methods. This review therefore describes how both propensity score and missing data methods are used in observational comparative effectiveness studies.

Methods: Published articles on observational comparative effectiveness studies were extracted from MEDLINE and EMBASE databases using the search terms (observational stud* and propensity score*). The searches were restricted to clinical studies done in the last 7.5 years.

Results: This search yielded 167 eligible articles. Of these, 100/167 (60%) used PS matching with 74/100 (74%) using 1:1 nearest neighbour matching. Majority of the articles (n = 126/167, 75%) checked overall balance after applying PS methods. Sixty-five percent (n = 108/167) of the articles used multivariable regression for outcome analysis on PS adjusted datasets. Only in 35/167 (21%)

articles were sensitivity analyses to examine possible influence of unmeasured variables on the results attempted. Majority of studies (114/167, 68%) conducted complete case analysis with only 53 of them stating this in the methods. Very few articles reported use of appropriate methods: multiple imputation (n = 16/167, 10%) and estimating PS within missing data patterns (n = 3/167, 2%). Thirty-seven percent of articles (n = 62/167) provided information on the amount of missing data and only seven percent (n=12/167) provided reasons for missingness. Only 8/167 articles (5%) fully adhered to STROBE guidelines when reporting missing data.

Conclusions: these findings on the use of propensity score methods are consistent with those found in the most recent systematic reviews. The most common PS method used is matching. Sensitivity analyses to examine possible influence of unobserved confounders were not common. Very few researchers use optimal methods for handling missing data or reported their missing data methodology which may lead to reporting biased findings. To improve the reliability of their findings, researchers should enhance the reporting and quality of their work by adhering to suggested guidelines for PS use and handling missing data. Recently published PS guidance is available on the EQUATOR website as part of STROBE guidelines.

2.2.1. Introduction

Existing systematic reviews (45, 46, 83) have focussed on the use and adequacy of reporting of PS methods with an aim of examining reproducibility of findings. The latest by Zakrison (2017) examined articles published up to July 2015 and Yao (2017) examined articles published up to November 2015. The methodological aspects examined in these reviews included: variable selection, methods for estimating and using PS, checking of covariate balance, methods for analysing outcome(s) after PS estimation among others. However, none of the reviews examined how PS methods are used in the presence of missing data. Analysis may result in biased estimates if missing data are not correctly handled (84) even with proper use and reporting of PS methods. On the other hand, existing missing data reviews have focused on RCTs or observational studies that do not focus on comparative evaluations (for example see (85-87)) – thus they have not examined the use of PS methods. Building on previous systematic reviews (45, 46, 83, 87, 88), I therefore systematically reviewed published observational comparative effectiveness studies primarily to:

iii) Describe the methodological aspects of PS use in the general clinical literature (as the latest reviews had focused only on acute care and cancer literature).

iv) Assess the methods for handling missing data in estimation of PS and if the methods used are in line with the STROBE guidelines (60). The STROBE guidelines make recommendations for how missing data should be reported in observational studies and in particular they

require researchers to report on the proportion of missing data, reasons why data are missing and how missing data are addressed.

2.2.2. Systematic review methods

2.2.2.1. Search terms and literature search databases

I used the search terms *(observational stud* and propensity score*)* – and with this literature searches were carried out in Embase[OvidSP] (1974 to 2017 June 30) and Medline[OvidSP] (1946 to 2017 June 30). The searches were further restricted to clinical articles published in the last 7.5 years (since January 2010). The restriction was motivated by the fact that the STROBE guidelines were first published in 2007 (60) and researchers would actively use them three years after dissemination and the lag to doing work to publication.

2.2.2.2. Inclusion and exclusion criteria

Primary research studies of observational comparative effectiveness studies, published in English, were included. Studies only discussing methods in randomised studies, meta-analyses and systematic reviews were therefore excluded, although bibliographies from secondary studies were hand searched for additional relevant primary observational studies. Conference abstracts were also excluded due to the limited information they provided. Further, studies on quasi experiments were also excluded since a researcher may sometimes have control over treatment assignment. All the studies excluded as described here are referred to as non – primary.

2.2.2.3. Screening of papers and data extraction[8]

Abstracts were selected for text review and final inclusion by myself. If any abstract did not offer sufficient information, then the full text was obtained (together with supplementary materials) and scanned through for the relevant aspects of method(s). I extracted data into structured forms (from the full text of the selected articles) which contained variables specific to: (i) article characteristics; (ii) propensity score methods and; (iii) missing data methods (see **Table 2.2.1** for full description of the variables). As some articles used more than one PS method, I considered the methods on which results were based as primary (the use of other PS methods is also reported). Methods used for sub-group analysis within these articles were not examined. A second reviewer (MO) randomly selected 25% (n = 42) of the articles that met inclusion and exclusion criteria and did an independent data abstraction then reviewed methods and the degree of agreement was examined using Kappa coefficient (89) and is reported. This comparison in agreement was examined for each of: PS variable selection method, PS method used, treatment estimand reported, method for estimating PS models, checking of balance after PS use, sensitivity approaches to unmeasured confounders, proportion of missing data reported, reasons for missing data indicated and method used to address missing data. Where there was any disagreement, LM (myself) and MO jointly reviewed the article and reached a consensus.

[8] In this sub – section, I refer to a colleague in the Health Services Unit as MO

Table 2.2.1: List of variables extracted

#	Variable	Response options
	Article characteristics	
1	Year of publication	
2	Title of publication	
3	Number of treatment groups compared	(1) 2 (2) > 2
4	Setting (country/continent)	
	Propensity score methods	
5	Variable selection method	(1) Clinical relevance (2) Statistical significance (3) Both clinical relevance and statistical significance (4) not mentioned
6	Propensity score method used	(1) Matching (2) Weighting (3) Sub-classification (4) Regression adjustment (5) not mentioned
7	Treatment estimand reported	(1) ATE (2) ATT (3) Both ATE and ATT (4) not mentioned
8	Matching algorithm used (if matching is used)	(1) Greedy (2) optimal (3) nearest-neighbour (4) not mentioned (5) matching not used
9	Type of matching used	(1) 1:1 (2) ratio (3) not mentioned
10	Weighting method (if weighting is used)	(1) IPTW (2) Weighting using PS odds (4) not mentioned
11	Number of subclasses used (if sub-classification is used)	
12	Proportion of effective sample size after using propensity score methods	
13	Method for estimating PS models	(1) Logistic regression (logit/probit) (2) Regression trees (3) not mentioned
14	Multiple propensity score methods compared to select the best performing one	(1) Yes (2) No
15	Checking of balance after the use of propensity methods	(1) Yes (2) No
16	Graphical methods used	(1) Yes (2) No
17	Standardised mean difference used	(1) Yes (2) No
18	Statistical methods used to check balance	(1) Yes (2) No
19	Common support examined	(1) Yes (2) No
20	Outcome analysis methods	(1) Multivariable regression (2) simple methods (T-test, relative difference, Wilcoxon, etc)
21	Sensitivity approach to unmeasured variables	(1) Instrumental variables (2) Rosenbaum bounds (3) No sensitivity method used
22	Software used for estimating PS	(1) R (MatchIt, Matching, twang, optmatch) (2) SAS (3) STATA (4) Other (5) Not mentioned
23	***Missing data methods with propensity score models***	
24	Proportion of missing data reported	(1) Yes (2) No
25	Missing data method reported	(1) Yes (2) No

26	Missing data mechanism mentioned	(1) Yes (2) No
27	Reason for missing data given	(1) Yes (2) No
28	Specific missing data mechanism	(1) MCAR (2) MAR (3) MNAR (4) Not mentioned
29	Specific missing data method used	(1) Complete case (2) Pattern mixture (3) Multiple imputation (4) Not mentioned
30	Missing data sensitivity conducted	(1) Yes (No)
31	Analysis compared between those with complete and incomplete data	(1) Yes (2) No
32	Variables included in MI explained (if MI used)	(1) Yes (2) No
33	Number of imputations specified (if MI used)	(1) Yes (2) No
34	Methods used to estimate propensity scores after MI	(1) Within (2) Across (3) Not mentioned
35	Software used for MI	(1) R (Hmisc, MICE, mi, etc) (2) SAS (MI) (3) STATA (MI) (4) Other (5) Not mentioned

2.2.2.4. Presentation of findings

Articles were stratified into four publication time periods; 2010 – 2011, 2012 – 2013, 2014 – 2015 and 2016 – 2017 to explore trends in practice. Frequencies and proportions were used to summarise findings. This review has incorporated various methodological aspects recommended in the PRISMA statement (90).

2.2.3. Results

2.2.3.1. Study characteristics

The process of identification and selection of articles is summarised in **Figure 2.2.1**. I identified 2422 articles after 973 duplicates were removed, 2255 articles did not meet inclusion criteria leaving 167 articles for full text review (see **Appendix B** – references for these articles).

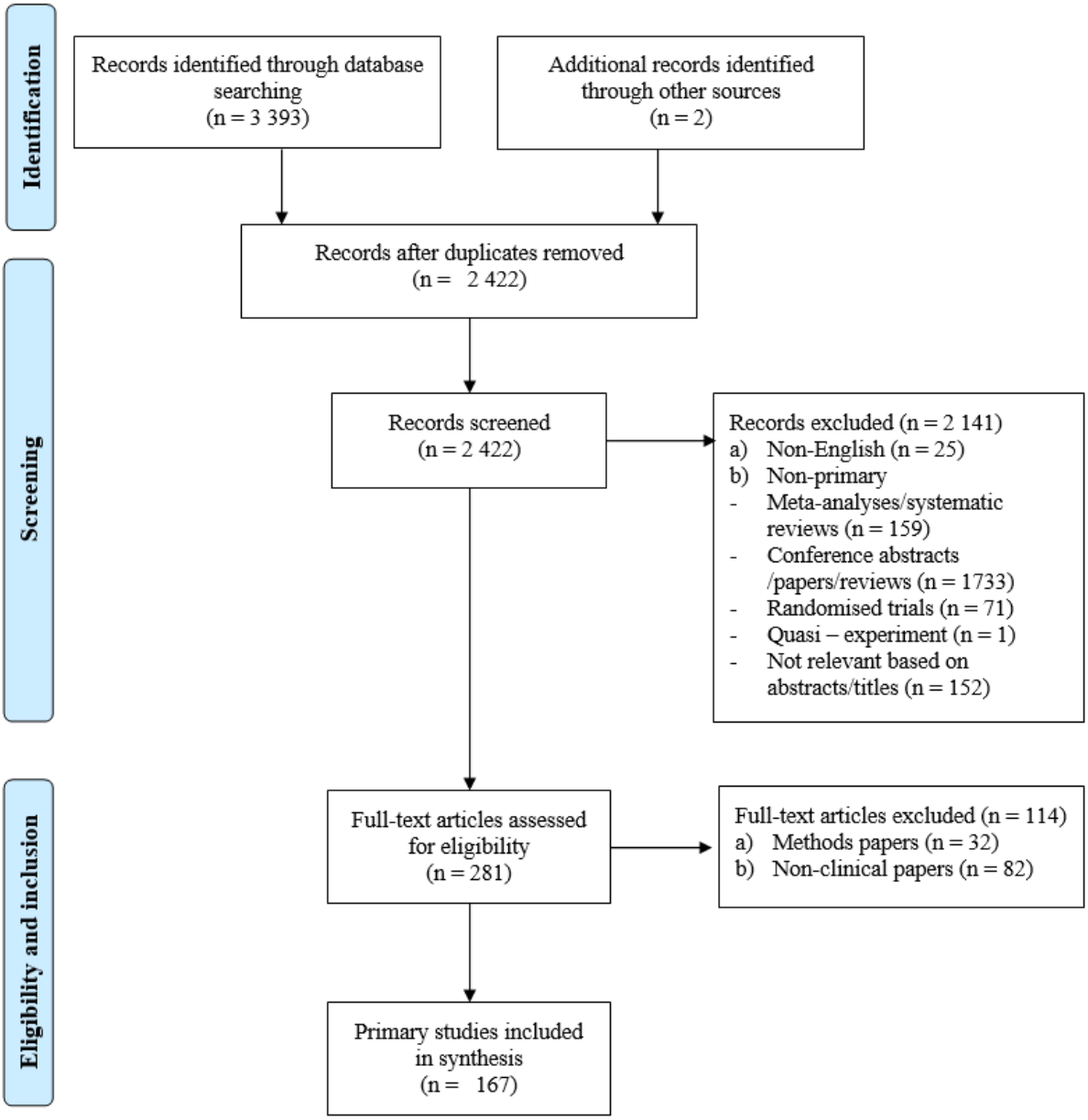


Figure 2.2.1: Selection process of primary observational studies

Seventy-one percent (n = 118/167) of the eligible articles retrospectively analysed routine datasets most of which were derived from medical registry databases. The remaining twenty-nine percent (n = 49/167) used prospective designs which were based in hospital settings. Most of the articles (90%, n = 151/167) conducted comparative analyses using two treatment groups while 10% (n = 16) reported

analyses using more than two groups. Almost half of the articles (n = 81/167) were based on data from North America (USA and Canada) – while 29% (n = 49/167), 18% (n = 30/167), 6 /167(n = 4%) and 1/167 (n = 1%) were based on data from Europe, Asia, Australia and Africa respectively (**Table 2.2.2**).

Table 2.2.2: Description of the studies

	n (%)
Design	
Retrospective	118 (71%)
Prospective	49 (29%)
Setting	
North America	81 (49%)
Europe	49 (29%)
Asia	30 (18%)
Australia	6 (4%)
Africa	1 (1%)
Number of treatments compared	
2	151 (90%)
>2	16 (10%)

The Kappa estimates of agreement on the selected PS and missing data methodological aspects independently identified by myself and MO ranged from 0.96 to 1 which indicated excellent agreements.

2.2.3.2. Propensity score methods used

2.2.3.2.1. PS estimation methods and variable selection

In estimating propensity scores, 78% (n = 130/167) of the articles used logistic regression models and 2% (n = 3/167) used tree based regression models. However, it was not clear in 20% (n = 32/167) of the articles which methods were used to estimate the propensity scores. Slightly more than half of the articles (n = 88/167, 53%) listed and justified the variables used in PS regression models; 42% (n = 70/167) cited clinical relevance and or significance in previous literature

and; 11% (n = 18/167) conducted significance testing prior to inclusion in the final PS models.

2.2.3.2.2. PS implementation methods

Table 2.2.3 summarises the PS methods articles used to adjust for or balance covariates between treatment groups. The most common method used was matching (n = 100/167, 60%), followed by regression adjustment (n = 30/167, 18%), then weighting (n = 23/167, 14%) and lastly sub-classification (n = 14/167, 8%). Across the time periods, the proportion of articles that used matching and weighting seemed to increase while regression adjustment and sub-classification seemed to decrease.

In 100 articles that created balanced treatment and control groups with PS matching, 28% (n = 28/100) used the greedy algorithm in selection of populations from treatment and (active) control groups, and four articles used the optimal algorithm. Many reports (n = 68/100, 68%) did not indicate the specific nearest neighbour algorithms implemented. For 23 articles that used weighting, sixteen and four used IPTW and weighting by odds respectively, the other 3/23 (13%) did not indicate the type of PS weights used. For 14 articles that used PS sub-classification, numbers of sub-classes used were: 3 – 5 (n = 9/14), 6 – 10 (n = 3/14), 30 (n = 1/14) and 60 (n = 1/14).

I also examined the proportion of the dataset that were eliminated from the analyses as a result of the method used for PS adjustment. Data were available in 114 articles using: PS matching (n = 81/114), weighting (n = 20/114) or sub-classification (n = 13/114)). For articles that used matching, 35/81 (43%) had eliminated more than 70% of the original sample with two studies restricting the

analysis to less than 10% of the original sample while the remaining studies (n = 44/81) use sample sizes between 31 and 90% of the total. The sample sizes were much reduced with matching as mostly articles used one-to-one matching (n = 74/100, 74%) where many patients in the (active) controls were dropped from the analyses. Sub-classification and weighting mostly retained 100% of the patients in analyses.

Table 2.2.3: Summary of PS implementation methods

	2010 – 2011	2012 – 2013	2014 – 2015	2016 - 2017	Total
Matching	8 (38%)	16 (44%)	37 (67%)	39 (71%)	100 (60%)
Regression adjustment	9 (43%)	9 (25%)	10 (18%)	2 (4%)	30 (18%)
Weighting	1 (5%)	5 (14%)	5 (9%)	12 (22%)	23 (14%)
Sub-classification	3 (14%)	6 (17%)	3 (5%)	2 (4%)	14 (8%)
Total	**21 (13%)**	**36 (22%)**	**55 (33%)**	**55 (33%)**	**167 (100%)**

2.2.3.2.3. Checking of common support and covariate balance after application of PS methods

Twenty articles (12%) examined the degree of common support through the use of density plots and histograms and only one article excluded from analysis observations from the region outside that of common support. Of the 167 articles, 126 (75%) checked balance in distribution of covariates between treatment and control groups (by PS method: matching (n = 90/126), weighting (n = 17/126), sub-classification (n = 11/126), regression adjustment (n = 8/126)). **Table 2.2.4** summarises the methods used for checking covariate balance. As some articles used more than one method, the total number of methods exceeded the actual number that used them. The most common method was the use of significance testing (n = 79/126, 63%), followed by standardised mean difference (n = 46/126,

37%) and graphical methods (n = 17/126, 13%). Nineteen articles (11%) used none of these three methods but presented variable means per treatment group.

Table 2.2.4: Summary of methods used for checking distribution of covariate balance

	Graphical methods	Standardised mean difference	Significance testing
Matching	11 (9%)	39 (31%)	55 (44%)
Regression adjustment	3 (2%)	2 (2%)	4 (3%)
Weighting	1 (1%)	4 (3%)	12 (10%)
Sub-classification	2 (2%)	1 (1%)	8 (6%)
Total	17 (13%)	46 (37%)	79 (63%)

* The denominator for calculation of all the percentages is 126

2.2.3.2.4. Outcome analysis methods with PS methods

After estimating PS on the datasets, what followed was the outcome analysis. Sixty-five percent (n = 108/167) of the articles used multivariable regression on the PS adjusted datasets. This number included those articles in which PS was used (together with other variables) as a covariate in the regression models. The remaining 35% (n = 59/167) used simple methods in outcome analysis. In general, the common simple methods used after PS adjustment included: t, chi-square, log-rank, Kruskal-Wallis, Mann Whitney, Wilcoxon and McNemar's tests.

2.2.3.2.5. Use of more than one PS method

Of the 167 articles, fourteen percent (n = 23/167) used more than one PS method. Only one of the 23 compared the quality of matched populations (for the different PS methods) before conducting the outcome analysis. The remaining 22/23 conducted outcome analyses without examining the PS method that would result in the most balanced population.

2.2.3.2.6. Treatment estimands reported

Two articles clearly reported estimating ATE and only one ATT. The two articles that estimated ATE had conducted PS matching and did not provide much methodological detail to assess the appropriateness of this estimand with matching (as different weighting schemes may be used to estimate ATT if they used optimal full matching). The only article that reported ATT had used 1:1 matching which was an appropriate method. The rest of the articles (n = 164/167, 98%) did not specifically mention any type of treatment estimand in their analyses. However, I implicitly inferred the type of estimands based on the PS methods implemented: 14/164 of these articles used IPTW which means they estimated ATE while some used weighting by odds (n = 4/164) and 1:1 or ratio matching (n = 85/164) which means they estimated ATT. It was not feasible to determine the type of estimands used in the remaining 61 articles as they did not provide sufficient specific details on the PS methods they used. Based on this, it seems the majority of articles estimated ATT.

2.2.3.2.7. Sensitivity methods to unmeasured covariates

To assess the possible influence of unmeasured confounders, seven articles used instrumental variable analysis, three used Rosenbaum and related approaches, and one fitted the same model to a dataset which had a key variable they did not measure. A further 24 articles attempted sensitivity analysis by using sub-groups or a subset of variables within their datasets. However, a majority of the articles (n = 132/167, 79%) did not report any sensitivity analyses – although 113/132 (86%) of these acknowledged the possibility of their results being biased due to known or unknown unmeasured confounders.

2.2.3.3. Missing data methods used

2.2.3.3.1. Reporting of missing data

Of the 167 articles, thirty-seven percent (n = 62/167) provided information on the amount of missing data and only 12/62 provided reasons for missingness. Of the 12, three articles linked the reasons to the type of missing data mechanism: MAR (n = 1/3) and MCAR (n = 2/3). None of the remaining articles that provided information on amount of missing data (n = 59/62) commented on any assumed missingness mechanism.

2.2.3.3.2. Missing data methods used

Missing data were addressed in 51% (n = 86/167) of the articles. The remaining 81 did not mention how they dealt with missing data (**Table 2.2.5**). The most common approach used was complete case analysis (62%, n = 53/86) – which was determined if articles mentioned they excluded individuals due to missing data. Among the 81 articles that did not mention any use of missing data methods, 61 used multivariable regression models to estimate propensity scores using standard software (R, SAS, SPSS and STATA). Therefore, I assumed that these articles must have used complete case analysis as this is the default method for handling missing data in these analytic software. Going by this then the number of articles that used complete case analysis would be 114/167. Some articles also used *ad hoc* methods that are likely to result in biased estimates: (a) imputation to most common category (n = 4); (b) mean value imputation (n = 3); (c) LOCF (n = 2); (d) missing indicator method (n = 1) and; (e) truncation which was used in a cohort study (n =1) where data were utilised only up to the earliest time when

the first dropout was experienced, and all the follow up data beyond this were excluded for everybody.

Nineteen articles (11%) used more appropriate methods: (a) estimation of PS by various patterns of missing data (n = 3) and; (b) multiple imputation which was used in 16 of the articles among those that reported missing data methods. Among the articles that used multiple imputation; 8/16 (50%) reported the number of imputed datasets – with the least being five and maximum 2000 and,12/16 (75%) explained the variables that were included in the imputation models. No article reported diagnostics on the plausibility of multiply imputed values. After conducting multiple imputation, only 5/16 (31 %) articles reported on how PS and outcome analysis proceeded: (a) four articles indicated to have estimated PS and outcome analysis per imputed dataset then pooled estimates of treatment effectiveness using Rubin rules and; (b) one article averaged PS across the imputed datasets – then used this in the adjusted analysis for all the imputed datasets with effects also pooled using Rubin rules.

Table 2.2.5: Summary of missing data methods

	Number of papers				
	2010 – 2011	**2012 – 2013**	**2014 – 2015**	**2016 - 2017**	**Total**
Methods not mentioned	11 (52%)	20 (56%)	25 (45%)	25 (45%)	81 (49%)
Methods mentioned	10 (48%)	16 (44%)	30 (55%)	30 (55%)	86 (51%)
Complete case	10 (100%)	11 (31%)	16 (29%)	16 (29%)	53 (62%)
Multiple imputation	0 (0%)	2 (6%)	5 (9%)	9 (16%)	16 (19%)
Pattern mixture	0 (0%)	1 (3%)	1 (2%)	1 (2%)	3 (3%)
Imputation to most common category	0 (0%)	0 (0%)	4 (7%)	0 (0%)	4 (5%)
Simple imputation	0 (0%)	0 (0%)	2 (4%)	1 (2%)	3 (3%)
Imputation (type not specified)	0 (0%)	0 (0%)	2 (4%)	1 (2%)	3 (3%)
LOCF	0 (0%)	1 (3%)	0 (0%)	1 (2%)	2 (2%)
Truncation	0 (0%)	0 (0%)	0 (0%)	1 (2%)	1 (1%)
Missing indicator	0 (0%)	1 (3%)	0 (0%)	0 (0%)	1 (1%)

* Some of the percentages did not add up to 100% due to rounding off.

*The percentages for missing data methods are based on the number of articles that mentioned use of methods per reporting time period.

Across the four reporting time periods, the proportion of articles using complete case analysis was higher in the articles published in 2010 – 2013 compared to the use in articles published in 2014 – 2017. The proportion of articles using multiple imputation increased across the four reporting time periods, from none in 2010 – 2011 to 16% in 2016 – 2017.

2.2.3.3.3. Missing data sensitivity analysis

The only form of sensitivity analysis that was conducted in five articles was the comparison of treatment effectiveness estimates derived from analyses of multiply imputed datasets and those from complete cases – or comparison of estimates between completers and non – completers in cohort studies. However, no thorough sensitivity analyses aiming to rule out the possibility of MNAR were reported.

2.2.3.3.4. Adherence to STROBE guidelines of reporting missing data

As presented above, adherence to different aspects of STROBE guidelines as pertains to reporting of missing data were as follows: indicating reasons for missing data (n = 12/167, 7%), reporting amount of missing data (n = 62/167, 37%), and indicating missing data method(s) used (n = 86/167, 51%). Overall, only 8/167 (5%) articles adhered to all these three aspects.

2.2.4. Discussion

The use of PS methods has earlier been examined in systematic reviews. Therefore, findings obtained in this review updates the existing literature on the use of PS methods and also provides the first insights on how missing data are addressed together with these methods in the clinical literature. These findings are consistent with the recent systematic reviews by Thoemmes (2011) (45) and partly those of Zakrison (2017) (46) in which the number of published articles using PS methods have increased over time and majority of researchers continue to:

- Use PS matching as the preferred method for creating balanced treatment groups
- Not provide sufficient justification for variable selection for PS models;
- Commonly use logistic regression as the PS estimation method;
- Check balance of covariate distribution;
- Use significance testing as the commonest method for assessing balance;
- Ignore checking of common support in their analyses.

As most researchers used 1:1 or 1:k matching, the resulting effective sample sizes were often reduced. Efforts should be made to utilise all the data by implementing several matching strategies including optimal full matching – and base outcome analysis on the method that minimises covariate imbalance to a desirable degree and at the same time should retain the largest number of patients in the analysis (91). Additional findings in this review show that there has been no increase in the number of articles using doubly robust methods in which multivariable regression models are used for outcome analysis after adjusting datasets with PS. This is in comparison to a previous review by Austin (2008) (92) which indicated that a third of studies in medical literature did not use proper methods in estimating treatment effects after creating PS adjusted datasets.

As sensitivity analyses to unmeasured confounders were rare in articles I reviewed, researchers should also be encouraged to use appropriate methods to conduct sensitivity analyses (after PS analysis) including – instrumental variables and Rosenbaum bounds to help increase confidence in their findings.

Among 167 studies with PS, only 86 (51%) discussed missing data issue and among them only 16 (19%) used multiple imputation methods to account correctly for the missing data. However, even in these studies reporting was incomplete as only 5 (31%) described how results were generated across the imputed datasets. Approximately 68% of the articles based effect estimates on complete case analysis but in the majority of cases this was an implicit not explicit analysis strategy. This result is consistent with the findings of Karahalios (2012) (87) and Eekhout (2012) (88) who reviewed the use of missing

data methods in epidemiologic studies (though not specific to the use of PS) and also found that most researchers used complete case analysis in practice. Complete case analysis may provide unbiased estimates only if data are missing completely at random (MCAR). However, only one of the studies provided evidence for MCAR and thus appropriately estimated PS and treatment effectiveness on complete cases. In other cases, researchers used complete case analysis but offered no explanation for why data were missing nor the underlying data generation mechanism. This use of complete case analysis may result in elimination of substantial numbers of observations and the implicit assumption of MCAR can potentially give biased results. In situations where it is not known how data were generated (as could be the case when using registry datasets), then it is plausible to assume that data are MAR – then use multiple imputation in estimating propensity scores. Thereafter sensitivity analysis should be used to help validate this assumption because the existence of the MNAR assumption may not be ruled out where MAR is thought to hold (93, 94).

Imputation is aimed at maximising the use of all available data. However, less principled methods like simple imputation, LOCF, imputation to most common category, truncation and missing indicator (as were used in some of the articles in this review) should be avoided (87).

This review has a number of strengths and limitations. It is the first study to systematically, jointly examine the use of PS and missing data methods in recent comparative effectiveness studies using observational datasets. As comparative effectiveness is often an implied concept rather than being explicitly stated, I did not restrict the search term to further include 'comparative effectiveness', and I

was able to obtain a potentially more representative number of articles than if this was included as part of the search. However, I do acknowledge that some articles may have been excluded as searches were not conducted for the grey literature. Study selection was only carried out by myself, however inclusion criteria were clearly defined and most studies were excluded as they were not observational studies comparing treatment outcomes. Moreover, additional review of a random selection of included articles resulted in a very high level of agreement.

2.2.5. Conclusion

Even though there are no formal guidelines on how PS methods should be reported, researchers are encouraged to observe the following reporting elements:

- First define the type of estimand they are interested in, then select appropriate PS method linked to the estimands;
- Examine the common support;
- Give detailed information on the process of achieving balanced groups – for instance they need to report on the model type (logistic or tree based regression), variables used in PS models, algorithms (in case of matching then either optimal or greedy);
- Method for examining balance of covariates;
- Proportion of sample size before and after PS adjustment among other elements.

Detailed information on these recommendations are documented in Stuart (2010), Austin (2011), Thoemmes (2011), Zakrison (2017) and Yao (2017,

published on EQUATOR website) (37, 45, 46, 62, 83). These recommendations are to help researchers in assessing the adequacy and reporting of the PS methods they use. In addition to the PS reporting needs, researchers should pay attention to how missing data are methodologically addressed to avoid unknowingly reporting biased results. These would include (but not limited to):

- Identifying the most appropriate missing data mechanism (either MCAR or MAR or MNAR). If unknown, then researchers may assume MAR.
- Applying the correct missing data method based on the identified/assumed missing data mechanism. Methods and possible mechanisms have been explained in sub – section 2.1.2.
- Then conduct sensitivity analysis in the MNAR framework if data generation mechanism was unknown.

As also suggested by Karahalios (2012) (87), authors and editors should follow STROBE guidelines to increase the reliability of findings. That is, reasons for missing data should be indicated, amount of and methods used to handle missing data should be reported.

Chapter 3

Using observational data to compare the effectiveness of antibiotic treatments for children hospitalised with pneumonia in Kenya

---

This chapter is subdivided into two sections which I briefly describe (and later discuss) as:

3.1. Analysis protocol – provides the background to the design for a comparative analysis of penicillin versus penicillin plus gentamicin (based on guiding principles explained in Chapter 1 sub – section 1.1.6). This section also explains the analysis steps to be undertaken after estimating propensity scores on multiply imputed datasets.

3.2. Results – describes and discusses the results of analyses based on the protocol in section 3.1.[9]

---

[9] The protocol and the analysis results were published in BMJ Open as separate papers:

i) The protocol: Malla L, Perera-Salazar R, McFadden E, et al. Comparative effectiveness of injectable penicillin versus a combination of penicillin and gentamicin in children with pneumonia characterised by indrawing in Kenya: protocol for an observational study. *BMJ Open* 2017;7:e016784. doi: 10.1136/bmjopen-2017-016784.

ii) The results: Malla L, Perera-Salazar R, McFadden E, et al. Comparative effectiveness of injectable penicillin versus a combination of penicillin and gentamicin in children with pneumonia characterised by indrawing in Kenya: a retrospective observational study. *BMJ Open* 2017;7:e019478. doi: 10.1136/bmjopen-2017-019478.

3.1. Comparative effectiveness of injectable penicillin versus a combination of penicillin and gentamicin in children with pneumonia characterised by indrawing in Kenya: A protocol for an observational study

Section summary

Introduction: WHO treatment guidelines are widely recommended for guiding treatment for millions of children with pneumonia every year across multiple low and middle income countries. Guidelines are based on synthesis of available evidence that provides moderate certainty in evidence of effects for forms of pneumonia that can result in hospitalisation. However, trials have included fewer children from Africa than other settings and it is suggested that African children with pneumonia have higher mortality. Thus despite improving access to recommended treatments and deployment with high coverage of childhood vaccines, pneumonia remains one of the top causes of mortality for children in Kenya. Establishing whether there are benefits of alternative treatment regimens to help reduce mortality would ideally utilize large pragmatic clinical trials. However, these remain relatively expensive and time consuming. This protocol describes an approach to using secondary analysis of a new, large observational dataset (CIN) as a potentially cheaper and quicker way to examine the comparative effectiveness of penicillin versus penicillin plus gentamicin in treatment of (indrawing) pneumonia. Addressing this question is important as although it is now recommended that this form of pneumonia is treated with oral medication as an outpatient it remains associated with non-trivial mortality that may be higher outside trial populations. Further, the value of using broad versus

narrow spectrum antibiotics has rarely been examined in trials (see Chapter 1 sub – section 1.1.4.1 on summary of evidence).

Methods and analysis: I used the CIN dataset described in sub – section 1.2.2 of Chapter 1. This analysis used inpatient CIN data collected up to March 2016 when the pneumonia treatment guidelines changed in Kenya. These data were generated as part of a large observational research exercise and represent the documentation of clinical findings of clinicians in routine practice and their treatment prescriptions. Although CIN is an established data collection system incorporating a number of data quality checks it still poses challenges of non-random treatment allocation and missing data. To overcome these challenges this analysis used a rigorous approach to study design, propensity score methods and multiple imputation to minimize bias.

Ethics and dissemination: The primary data are held by hospitals participating in the Kenyan Clinical Information Network (CIN) project with de-identifed data shared with the KEMRI-Wellcome Trust Research Programme for agreed analyses. The use of data for the analysis described that are based on de-identified data collected after patient discharge and that focus on the effectiveness of routine treatments was approved by the Kenya Medical Research Institute Scientific and Ethical Review Unit.

3.1.1. Introduction

Kenya has developed and disseminated national treatment guidelines largely drawing on those of WHO for a number of childhood diseases including pneumonia (22, 33). These pneumonia guideline recommendations are based on synthesis of available evidence that provides moderate certainty in evidence of effects of treatments for forms of pneumonia that can result in hospitalization (22, 33). Such guidelines have been shown to be effective in reducing pneumonia related mortality and thus Kenyan clinicians are supposed to use them in routine practice to treat pneumonia (and other diseases) (43, 95). However, although the guidelines are based on the best available evidence, the evidence available from trials conducted in Africa remains limited (19). There has also been little thorough investigation of the effectiveness of treatments in non-trial populations in routine settings that may often differ from those enrolled in formal clinical trials. For example, many children admitted with pneumonia may have co-morbidity that might exclude them from trials (96). These issues can prove problematic when making national guidelines where study generalisability can be contested (97).

The WHO and Kenyan pneumonia treatment guidelines are implicitly based on risk stratification of illness with children deemed at higher risk of severe illness and mortality offered broad spectrum antibiotic regimens and those at lower risk narrower spectrum antibiotics (27, 28, 33, 98). This risk stratification approach is operationalized by requiring clinicians to look for specific features in the clinical history and examination that are used to define illness severity and therefore recommended treatment (see **Box 1.1.1** of Chapter 1 for pre – 2016

pneumonia severity classification and treatment algorithms). Previous studies conducted in Kenya have, however, indicated that clinicians do not always follow guideline recommendations in treating pneumonia (43). Variation from the guideline recommended approach can occur at the point of pneumonia severity assignment (clinicians do not follow the rules linking clinical signs and severity category) and at the point of treatment assignment (clinicians do not follow the rules linking treatment and severity). This variability in treatment assignment provides the opportunity for comparative effectiveness evaluation if similar populations of children with pneumonia are prescribed different treatments. Clinicians may create such a situation by not following recommendations because they have inadequate knowledge or if they believe (potentially contrary to the evidence) that certain treatments result in better health outcomes.

In particular, a previous study showed that clinicians over-prescribed gentamicin, adding this to penicillin for the treatment of pneumonia characterized by lower chest wall indrawing but no other signs of severe illness instead of penicillin alone as was recommended (43)[10]. Therefore, this protocol was developed to explore whether there is any benefit from adding gentamicin to penicillin in treating children with indrawing pneumonia. Such a benefit could accrue if bacterial causes of pneumonia that were previously (prior to introduction of new vaccines) proportionately less common (eg. *S. aureus* and gram negative bacteria) are now accounting for an increased proportion of pneumonia deaths – as in such cases, the addition of gentamicin might provide

[10] The fact that inadequate knowledge in handling childhood pneumonia may result in inconsistent treatment allocation is supported by a survey conducted in seven developing countries showing that 56% of nurses and doctors had inadequate knowledge in managing pneumonia in children (99).

effective treatment for a broader spectrum of pathogens. Tackling this question is of importance as WHO have recently changed indrawing pneumonia treatment guidance based on trials that suggest equivalence of oral amoxicillin and injectable penicillin (100-103). New guidance recommends outpatient oral treatment for a population of children previously admitted to hospital (27). However, mortality from pneumonia has been reported to be higher in African settings (7, 104) despite the increasing use of multiple vaccines spanning: measles, pertussis, HiB and pneumococcal conjugate vaccines. It remains possible therefore that for a small number of children a broader spectrum antibiotic regimen might be of benefit. This study addressed this question that has not been the subject of prior community and pragmatic clinical trials.

3.1.1.1. Objectives

Primary

1) Experiment 1: To compare the effectiveness of injectable penicillin versus penicillin plus gentamicin (both injectable) in treatment of indrawing pneumonia; where severity level is constructed (assigned) using data recorded on each child's clinical signs (hospitals use a structured record form that supports recording of signs highlighted in guidelines) such that the assigned severity classification is consistent with guideline recommendations.

Secondary

2) Experiment 2: To compare effectiveness of injectable penicillin versus penicillin plus gentamicin in treatment of indrawing pneumonia; where I used clinician documented severity level.

3) Experiment 3: To compare effectiveness of injectable penicillin versus penicillin plus gentamicin in treatment of all cases of pneumonia admitted to hospital.

Experiment 1 was primary as it most approximated a typical randomised trial where recruitment would be based on specified clinical signs. This scenario provided an evaluation of alternative therapies within a guideline class (where children had very similar clinical signs) and thus was the best mimic of a prospectively designed comparative evaluation in which clinicians stick to the rules of severity classification (see (31) for an example of a RCT in Kenya that this would be similar to – where classification is based on clinical signs). Recommended treatment for this disease classification was penicillin alone, treatment with combination therapy may therefore represent over-treatment. Alternatively, the combination treatment that provides broader antimicrobial cover could provide an advantage in a small proportion of cases that would only be detected in moderately large studies – where the addition of gentamicin offers improved treatment for specific organisms not susceptible to penicillin alone.

Experiment 2 provided a test of alternative therapies amongst those where clinicians used their own judgement (possibly including gut feeling) to classify and treat (55) and have on occasions (potentially) over-ridden or ignored the

guideline recommendations. In this case although the same label of indrawing pneumonia is given to all, the treatment selected may be an indicator of perceived severity and there may be a potential bias as a result – and the propensity score distributions (see below) may help demonstrate this and in theory may overcome this potential bias. Here if there is no clinically relevant difference between treatments within a group of patients that reflects clinicians' actual classification decisions this could reassure them that monotherapy with penicillin (or amoxicillin) would be acceptable. Lastly, experiment 3 was an extension of the logic of experiment two. To date there have been no pragmatic trials of penicillin alone compared with alternative combination therapies for all forms of inpatient pneumonia, and addressing this question was relevant possibly for two reasons. First, the population of children admitted with severe forms of pneumonia is now largely one that has received *H. influenzae* Type B and pneumococcal conjugate vaccines that have likely changed the aetiology of this illness. Second, if clinicians are poorly trained and unable to classify illness severity – resulting in non-adherence to guidelines - it would be useful to explore the potential impact of this across all levels of severity of pneumonia. This analysis had the largest numbers of patients.

3.1.2. Methods and analysis

To answer these three questions, I used the Kenyan Clinical Information Network (CIN) dataset that provided observational data on all admissions to 13 Kenyan County (plus one sub - county) hospitals (see Chapter 1 sub – section 1.2.2 on CIN introduction). The analysis proceeded in two stages – design and

outcome analysis as suggested by Rubin (2008) (39) as an objective way for analysing observational datasets.

3.1.2.1. Study Design

This was an observational study conducting secondary analyses of data routinely collected from hospital paediatric wards in Kenya's CIN. The design process for the three experimental scenarios were similar and broadly consisted of the following steps suggested in Rubin (2008) (39):

a) Definition of inclusion and exclusion criteria.
b) Understanding the pneumonia diagnosis and treatment assignment processes. This was to help understand key and auxiliary variables required for analysis.
c) Verification of sample size if sufficient for any meaningful analyses.
d) Creation of comparable treatment arms – which aimed to analytically overcome non – random treatment assignment and deal with missing data.
e) Outcome analysis followed after conceptualisation of design in steps a – d.

3.1.2.1.1. Inclusion and exclusion

This analysis included all children aged 2 – 59 months and excluded children with any co-morbidity of HIV, meningitis, tuberculosis or severe acute malnutrition as there are specific antibiotic treatment rules for these children that supersede those for pneumonia. Specifically, Kenyan guidelines for the inpatient treatment of pneumonia in children that are HIV infected recommend only combination therapy. Importantly therefore children with other co-

morbidities such as mild anaemia, diarrhoea and malaria would not necessarily be excluded from the analysis.

3.1.2.1.2. Understanding the diagnosis and treatment assignment rules for pneumonia paediatric patients

Clinicians are supposed to use guidelines widely disseminated as the 'Basic Paediatric Protocols' in Kenya (28) that are adapted from WHO guidance, based on available evidence and developed by consensus by a national guideline panel (see (40-42)). However, studies have shown that some clinicians do not follow treatment guidelines (see Chapter 1 on non – guideline treatment approaches for pneumonia). In Kenya, as in many low and middle income countries these recommendations reflect the absence of access to further diagnostic tests. Thus pulse oximetry, blood culture or tests for inflammatory markers are not routinely available (96). As indicated in Chapter 1, clinicians may fail to adhere to guideline recommendations by making errors or over-riding recommendations at any of the three steps of assessment, severity classification and treatment assignment. However, based on the clinical symptoms and signs recorded it was possible to assign a severity classification (and thus expected treatment) based on the data. It was a data derived and investigator assigned classification as indrawing pneumonia (not necessarily the clinician defined classification) that was used in the primary analysis (experiment 1).

3.1.2.1.3. Analysis Variables

a) Outcome variable

Mortality was used as the outcome variable in all the three experiments.

b) Independent variables

These variables were grouped into key and auxiliary. Key variables were defined as those that would influence pneumonia severity classification and hence treatment based on the treatment protocol (28) (see **Box 1.1.1** in Chapter 1). Auxiliary variables were defined as those that might, *a priori,* be expected to influence treatment assignment based on clinical reasoning (for example they might make a clinician concerned for severe illness), although according to the formal rules (the guidelines) they are not considered reasons to alter treatment assignment. Such auxiliary variables were identified from those clinical symptoms and signs that were routinely collected within CIN. See **Table 3.1.1** for a summary of key and auxiliary variables that were used in the analyses.

Table 3.1.1: Summary of key and auxiliary independent variables for experiments 1, 2 and 3[11]

Experiment 1 and 2 key variables	**Experiment 3 key variables**	**Auxiliary variables for experiments 1, 2 and 3**
Age (2 – 59 months)	Age (2 – 59 months)	Gender (male/female)
Indrawing (present/absent)	Indrawing (present/absent)	Cough duration (days)
History of cough (yes/no)	History of cough (yes/no)	Crackles (present/absent)
Difficulty breathing (present/absent)	Difficulty breathing (present/absent)	Weight (Kg)
Level of consciousness – AVPU (alert/verbal response/pain response/unresponsive)	Level of consciousness – AVPU (alert/verbal response/pain response/unresponsive)	Pallor (0, +, +++)
	Oxygen ordered (yes/no)	Capillary refill (immediate, 1 – 2 secs, 3 – 6 sec, > 6 secs)
	Cyanosis (present/absent)	Fever (present/absent)
	Inability to drink/breastfeed (yes/no)	Diarrhoea (present/absent)
	Grunting (present/absent)	Convulsions (present/absent)
	Respiratory rate (breaths/min)	Vomiting (yes/no)
		Referral (yes/no)
		Length of illness (days)
		Number of fits
		Thrush (present/absent)
		Quinine/artesunate (prescribed/not prescribed)
		Weight for age z – score
		Wheeze (present/absent)
		Comorbidities (Malaria and or diarrhoea)

3.1.2.1.4. Sample size verification

Here, sample size verification uses the formula cited in (105):

$$ns = \frac{r+1}{r}\frac{\bar{p}(1-\bar{p})(Z_{\beta}+Z_{1-\alpha/2})^2}{(p_1-p_2)^2}$$, where:

[11] Experiment 3 has more key variables than experiment 2 as it considers patient populations with "very severe, severe and non –severe pneumonia" – as classified in the previous WHO and Kenyan treatment guidelines. Therefore, in addition to variables used to classify severe pneumonia, other variables used to classify very severe and non-severe pneumonia are considered.

ns = size of smaller group.

r = ratio of smaller group.

$p_1 - p_2$ = clinical difference in proportions of the outcome.

Z_β = corresponds to power of 80%.

$Z_{1-\alpha/2}$ = corresponds to two – tailed significance level (1.96 for α = 5%)

$\bar{p}$ = corresponds to average of outcome proportions in two groups.

The value for $\bar{p}$ is estimated from studies – two of which formed evidence for earlier WHO indrawing pneumonia treatment guidelines. See **Table 3.1.2** that shows the number of deaths per treatment arm reported in these studies.

Table 3.1.2: Summary of some of pneumonia studies that informed previous WHO guidelines

Study	**Treatment arms**	**Mortality**	$\bar{p}$
Shann et. al(1985)	Chloramphenicol alone	48/377	0.1470
	Chloraphenicol+Penicillin	62/371	
Addo – Yobo et. al (2002)	Injectable penicillin	7/845	0.0050
	Oral amoxicillin	2/857	
Agweyu (2015)	Injectable penicillin	3/264	0.008
	Oral amoxicillin	1/263	

For assessment of sample size for indrawing pneumonia experiments, a weighted[12] $\bar{p}$ of 0.041 from these studies was used. The ratio r is varied between 1 and 3. **Figure 3.1.1** was generated by fixing power and significance level at 80% and 5% respectively. Estimates of $\bar{p}(1-\bar{p})$ derived from WHO studies were substituted in the sample size formula and data simulated in order to see what detectable differences would be achieved by different sample sizes. A total sample size of about 4000 would be sufficient to detect a minimum difference of

[12] Weighting was done using the total sample sizes per experiment.

1.5% (absolute difference e.g. a reduction of mortality from X% to X – 1.5%) in any of these experiments[13].

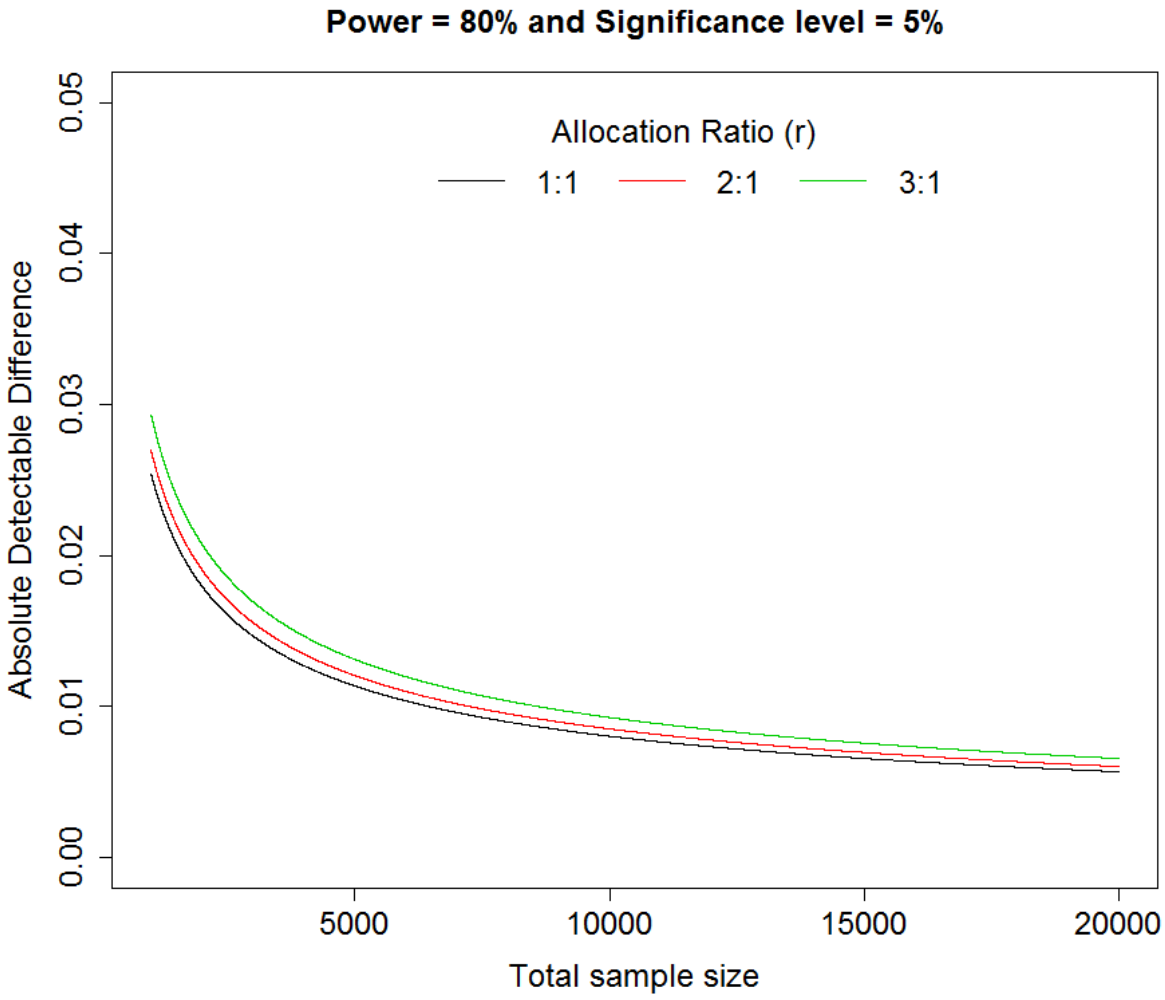


Figure 3.1.1: Sample size verification.

[13] A sample size of at least 4000 was required for experiment 3 as this was the minimum sample for experiments 1 and 2 which were nested in experiment 3.

3.1.2.2. Statistical and outcome Analysis

Statistical analysis proceeded in the following four steps:

Step 1 – subsets of patients of interest for the experiments were obtained.

- Experiment 1: First, missing clinical signs data were multiply imputed[14] (excluding outcome data[15]) – and then key clinical signs data used to construct (assign) a pneumonia severity level for all patients based on the algorithms in the pneumonia treatment protocol (28). Thereafter, a subset of patients with guideline-defined indrawing pneumonia (for each of the imputed datasets) was obtained for further analyses.
- Experiment 2: A subset of indrawing pneumonia patients (where this level of severity was indicated by the clinicians) was obtained from the raw dataset – and clinical signs data imputed using multiple imputation (without the outcome data).
- Experiment 3: The raw dataset containing all the patients with all forms of pneumonia severity was used –and clinical signs data imputed using multiple imputation (without the outcome data).

Step 2 – patient populations in the alternative treatment arms were adjusted using propensity score (PS) methods to overcome non – random treatment allocation. Standardised mean differences (and where was necessary density plots) were used as diagnostic checks for covariate balance and overlap (107, 108) between penicillin and penicillin plus gentamicin treatment groups. PS methods

[14] For the three experiments, 20 datasets were multiply imputed using chained equations (106).

[15] Rubin (2008) advises the hiding of outcome data during the design phase of comparative effectiveness studies using routine datasets and by extension this implies exclusion of outcome data in multiple imputation.

that utilised all the data (PS optimal full matching, weighting and sub classification) were examined in experiments 1 and 2 and the method that resulted in the minimum average absolute standardised mean differences for the majority of the variables and retained the largest number of patients in the analysis was considered appropriate (37). While only PS sub-classification was used for experiment 3. As experiment 3 aimed to investigate comparative effectiveness in all cases of pneumonia, propensity score was used as a proxy for disease severity thus patients with lower propensity scores were considered less ill while those with higher propensity scores were considered more ill (grouped in propensity score subclasses for analysis).

Step 3: conducting outcome analysis.

Bayesian log binomial regression models (109) were used to estimate average treatment effect for the treated[16][17]. A hospital variable was modelled as a fixed effect in the log binomial regression that measured treatment effects on pooled data. These models were fitted on each imputed dataset (adjusting for other variables used in PS models) and results pooled using Rubin rules (59).

3.1.2.3. Ethics and dissemination

The primary data are held by hospitals participating in the Kenyan Clinical Information Network (CIN) project with de-identifed data shared with the KEMRI-Wellcome Trust Research Programme for agreed analyses. The analyses described in this protocol were part of this larger project (CIN) which was approved by the Kenya Medical Research Institute Scientific and Ethical Review

[16] Treatment group (penicillin plus gentamicin) and active control group (penicillin).
[17] Bayesian models will be used to overcome any bias due to sparsity of data as PS sub-classification in itself reduces the effective sample size.

Committee (Protocol number: 2465). This committee agreed to the use of de-identified patient data derived from retrospective case record review without gaining individual patient consent as is common practice in service evaluation research. The findings were useful in understanding the external validity of current treatments – and provided a platform on which to do more similar analyses for different (combinations of) treatments.

Having presented the background and methods to this comparative analysis, the next section 3.2 presents the primary results of this analysis.

3.2. Using observational data to compare the effectiveness of antibiotic treatments for children hospitalised with pneumonia in Kenya

Section summary

Background: Kenyan guidelines for antibiotic treatment of pneumonia (based on severity) recommended treatment of pneumonia characterised by indrawing with injectable penicillin alone in inpatient settings until early 2016. At this point, they were revised becoming consistent with World Health Organisation guidance after results of a Kenyan trial provided further evidence of equivalence of oral amoxicillin and injectable penicillin. This change also made possible use of oral amoxicillin for outpatient treatment in this patient group. However, given non-trivial mortality in Kenyan children with indrawing pneumonia it remained possible they would benefit from a broader spectrum antibiotic regimen, a question not previously adequately addressed. Therefore, the objective of the present study is to compare the effectiveness of injectable penicillin monotherapy with a regimen combining penicillin with gentamicin.

Setting and Methods: I used the CIN dataset that captures data on all admissions to 13 Kenyan county (plus one sub – county) hospitals. To overcome the challenges associated with the non – random allocation of treatments and missing data in observational data, I used propensity score (PS) methods and multiple imputation to minimize bias. Selection of the study population was based on inclusion criteria typical of a prospective trial for the primary analysis (experiment 1). I also explored more pragmatic inclusion criteria (experiment 2) as part of a secondary analysis. And as an extension of the logic in experiment 2, further analysis using data on all children receiving penicillin alone or penicillin

and gentamicin in combination was conducted for all forms (levels of severity) of pneumonia in experiment 3.

Results: The analyses included children aged 2 – 59 months with the following sample sizes per experiment: experiment 1 (n = 4002), experiment 2 (n = 6420) and experiment 3 (n = 10565). Using propensity score adjusted log binomial regression, the estimated risk of dying, in experiment 1, in those receiving penicillin plus gentamicin was 1.46 [0.85, 2.43] compared to the penicillin monotherapy group. In experiment 2 (secondary analysis), the estimated risk was 1.04 [0.76, 1.40]. And in experiment 3, the benefit of using gentamicin plus penicillin was seen to increase from those who had lower propensity scores to higher as was defined by the PS subclasses.

Conclusion: There is no statistical difference in the treatment of indrawing pneumonia with either penicillin or penicillin plus gentamicin. By extension it is unlikely that treatment with penicillin plus gentamicin would offer an advantage over treatment with oral amoxicillin the currently recommended replacement therapy for penicillin. However, populations currently defined as having indrawing pneumonia may experience mortality rates as high as 2%. Efforts should be made to identify additional risk factors of severe illness that might guide clinicians on the need for an inpatient diagnostic workup and enhanced supportive care as policy now recommends outpatient treatment for indrawing pneumonia.

3.2.1. Introduction

The WHO recommended guidelines for treating pneumonia have considerable influence on policy and practice in low and middle income countries. While the evidence base and rigour of guideline development have improved considerably, there remain few data on their effectiveness when implemented in non-trial settings as discussed in sub – section 1.1.4.1 of Chapter 1. These analyses therefore aim to provide extra data from Kenya on comparative effectiveness of broader spectrum antibiotic use versus narrower spectrum antibiotic use for children with pneumonia. The rationale has been discussed in sub – section 3.1.1 of the protocol. And in brief, in a recent change to guidance, it is now recommended that pneumonia characterized by lower chest wall indrawing be treated in outpatient settings with oral medication (24, 27). Yet indrawing pneumonia remains associated with non-trivial mortality that may be higher outside trial populations (110). Residual mortality may be associated with causes that are not prevented by currently available conjugate vaccines and organisms that are not susceptible to the antibiotics currently recommended. It is therefore possible that children with indrawing pneumonia may benefit from broader spectrum antibiotic treatment. Hence I examine the comparative effectiveness of gentamicin plus penicillin versus penicillin alone. Also as demonstrated in the descriptive analysis in sub – section 1.2.6.3 of Chapter 1, clinicians use gentamicin plus penicillin for all severity levels of pneumonia – some contrary to the guidelines. Based on these analyses, if there is no advantage of adding gentamicin to the treatments, it might be possible to encourage the clinicians not to use gentamicin as to prevent overuse and avoid possible toxicity.

3.2.2. Methods

3.2.2.1. Clinical definitions of pneumonia, primary and secondary analyses.

As explained in the protocol (sub – section 3.1.1), the WHO and Kenyan pneumonia treatment guidelines are implicitly based on risk stratification of illness with children deemed at higher risk of mortality offered broader spectrum antibiotic regimens and those at lower risk narrower spectrum antibiotics (27, 28, 33, 98). Three categories of clinically diagnosed pneumonia have been presented in sub – section 1.2.5. Categorization in this sub – section outlines previous and recently revised WHO and Kenyan pneumonia treatment guidelines (27, 33). What I refer to as indrawing pneumonia may be associated with low but clinically significant mortality rates (110, 111). Prior to March 2016 recommended treatment for this group was penicillin monotherapy and the aim is to examine whether there is any advantage of broader spectrum antibiotics in this group. Since March 2016 new guidelines recommend outpatient treatment with oral amoxicillin for this group on the basis of trials suggesting equivalence of amoxicillin and penicillin. However, as indicated above in sub section 3.2.1, very few patients had been included in studies comparing narrow (amoxicillin or penicillin) and broader spectrum antibiotic regimens. As indicated above, beyond the confines of clinical trials amongst all children being treated for indrawing pneumonia, clinical outcomes (including mortality) are worse than seen in the trials (7) and clinicians are often choosing not to use a single drug regime and are in fact often opting to use the combination of gentamicin and penicillin in the group meeting criteria for indrawing pneumonia in real life settings (43). As mortality is higher in real life settings than in trials and as the possibility that

broad spectrum antibiotics could have an advantage over monotherapy with penicillin (or amoxicillin) has not been explored in previous trials, I feel that examining whether broad spectrum antibiotics confer an advantage is an important question (see sub – section 1.1.4.1 on trials included in systematic reviews of antibiotic use for pneumonia).

The ability to use routine data to compare treatment effects requires that patients with similar problems receive different treatments. Previous studies conducted in Kenya and elsewhere have indicated that clinicians often do not follow guideline recommendations in treating pneumonia (43). Variation from the guideline recommended approach can occur at the point of pneumonia severity assignment (clinicians do not follow a nationally approved protocol linking clinical signs and severity category outlined in **Box 1.1.1**) and at the point of treatment assignment (clinicians do not follow this protocol that links treatment and severity). This variability in adherence to protocols provides the opportunity for comparative effectiveness evaluation. More specifically, the adherence and non – adherence to treatment protocols by clinicians allows us to classify indrawing pneumonia admissions in two ways:

1) Those with clinical signs placing them in the group of indrawing pneumonia irrespective of the category or classification assigned to the child by the clinician.
2) Those given a clinician classification of indrawing pneumonia irrespective of the actual clinical signs observed by the clinician.

Based on these two possibilities 2 experiments (plus a third experiment based on the logic of experiment 2) were designed (see protocol above (112)) with specific objectives as follows[18]:

1) **Experiment 1:** To compare effectiveness of injectable penicillin versus penicillin plus gentamicin (both injectable) in treatment of indrawing pneumonia; where the child is identified as belonging to a population of children with indrawing pneumonia on the basis of data on their recorded clinical signs. The Experiment 1 population of indrawing pneumonia is therefore clinically consistent with the pre-2016 clinical guideline recommendations that define this condition.
2) **Experiment 2:** To compare effectiveness of injectable penicillin versus penicillin plus gentamicin in a population in which we use the clinician assigned categorisation of indrawing pneumonia, which may not be consistent with clinical guideline recommendations.
3) **Experiment 3:** To compare effectiveness of injectable penicillin versus penicillin plus gentamicin in treatment of pneumonia regardless of severity level.

I defined Experiment 1 as my primary analysis as I propose it would identify a population similar to that recruited to a randomised trial where the inclusion criteria would be based on specified clinical signs. Experiments 2 and 3 offer scenarios that may represent a more pragmatic study design with inclusion criteria based around a clinician led classification or, for experiment 3, a

[18] All children with danger signs were excluded from experiment 1 and in general (both in experiments 1 and 2), children with the following comorbidities were excluded: HIV, meningitis, tuberculosis and or acute severe malnutrition.

situation in which antibiotic selection seems divorced from recommended clinical classifications.

3.2.2.2. Statistical analysis

3.2.2.2.1. Defining per protocol and intention to treat populations

In typical randomised controlled trials, types of analyses to be conducted are defined beforehand – and this involves defining the type of patient populations that are included in the analyses. Intention to treat and per protocol populations derived from observational datasets have been described in Danaei (2013) (113). I defined per protocol and intention to treat populations based on the dates actual treatments were recorded as prescribed for patients included in my primary (experiment 1) and secondary (experiments 2 and 3) analyses. Within each experiment, and after applying inclusion and exclusion criteria, I define the per protocol population as those whose prescription of one of the two study regimens did not change during the admission. The intention to treat population is defined by the original treatment assignment and included children in whom treatment was subsequently changed (see **Figure 3.2.1** in the Results section).

3.2.2.2.2. Dealing with missing data and propensity score matching

As CIN comprises data from routine care settings it faces challenges of non – random treatment allocation and missing data. The missing data and propensity score methods for this analysis have been detailed in the protocol (112). In brief, after exploring the patient populations, 20 datasets[19] (114) were derived using multiple imputation (with chained equations) for each experiment (variables in

[19] The current literature (114) recommends the use of more than 5 imputed datasets and therefore 20 should be sufficient.

all the experiments had missing data less than 30% – see **Appendix C.1**). Clinical signs and symptoms data considered were those recorded by clinicians as patients were admitted. The multiple imputation excluded outcome data as guidance on the use of observational datasets for comparative effectiveness analysis recommends exclusion of outcome data in the design phase (39). Following this, those with missing outcome data were excluded from the analysis (missingness in the outcome data were 0.5%, 0.8% and 1.2% for experiments 1, 2 and 3).

For each imputed dataset, patient populations in the alternative treatment groups (penicillin monotherapy versus penicillin plus gentamicin) were then adjusted using propensity score (PS) methods to overcome non – random treatment allocation (steps detailed in the protocol sub – section 3.1.2.2). PS is a distance measure (37) which is used as a means to overcome allocation bias as treatment outcomes in children with similar propensity scores can then be compared. In experiments' 1 and 2 analyses, I compared three approaches to reducing possible bias based on PS – optimal full matching, weighting and sub-classification (37, 62). All are aimed at creating groups of patients that are comparable in terms of the distribution of observed signs and symptoms. For each experiment, in order to select the optimum PS implementation method, absolute standardised mean differences (ASMD) were used as diagnostic checks for covariate balance and overlap (107, 108) between the alternative treatment groups. PS methods that resulted in the minimum average absolute standardised mean differences for the majority of the variables while retaining the largest number of patients in the analysis were considered the most

appropriate (37). While only PS sub – classification was used in experiment 3. Results of all PS modelling approaches are provided in full in Chapter 4.

3.2.2.2.3. Analytic modelling

In sample size calculations conducted prior to the experiments (presented in greater detail in the protocol sub – section 3.1.2.1), it was estimated that a sample size of at least 4000 would be sufficient for the planned experiments to detect a minimum difference of 1.5% in mortality between the two treatment groups. The sample size for experiment one was 4002 and experiment two 6420 (including 3312 of those that were also in experiment 1). In other words, experiment 2 largely included those in the experiment 1 population but also children not meeting eligibility criteria for experiment 1. And experiment 3 had a sample size of 10565. For each of the experiments, after multiple imputation, multivariable log-binomial regression models were fitted to PS adjusted datasets and adjusting for all the variables also used in the PS models (also as a form of sensitivity analyses, treatment effects were estimated on PS unweighted datasets). Only pooled treatment effect estimates (across the imputed datasets) are reported.

3.2.3. Results

3.2.3.1. Creating per protocol and intention to treat populations

Examining the dates treatments were given, five treatment arms (per experimental scenario) were defined – specifically those who received: (1) penicillin alone without changes, (2) a combination of penicillin plus gentamicin without changes, (3) penicillin but switched to a combination of penicillin plus gentamicin, (4) penicillin but switched to ceftriaxone, and (5) a combination of penicillin plus gentamicin but switched to ceftriaxone (ceftriaxone is the recommended second line treatment for severe pneumonia). Therefore, per protocol analyses would compare patients in treatment arm 1 versus 2, while intention to treat analyses would compare patients in treatment arms 1, 3, and 4 versus 2 and 5 (**Figure 3.2.1**).

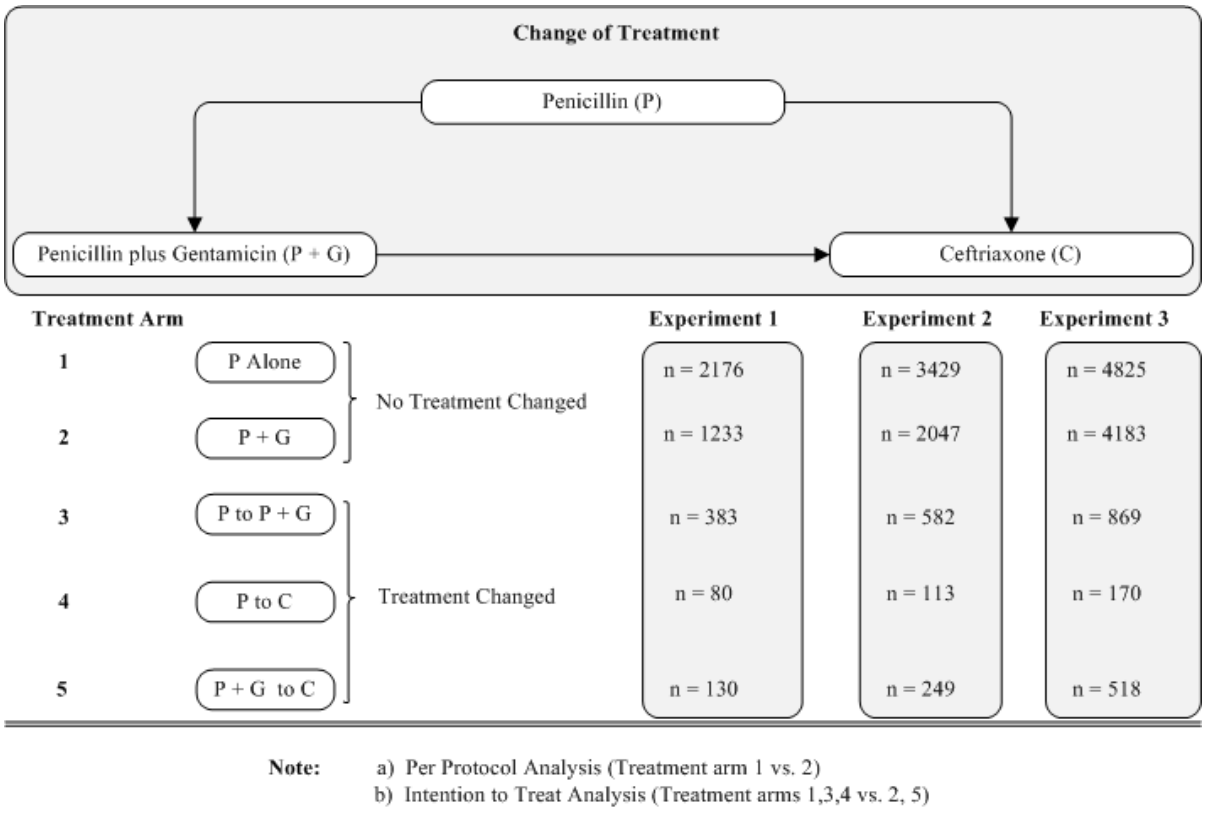


Figure 3.2.1: Summary of patients per treatment arm in experiments 1 – 3

In this analysis, intention to treat populations were considered primary and are reported in experiments 1 – 3 in keeping with clinical trial reporting guidelines. These analyses include a relatively larger number of patients compared to per protocol analyses. The recommended doses of penicillin and gentamicin in these hospitals are 50,000 iu/Kg and 7.5 mg/Kg given four times and once daily respectively. Additional data suggest most clinicians prescribed these doses correctly (see **Appendix C.2**).

3.2.3.2. Comparing performance of optimal full matching, weighting and PS sub-classification in experiments 1 and 2 respectively

For experiment 1 and 2, the three PS implementation methods were compared to determine the one which would result in the least absolute standardised mean differences for most of the variables in the analysis (even though all the three methods resulted in variables with ASMD<=10%). For experiment 1, PS weighting performed better than PS optimal full matching and sub-classification and for experiment 2, the performance of weighting was comparable to that of optimal full matching (see **Figures 3.2.2 and 3.2.3**). In both experiments, PS sub-classification reduced covariate imbalance the least. Thus, in the subsequent sections, outcome analyses are based on PS weighted datasets for experiments 1 and 2.

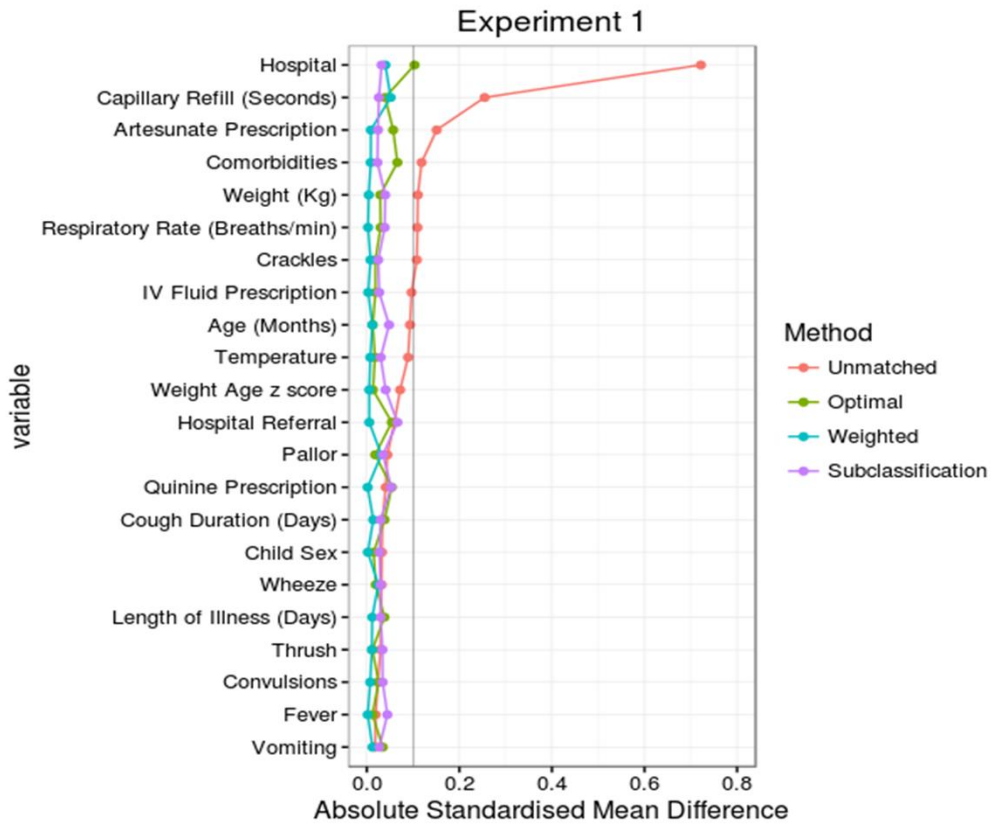


Figure 3.2.2: Comparing performance of the three PS implementation methods in experiment 1:

The y – axis contains all the variables used in the PS models. While x – axis shows absolute standardised mean difference (ASMD) which is a measure of covariate balance between the two treatment groups. An ASMD value of <= 10% indicates the method has performed well in creating comparable groups.

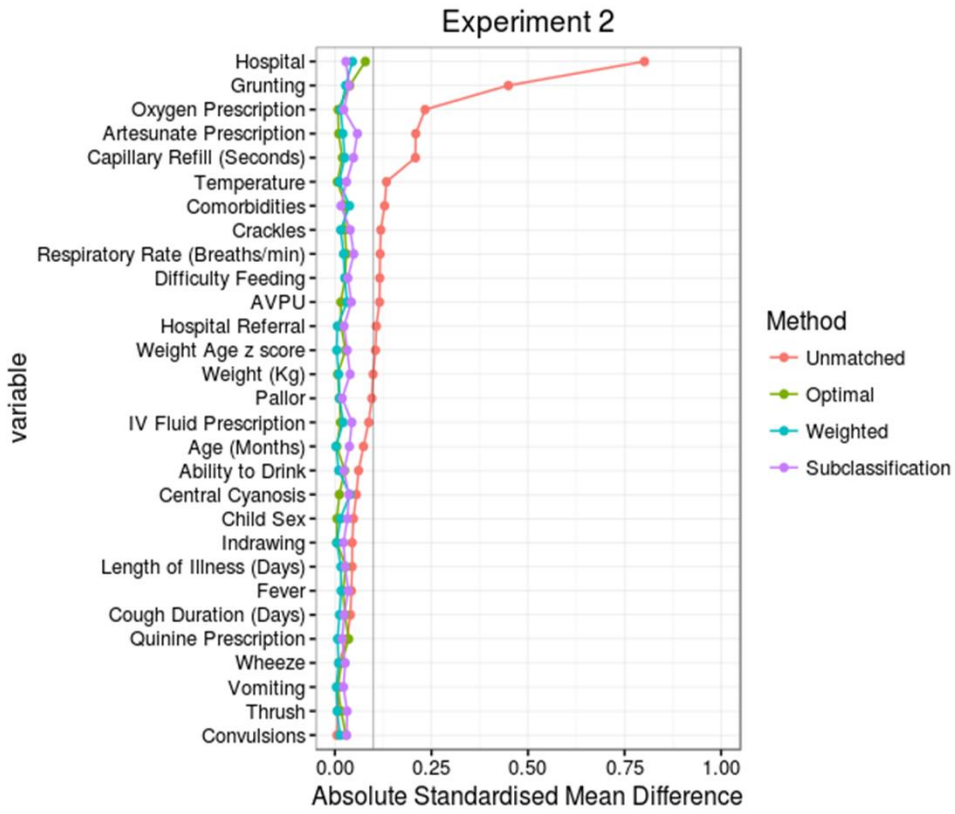


Figure 3.2.3: Comparing performance of the three PS implementation methods in experiment 2

3.2.3.3. Examining ASMD for PS sub-classification in experiment 3

Five subclasses were used in experiment 3. After estimating PS for all the patients in a dataset, the scores were then sorted and the number of patients who received penicillin plus gentamicin subdivided equally (based on PS quintiles) within each subclass while the number of patients who received penicillin alone varied in the subclasses. Only subclass 2 had all variables with ASMD < 10% (**Table 3.2.1**). While the remaining subclasses had varying number of variables with ASMD > 10% (subclass 1 (n = 3), subclass 3 (n = 2), subclass 4 (n = 4), and subclass 5 (n = 6)).

Table 3.2.1: Experiment 3 – ASMD per subclass

Variable	Subclass				
	1	2	3	4	5
Hospital	0.01	0.05	**0.12**	0.04	0.04
Gender	0.02	0.06	0.02	**0.11**	0.10
Age (Months)	0.10	0.06	0.02	0.05	0.09
Weight (Kg)	**0.11**	0.06	0.02	0.04	0.16
Central Cyanosis	0.01	0.01	0.04	0.04	0.05
Indrawing	**0.16**	0.02	**0.11**	0.06	0.22
Grunting	0.00	0.01	0.07	0.07	0.01
Pallor	0.05	0.02	0.06	0.03	0.04
Ability to Drink	0.05	0.04	0.03	0.06	0.04
AVPU	0.01	0.06	0.05	**0.15**	0.07
Respiratory (Breaths/min)	**0.16**	0.03	0.03	**0.12**	**0.19**
Capillary refill (Seconds)	0.08	0.08	0.05	0.04	**0.17**
Fever	0.01	0.07	0.10	0.04	0.03
Comorbidities	0.01	0.03	0.01	0.01	0.02
Convulsions	0.01	0.02	0.02	0.06	0.03
Vomits	0.02	0.02	0.04	0.03	0.07
Temperature	0.02	0.05	0.03	0.07	0.03
Cough Duration (Days)	0.02	0.09	0.06	0.05	0.10
Difficulty in Feeding	0.08	0.04	0.02	0.09	0.05
Crackles	0.06	0.02	0.03	0.06	0.09
Hospital Referral	0.01	0.02	0.05	0.05	0.05
Length of Illness (Days)	0.02	0.08	0.03	0.07	0.09
Thrush	0.02	0.02	0.04	0.03	0.04
Wheeze	0.02	0.05	0.05	**0.15**	**0.12**
Oxygen Order	0.00	0.01	0.02	0.02	**0.17**
IV Fluid	0.02	0.02	0.02	0.07	0.09
Quinine Prescription	0.07	0.02	0.08	0.08	0.07
Artesunate Prescription	0.02	0.03	0.07	0.02	0.02
Weight Age z score	0.08	0.04	0.07	0.07	0.02

3.2.3.4. Outcome Analysis Results

3.2.3.4.1. Exploring mortality in raw datasets

Examining the raw datasets without PS adjustments in experiment 1, the average number of pneumonia deaths (across the 20 imputed datasets) in the penicillin plus gentamicin group was 33/1363 (2.42%) and in the penicillin monotherapy group was 26/2639 (0.99%). And for experiment 2, the average number of deaths were 87/2296 (3.79%) and 50/4124 (1.21%) in penicillin plus

gentamicin and penicillin monotherapy groups respectively. Overall, the average number of pneumonia deaths in the penicillin plus gentamicin group was approximately two and a half to three times the number of mortality events in the penicillin monotherapy group in experiments 1 and 2 respectively. While in experiment 3, the average proportion of children who died increased consistently from PS subclass one to five. The denominator in the penicillin plus gentamicin group is consistent across the five subclasses – as the process of sub-classification ensured equal number of treated (penicillin plus gentamicin) individuals per subclass. While the denominator for penicillin monotherapy group reduced from subclass 1 to 5 indicating that the more severe the disease the less likely use of penicillin alone as treatment became. PS in this sense become a proxy for disease severity, children in subclass 1 were likely to have less severe pneumonia (fewer variables with a positive value that may be associated with possible risk) and children in subclass 5 were likely to have more severe pneumonia (more variables with a positive value that may be associated with possible risk) (**Table 3.2.2**). Therefore, this relationship of PS subclass with mortality is expected.

Table 3.2.2: Pneumonia deaths in experiment 3

Subclass	Penicillin plus Gentamicin	Penicillin
1	21/940 (2.23%)	23/3389 (0.68%)
2	34/940 (3.62%)	14/1262 (1.11%)
3	38/940 (4.04%)	17/695 (2.45%)
4	63/940 (6.70%)	14/376 (3.72%)
5	108/941 (11.48%)	14/142 (9.86%)
Total	264/4701 (5.62%)	82/5864 (1.40%)

3.2.3.4.2. Modelling mortality risk ratios

The analysis considered penicillin monotherapy as the reference group and mortality as the outcome – and therefore a risk ratio (RR) greater than one would be interpreted to favour penicillin over penicillin plus gentamicin. For all experiments, the treatment RRs estimated on the unmatched datasets were larger than the RR estimated on datasets obtained through PS weighting (see **Table 3.2.3** for experiment 1 and 2 results). In experiment 2, the PS unadjusted analysis showed that penicillin monotherapy was significantly more effective than penicillin plus gentamicin (1.68 [1.15, 2.36]). However, the PS weighted effect estimate (1.04 [0.76, 1.40]) was much reduced and suggested that use of PS had corrected (to a degree) for allocation bias indicating that there was no statistical difference in mortality outcomes between penicillin plus gentamicin and penicillin monotherapy treatments. It was also observed that the adjusted point estimate for any effect difference in experiment 2 (1.04 [0.76, 1.40]) was less than that in experiment 1 (1.46 [0.85, 2.43]). This may be due to an increase in the number of covariables available for PS weighting that could be used in Experiment 2 resulting in closer matching (see **Table 3.1.1** of the protocol).

Table 3.2.3: Treatment effect estimates (RR (95% C.I))

	Experiment 1	**Experiment 2**
Regression without PS adjustment	1.75 [0.94, 2.77]	1.68 [1.15, 2.36]
PS Weighted data	1.46 [0.85, 2.43]	1.04 [0.76, 1.40]

The pooled PS subclassification treatment effect estimates for experiment 3 was statistically significant (1.75 [1.36, 2.27]) – see **Figure 3.2.4**. This implies that overall the use of penicillin monotherapy has a protective effect. This most likely demonstrates that PS sub-classification did not completely eliminate imbalance

in unobserved covariate distribution in experiment 3 as clinically gentamicin plus penicillin would be expected to be at least as effective as penicillin monotherapy. However, as pneumonia severity for patients in experiment 3 varied more widely as opposed to patient populations in experiments 1 and 2 (that were considered to have more clinically similar pneumonia severity), interpretation should be restricted to the subclass level. When examined as sub-classes and moving from subclass one to five in experiment 3 – that is moving from less to more severe pneumonia - the apparent benefit from penicillin alone compared with the use of penicillin plus gentamicin gradually declined (also see the **Appendix C.3** for convergence diagnostics for the Bayesian models fitted in experiments 1 – 3).

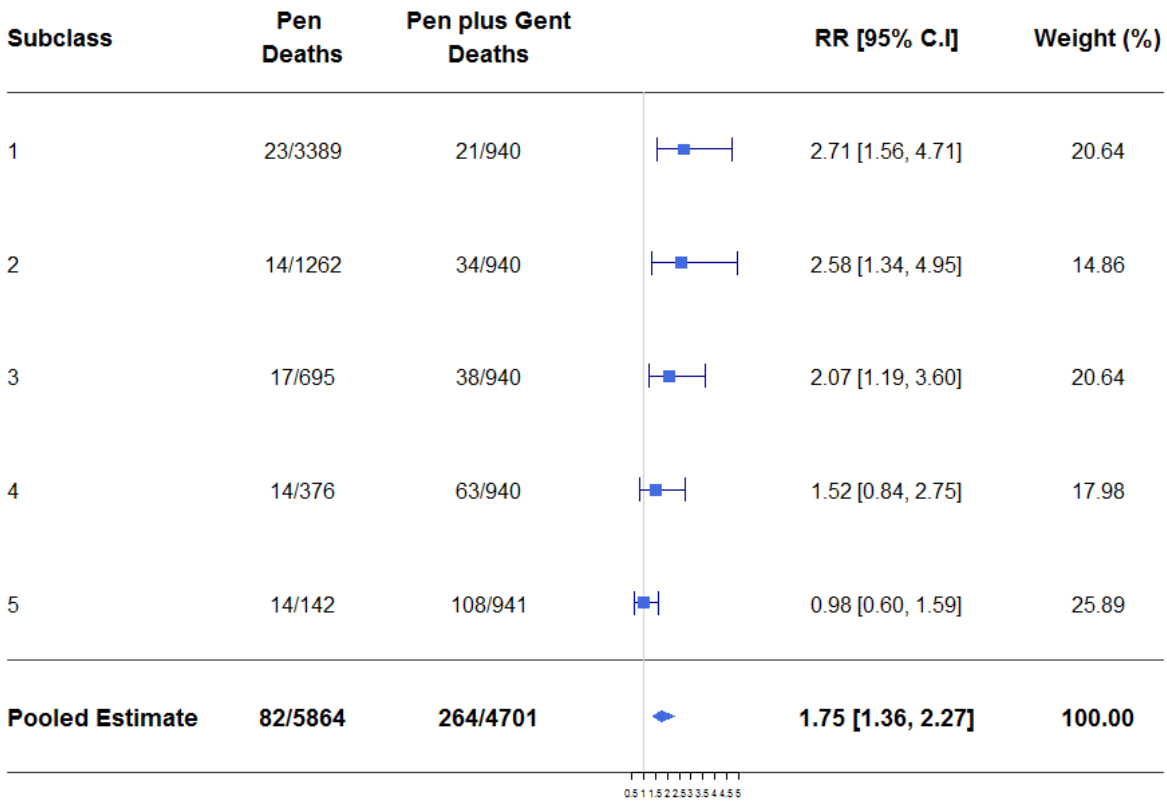


Figure 3.2.4: Experiment 3 treatment effectiveness by sublcass: The reference category in the analysis is the penicillin group – and therefore any mortality effect of RR > 1 would favour the penicillin group

3.2.4. Discussion

I compared penicillin alone with penicillin plus gentamicin in treatment of indrawing pneumonia in populations with overall mortality of 1.5% and 2% in experiments 1 and 2 respectively. There were more fatal events in the penicillin plus gentamicin group than the penicillin group (approximately 2.5 times) and unadjusted analyses pointed, therefore, to a protective effect of penicillin treatment. However, adjusted analyses, both in experiments 1 and 2, that aim to account for allocation bias that can result from non-random treatment allocation suggest that there is no appreciable difference in outcomes between penicillin and gentamicin plus penicillin when treating indrawing pneumonia.

Analyses in experiment 3 demonstrated that children may gain less from addition of gentamicin. The lower subclasses (1 – 3) showed counter intuitive results that the use of penicillin plus gentamicin was associated with worse outcomes – strongly suggesting the persistence of residual bias. One possibility would be that gentamicin plus penicillin is mostly prescribed to patients who are more likely to die – as a sense of gut feeling by the clinicians (though this is not captured as part of the observed covariates in the CIN dataset). The influence of gut feeling may extend to experiments 1 and 2 though to a lesser degree as patients included may be perceived to be as homogenous as possible in terms of danger signs.

These analyses were conducted using data from over 4,000 children, one hundred times more participants than were included in the only prior randomised controlled trial of penicillin monotherapy and penicillin plus gentamicin in treatment of pneumonia in an Asian population (115). There are continuing

concerns of clinically important mortality in children with indrawing pneumonia in Africa (111). This has led to hesitation to adopt new WHO and Kenyan guidelines that now recommend the treatment of indrawing pneumonia as an outpatient using amoxicillin (27, 33). These results suggest that there are likely to be two distinct issues. Firstly, they suggest that offering broader spectrum injectable antibiotic treatment to children with carefully assessed and classified indrawing pneumonia may not improve outcomes compared to treatment with penicillin monotherapy. As other studies have suggested equivalence between oral (high dose) amoxicillin therapy and injectable penicillin therapy (31, 100-103) it seems likely therefore that oral amoxicillin and penicillin plus gentamicin combination therapy would result in similar outcomes when used to treat carefully assessed and classified indrawing pneumonia. Clinicians should therefore carefully adhere to guidelines for treatment of indrawing pneumonia and if they do they could avoid using gentamicin helping to prevent any possible toxicity.

Secondly, however, these results suggest that children fulfilling a definition of indrawing pneumonia based on clinical signs, and having excluded serious co-morbidities, may still have an appreciable risk of mortality irrespective of their antibiotic treatment (1.5% in all children in experiment 1). When clinicians categorise children with indrawing pneumonia and imperfectly adhere to clinical sign based guidance mortality tends to be higher (2% in all children in experiment 2). These findings point to as yet uncharacterised risk factors that could be important in determining which children need admission to hospital as current guidance indicates that all those with indrawing pneumonia can now be

treated as an outpatient. While offering an alternative antibiotic to amoxicillin to this group may not improve outcomes, it is possible that closer and continuing observation in hospital may help identify co-morbid or alternative conditions that are contributing to this mortality and that may be treated.

The trials that informed the basis for the revised WHO guidelines (100-103) showed extremely low mortality (0 – 0.2%) suggesting that the populations included in such trials may not be directly representative of all those to whom guidelines are applied in routine settings. In the trial by Agweyu (2015) conducted in Kenya (which compared penicillin versus oral amoxicillin for indrawing pneumonia) overall mortality was 0.8% (31). In a parallel observational cohort providing data from the same hospitals over the same time period for children treated with penicillin alone but not included in the Kenyan trial mortality was not significantly different but marginally higher at 1.2% (Agweyu (2017), submitted) perhaps suggesting that even the limited exclusion criteria in this pragmatic trial might result in exclusion of some sicker children. Taken together with data from the analyses presented here it does appear there is a need to explore whether guidelines might be modified to accommodate additional clinical risk factors for possible life-threatening illness that should prompt admission. In a population with high coverage with conjugate vaccines this may more usefully be for more rigorous evaluation to identify alternative diagnoses or for improved supportive care than for different antibiotics as my analyses suggest limited scope for benefit from adding gentamicin to treatment.

3.2.5. Conclusion

These results suggest that children with indrawing pneumonia may gain little benefit from treatment with gentamicin plus penicillin. And as gentamicin plus penicillin had been demonstrated to be superior to chloramphenicol, we can also infer that chloramphenicol may be less beneficial too. However, the results also suggest that further work is needed to identify those who are at higher risk of death who might be prioritised for an inpatient diagnostic work up and improved supportive care rather than treated as outpatients.

Chapter 4

Sensitivity analyses on multiple imputation strategies and unmeasured confounders in comparing effectiveness of penicillin monotherapy and penicillin plus gentamicin for pneumonia treatment

---

As Chapter 3 outcome analyses were based on propensity score weighting for experiments 1 and 2, analyses using other propensity score methods, for these two experiments, including optimal full matching and sub – classification are conducted in this chapter to examine consistency of treatment effect estimates across the methods. Also as imputation methods assumed that data were missing at random (MAR), departures from this assumption are examined for the three experiments by conducting analyses under missingness not at random (MNAR). In addition, there is no clear guidance on whether to include or exclude outcome data when conducting multiple imputation and estimating propensity scores with these methods – in which this Chapter tests methods which would yield treatment estimates with minimal bias. Further, influence due to unobserved confounders is examined using propensity score trimming and instrumental variable analyses.

This chapter is therefore subdivided into two sections. First, I present sensitivity analyses focusing on testing my assumptions when using multiple imputation. Second, I conduct sensitivity analyses exploring the possibility that confounding

not addressed by the PS methods I used may bias my results. In section 4.1, I examine whether including patient's outcome in multiple imputation strategies results in treatment effect estimates that suggest improved prevention of bias. Further, I examine the amount of bias eliminated and treatment effect estimates obtained when using across and within methods for estimating propensity scores (PS) after multiple imputation. As presented in Chapter 2 sub – section 2.1.2 but briefly described here – in the across method, propensity scores are averaged across the imputed datasets – then PS adjustments are performed using the average scores before analysis of outcome data. While in the within method, PS adjustment and outcome analysis are conducted for each imputed dataset, with effects pooled using Rubin's rules (59). These methodological aspects are examined for all three PS adjustment methods – weighting, optimal full matching and sub classification. The final part of this section conducts analyses of a different kind to examine whether it is reasonable to assume a missingness at random (MAR) mechanism to generate the treatment effect estimates by exploring the possibility of missingness not at random.

As analyses of the Clinical Information Network dataset were limited to the observed variables, further analyses are conducted using PS trimming and instrumental variables in section 4.2. These methods help in examining the sensitivity of primary estimates (derived in Chapter 3) to the potential effects of unmeasured variables.

As is the case in section 4.1 and for completeness in the majority of these sensitivity analyses, I conducted analyses presented in section 4.2 using datasets

derived by including and excluding the outcome variable in the imputation models.

4.1. Exploring the sensitivity of effect estimates to different multiple imputation strategies

As explained in Chapter 1 sub – section 1.1.6, comparative evaluations using observational datasets should strive to have two phases: (a) design and (b) outcome analysis. And to have a credible design, the literature (39, 81) emphasizes that patient outcome data would not typically be known at the design phase. Just like in RCTs, therefore outcomes should only be observed when the design is fully set and this is an important feature that should be implemented when designing observational studies (39). By extension, multiple imputation methods ought to exclude outcome data in imputation models before conducting any propensity score adjustments (81). Simulation studies have been conducted by Hill (2004), Mitra (2016), and Leyrat (2017) (81, 116, 117) to examine the performance of including/excluding outcome data in imputation models and when implemented with across and within approaches for estimating PS after multiple imputation. According to Hill (2004) (117), the inclusion of outcome data in imputation models with propensity scores estimated using the within method eliminated the greatest amount of bias in PS matching. However, Mitra (2016) (81) found that exclusion of outcome data in imputation models with PS estimated using the across method had the potential of eliminating the largest amount of bias in PS matching. In a further report, Leyrat (2017) (116) demonstrated that PS weighting performed better when implemented with inclusion of outcome data in imputation models with PS estimated using the

within method. The existing literature therefore presents conflicting results suggesting there is as yet no defined best practice to estimate PS after multiple imputation, and whether to include or exclude outcome data in the imputation models. This section hence reanalyses the pneumonia data presented in Chapter 3 and explores effects of analysing treatment effects under different scenarios which I denote as follows:

- MIA – multiple imputation excluding outcome data with PS estimated using the across method.
- MIAY – multiple imputation including outcome data with PS estimated using the across method.
- MIW – multiple imputation excluding outcome data with PS estimated using the within method (which is the scenario that was used in Chapter 3 analysis).
- MIWY – multiple imputation including outcome data with PS estimated using the within method.

These analyses are conducted for each of the three PS implementation methods – PS weighting, optimal full matching and sub – classification.

4.1.1. Measuring bias

In Chapter 3 I used absolute standardised mean difference (ASMD) as a measure of possible bias to select the best performing method among the three PS methods (weighting, matching and sub – classification). ASMD illustrates any difference in distribution between treatment and (active) control groups for each variable of interest. This kind of bias measure can be used to determine a measure reflecting the percent reduction in bias. Based on the definition by Hill

(2004) (117), if a quantity (B) is measured in both treatment and control groups, then bias is defined as: $Bias\ (B) = |\bar{B}_t - \bar{B}_c|$, where $\bar{B}_t\ and\ \bar{B}_c$ are means of quantity B in treatment and control groups respectively. If $|\bar{B}_t - \bar{B}_c|$ represents bias in the raw dataset and $|\bar{B}_{tp} - \bar{B}_{cp}|$ bias in the propensity score adjusted dataset, then the percent reduction in bias (with respect to the raw dataset) is expressed as:

$$= \frac{|\bar{B}_t - \bar{B}_c| - |\bar{B}_{tp} - \bar{B}_{cp}|}{|\bar{B}_t - \bar{B}_c|}$$

I examine the percent reduction in bias for the different approaches for sensitivity analysis introduced in sub – section 4.1 to determine their different ability to minimise bias. This is examined for each of the variables used in the respective PS models for experiments 1 – 3.

4.1.2. Performance of different methods to minimise bias

For each of the four scenarios (MIW, MIWY, MIA, MIAY), the analysis process follows the steps presented in the protocol in Chapter 3, also summarised in **Box 4.1.1**.

Box 4.1.1: Summary of analysis steps

Step 1 – obtaining subsets of patients.

- Experiment 1: First, missing clinical signs data are multiply imputed – and then key clinical signs data are used to construct (assign) a pneumonia severity level for all patients based on the algorithms in the pneumonia treatment protocol (28). Thereafter, a subset of patients with guideline-defined indrawing pneumonia (for each of the imputed datasets) is obtained for further analyses.
- Experiment 2: A subset of indrawing pneumonia patients (where this level of severity is indicated by the clinicians) is obtained from the raw dataset – and clinical signs data are imputed using multiple imputation.
- Experiment 3: The complete dataset containing all the patients with all forms of pneumonia severity is used –and clinical signs data imputed using multiple imputation.

Step 2 – propensity score methods (weighting, full matching and sub-classification) are used in an effort to create balanced treatment groups and so overcome non – random treatment allocation. Bias reduction is then calculated for each PS method, with only PS sub-classification methods used in experiment 3 as previously. It is important to note that comparisons between different sensitivity analyses are done within each experiment.

Step 3: conducting outcome analysis for each scenario. Bayesian log binomial regression models are used to estimate average treatment effects for the treated[20].

Using the method outlined above in sub – section 4.1.1, I estimated percent bias reduction for each of the four scenarios (MIA, MIAY, MIW, MIWY) also defined at the beginning of this section (see results presented in **Appendices D.1, D.2 and**

[20] Treatment group (penicillin plus gentamicin) and active control group (penicillin).

D.3). With these results and for each experiment, I define the following distances:

- D 1 to represent the difference in percent bias reduction between MIAY and MIA scenarios. Here, more positives (on more than half of the variables) would imply that inclusion of outcome data in imputation models contributes to prevention of bias when PS are estimated with the across method;
- D 2 – difference in percent bias reduction between MIWY and MIW scenarios – with more positives implying the benefit of outcome inclusion in the imputation models when PS are estimated using the within method;
- D 3 – Difference in percent bias reduction between MIWY and MIAY (analyses with outcome included in imputation);
- D 4 – Difference in percent bias reduction between MIW and MIA (analyses with outcome excluded in imputation).

The distances D 3 and D 4 aim to examine the benefit of using the within method compared to using the across method of PS estimation. Based on these measures (D 1 – D 4) of bias reduction and as can be deduced from **Tables 4.1.1 – 4.1.3** (across the three experiments defined in Chapter 3 section 3.1), the main findings of these sensitivity analyses across these scenarios suggest that in:

- **PS weighting** – inclusion of outcome data in imputation models contributes to prevention of bias compared to the scenario where outcome data are excluded. And estimating PS using the within method gives more desirable results than estimation using the across method.

- **Optimal full matching** – the across method performs better with the exclusion of outcome data in imputation models – while the within method seems to perform better with the inclusion of outcome data in the imputation models. PS estimation using the across method seems to give better results than when estimated using the within method.
- **Sub – classification** – comparable results are achieved when outcome data are excluded and included in the imputation models. The across method does seem to perform better than the within method of PS estimation.

Table 4.1.1: Estimation of distances D1, D 2, D 3 and D 4 in experiment 1

Variables	PS weighting				PS optimal full matching				PS sub – classification			
	D 1	D 2	D 3	D 4	D 1	D 2	D 3	D 4	D 1	D 2	D 3	D 4
Child sex	12.3	12.3	-1.9	-1.9	-2.1	232.9	-228.0	-463.0	-25.9	-76.3	-28.9	21.5
Age (months)	-4.3	9.2	10.7	-2.9	-0.5	-15.9	-138.2	-122.8	-60.7	-5.1	-54.9	-110.5
Weight (Kg)	5.6	15.6	8.4	-1.7	8.4	-12.5	-84.3	-63.4	12.0	5.9	-91.2	-85.1
Pallor	35.4	19.4	-3.4	12.6	3.3	20.0	-29.2	-45.9	31.1	42.6	-60.2	-71.8
Respiratory rate	-8.6	-6.1	6.4	3.9	6.2	14.0	-1.9	-9.7	-4.1	9.1	-9.5	-22.7
Difficulty in breathing	-6.3	-5.0	3.5	2.2	-2.5	6.4	-179.5	-188.4	-1.0	-1.1	0.4	0.5
Capillary refill	1.3	2.8	11.7	10.3	-0.6	8.3	-51.5	-60.4	3.3	-2.6	-21.5	-15.6
Fever	249.3	396.5	112.4	-34.8	182.2	768.3	-849.1	-1435.1	328.1	485.4	-398.8	-556.0
Convulsions	6.5	-56.2	-86.6	-23.9	121.4	39.3	370.5	452.6	-56.8	-69.8	272.5	285.5
Vomiting	-157.2	-305.1	-425.5	-277.6	-483.3	-1166.1	-1552.2	-869.4	-233.3	-269.7	48.6	85.1
Cough duration (days)	3.7	11.1	-3.8	-11.2	35.4	-49.4	-153.1	-68.3	6.6	-10.6	55.9	73.1
WAZ score	-0.3	0.4	2.2	1.5	-0.3	-1.8	-13.2	-11.7	0.2	0.4	-10.2	-10.5
Co – morbidities	-4.2	0.3	5.1	0.6	-3.1	10.0	65.7	52.6	-3.1	-37.8	-96.5	-61.8
Crackles	7.0	0.7	-2.8	3.5	13.8	-10.3	-124.9	-100.8	0.2	33.6	-72.6	-105.9
Hospital referral	-11.9	-2.4	22.9	13.5	-34.0	-21.4	-74.7	-87.4	-5.5	24.5	2.6	-27.4
Length of illness	-2.0	100.8	-25.0	-127.8	336.0	206.9	-165.2	-36.1	451.5	100.3	34.9	386.2
Thrush	18.5	33.5	-3.5	-18.6	12.8	35.5	-83.0	-105.7	27.7	-15.7	-132.9	-89.4
Wheeze	-26.9	-0.4	10.0	-16.6	-46.4	-59.3	-80.7	-67.9	-34.6	-42.2	62.7	70.3
% of positive distances	**50%**	**67%**	**63%**	**50%**	**50%**	**63%**	**13%**	**13%**	**50%**	**44%**	**39%**	**39%**

* The numbers presented in the table are expressed in terms of percentages. Positive (highlighted in blue) values indicate benefit of a method while negative (highlighted in brown) shows reduced benefit. D 1 is the difference in percent reduction in bias between MIAY and MIA; D 2 = MIWY – MIW; D 3 = MIWY – MIAY and D 4 = MIW – MIA.

* In estimating the distances (D 1, D 2, D 3 and D 4), it is possible to have large percentage reduction/increase in bias even if the estimated ASMDs both in the raw data and PS adjusted datasets are small (< 0.1). For example, if ASMDs in raw and PS adjusted datasets are 0.0001 and 0.01, then the estimated percentage increase in bias would be 9900% - though both ASMDs are within tolerable limits of imbalance (< 0.1).

Table 4.1.2: Estimation of distances D1, D 2, D 3 and D 4 in experiment 2

Variables	PS weighting				PS optimal full matching				PS sub – classification			
	D 1	D 2	D 3	D 4	D 1	D 2	D 3	D 4	D 1	D 2	D 3	D 4
Child sex	0.3	-7.4	-0.3	7.3	-12.1	-20.8	-295.9	-287.2	-14.6	3.4	-93.4	-111.4
Age (months)	-0.2	12.9	18.0	5.0	-5.9	-23.0	-124.9	-107.9	-6.7	-5.6	-92.3	-93.4
Weight (Kg)	2.3	9.7	9.6	2.3	-16.5	-33.1	-72.7	-56.1	-12.1	-0.8	-76.4	-87.7
Pallor	5.8	11.6	4.8	-1.0	-19.5	-21.2	-27.8	-26.1	-2.0	14.0	-7.3	-23.4
Respiratory rate	-4.7	-13.4	-11.3	-2.7	-14.9	-5.5	-62.9	-72.4	2.3	-9.7	-12.7	-0.8
Difficulty in breathing	-1.9	-1.6	-2.9	-3.2	-7.7	-22.7	-63.8	-48.7	4.2	3.9	-1.1	-0.8
Capillary refill	-1.4	1.5	-3.9	-6.8	4.1	7.2	-43.2	-46.3	-0.7	-0.7	-25.7	-25.7
Fever	-1.5	-5.5	6.4	10.4	21.9	2.8	-154.1	-135.1	-0.6	-16.5	-92.1	-76.3
Convulsions	31.6	58.7	46.3	19.2	-674.4	206.9	886.6	5.3	153.2	-336.1	-557.1	-67.8
Vomiting	408.1	303.0	10.8	115.9	1924.4	-412.9	-3408.6	-1071.3	906.9	1562.3	-2919.9	-3575.4
Cough duration (days)	-2.0	9.4	17.9	6.5	-5.5	165.8	-2559.2	-2730.6	-9.9	28.6	-173.9	-212.4
WAZ score	-1.1	-0.2	3.5	2.6	1.6	-3.1	-8.8	-4.1	-9.8	-0.8	-4.8	-13.8
Co – morbidities	-0.1	0.1	1.0	0.8	16.0	-8.5	35.6	60.0	0.6	-15.1	-63.7	-48.0
Crackles	-0.9	1.4	-5.6	-8.0	6.4	29.0	-21.5	-44.1	2.4	9.0	-8.0	-14.6
Hospital referral	2.4	-1.9	10.1	14.5	9.9	-26.0	-68.6	-32.7	5.3	-6.3	-6.3	5.2
Length of illness (days)	-7.8	11.3	14.7	-4.4	-13.3	75.4	-1790.7	-1879.5	12.5	27.7	-93.0	-108.3
Thrush	525.3	1546.4	186.2	-834.9	1048.3	1827.5	-87.1	-866.3	1190.4	2255.0	-656.0	-1720.7
Wheeze	14.1	25.6	15.2	3.7	-46.0	185.1	-1056.5	-1287.6	37.9	-34.1	-26.5	45.5
Indrawing	15.5	4.8	-27.0	-16.3	45.1	147.8	-283.5	-386.2	12.1	-2.8	-34.8	-19.8
Oxygen order	-0.7	2.6	-0.7	-4.0	1.3	10.1	-94.0	-102.8	-1.0	-4.3	-22.8	-19.4
Ability to drink	6.3	-11.9	-3.2	14.9	-12.0	-81.1	-61.3	7.8	-4.1	-57.2	-30.7	22.4
Central cyanosis	8.9	11.2	0.8	-1.5	-40.0	336.5	-1857.3	-2233.8	-9.3	-23.4	-128.7	-114.6
Grunting	0.5	0.4	-5.5	-5.5	-0.7	11.0	-44.1	-55.7	-2.6	5.8	-31.2	-39.6
% of positive distances	**52%**	**65%**	**61%**	**52%**	**43%**	**52%**	**9%**	**13%**	**47%**	**39%**	**0%**	**13%**

Table 4.1.3: Estimation of distances D1, D 2, D 3 and D 4 in experiment 3

Variables	PS sub-classification			
	D 1	D 2	D 3	D 4
Child sex	44.5	-15.3	-61.0	-1.2
Age (months)	0.1	-6.7	-10.8	-4.0
Weight (Kg)	-1.7	4.1	-1.6	-7.3
Pallor	10.5	30.6	-10.8	-30.9
Respiratory rate	-0.2	1.1	6.9	5.7
Difficulty in breathing	-0.4	-5.0	-7.0	-2.3
Capillary refill	0.8	4.9	-3.7	-7.8
Fever	8.3	-36.8	-22.9	22.2
Convulsions	-14.0	-21.2	-14.5	-7.3
Vomiting	5.4	-6.4	-35.2	-23.4
Cough duration (days)	-18.9	127.9	-33.4	-180.2
WAZ score	-0.4	-0.2	-1.8	-2.0
Co – morbidities	3.3	28.6	90.2	64.9
Crackles	0.4	0.3	-1.9	-1.8
Hospital referral	-3.9	-1.4	0.7	-1.9
Length of illness (days)	-69.4	42.1	-35.5	-147.1
Thrush	-475.1	-296.6	35.9	-142.5
Wheeze	-0.5	-17.7	-1.3	15.8
Indrawing	0.7	-3.5	-2.0	2.1
Oxygen order	-0.8	-2.6	-9.4	-7.5
Ability to drink	3.8	1.2	0.6	3.2
Central cyanosis	-0.4	11.2	-27.2	-38.7
Grunting	0.2	-4.6	-13.0	-8.3
% of positive distances	**43%**	**43%**	**22%**	**26%**

4.1.3. Treatment effect estimates obtained in the various sensitivity analysis approaches

Next, in this sub – section I present the treatment effect estimates in **Tables 4.1.4 – 4.1.6** for PS weighting, optimal full matching and sub – classification under the defined scenarios above (MIA, MIAY, MIW and MIWY). All the treatment effect estimates in experiments 1 and 2 (**Tables 4.1.4 and 4.1.5**) support the conclusions in Chapter 3 of no statistical difference between penicillin monotherapy and penicillin plus gentamicin combination therapy in treatment of indrawing pneumonia (Chapter 3 results are highlighted in **Tables**

4.1.4 and 4.1.5). Also approximately similar results (to those obtained in Chapter 3 – highlighted in Table 4.1.6) were observed with sub classification under the four scenarios of MIA, MIAY, MIW and MIWY in experiment 3 – that is the apparent benefit from penicillin alone compared with the use of penicillin plus gentamicin gradually declined moving from less to more severe pneumonia (subclass 1 through to 5) – see Table 4.1.6.

Table 4.1.4: Treatment effect estimates (experiment 1)

	Weighting	Full matching	Sub classification
MIA	1.18 [0.61, 2.26]	0.83 [0.50, 1.32]	1.31 [0.70, 2.44]
MIAY	1.07 [0.58, 1.99]	0.97 [0.58, 1.62]	1.38 [0.73, 2.61]
MIW	1.46 [0.85, 2.43]	1.28 [0.79, 2.07]	1.46 [0.82, 2.61]
MIWY	1.23 [0.69, 2.20]	1.05 [0.65, 1.71]	1.31 [0.72, 2.38]

* The highlighted estimates are derived from the models used in the primary analyses in Chapter 3

Table 4.1.5: Treatment effect estimates (experiment 2)

	Weighting	Full matching	Sub classification
MIA	1.05 [0.69, 1.58]	0.92 [0.69, 1.24]	1.35 [0.91, 2.00]
MIAY	1.06 [0.70, 1.57]	1.10 [0.80, 1.51]	1.47 [0.99, 2.18]
MIW	1.04 [0.76, 1.40]	1.21 [0.87, 1.66]	1.56 [0.92, 2.37]
MIWY	1.05 [0.71, 1.54]	1.18 [0.83, 1.66]	1.45 [0.97, 2.18]

Table 4.1.6: Treatment effect estimates (experiment 3)

	MIA	MIAY	MIW	MIWY
Subclass 1	3.77 [1.29, 11.01]	3.29 [1.20, 8.94]	2.71 [1.56, 4.71]	3.05 [1.08, 8.65]
Subclass 2	2.16 [1.28, 3.63]	1.99 [1.11, 3.60]	2.58 [1.34, 4.95]	2.28 [1.32, 3.94]
Subclass 3	2.12 [1.18, 3.79]	2.00 [1.17, 3.44]	2.07 [1.19, 3.60]	1.90 [1.11, 3.27]
Subclass 4	2.29 [1.25, 4.16]	2.10 [1.18, 3.71]	1.52 [0.84, 2.75]	1.97 [1.10, 3.51]
Subclass 5	0.87 [0.55, 1.42]	0.90 [0.57, 1.43]	0.98 [0.60, 1.59]	0.98 [0.61, 1.58]
Pooled	1.69 [1.31, 2.21]	1.64 [1.28, 2.12]	1.75 [1.36, 2.27]	1.69 [1.31, 2.18]

4.1.4. Exploring whether analyses assuming MAR are appropriate

The plausibility of data missing at random (MAR) is examined by conducting analyses in the framework of a missingness not at random (MNAR) mechanism. As defined in Chapter 2 in sub – section 2.1.2, in MAR mechanism, observed data may be used to predict the missing values as part of multiple imputation. In the analysis of CIN data in Chapter 3, where there were missing data, I assumed the clinicians unintentionally failed to record such clinical data or fill some sections of the treatment sheets potentially resulting in data that are missing at random (and the missingness process is therefore ignorable). Under this assumption, the recorded data by clinicians would plausibly be predictive of missing data. However, it is also possible that data are missing for reasons that are not random but linked to the data recording process. For example, it may be related to unavailability of devices (such as thermometers) that support clinical measurements. These reasons are not necessarily known but could create a scenario in which data are missing not at random. The use of patterns of missingness within the data supposes that there are unmeasured factors that create much more missingness in some medical records than in others so there is an underlying non – random factor. Thus conducting multiple imputation and subsequent analyses on data subsets that display different patterns of missingness may provide an indication (if results vary) that assumption of MAR may not be feasible. The results from analyses of subsets of data with different missingness patterns may then be pooled across the different scenarios (118) and compared with results obtained under MAR mechanism. The MAR mechanism may be considered plausible if MAR and these MNAR results give similar

conclusions (119). Otherwise results may have to be interpreted in the context of a MNAR mechanism.

Even though there are different models (see (94, 118) for a discussion) that may be fitted in the MNAR context, Molenberghs (2004) (94) indicate that the use of pattern mixture models may play a major role in missing data sensitivity analysis. Pattern mixture models are implemented in Bell (2014) where it is indicated that fitting of other MNAR models may be complicated and require specialised software (118). In this analysis, pattern mixture models involve formulation of different missing data patterns – and outcome analyses are then conducted for each pattern and pooled. This form of grouping of patients for MNAR was implemented in Gathara (2017)(53). The pneumonia analysis involved missing data in more than one covariate and therefore patterns are formulated by grouping of patients according to the number of variables for which they have missing data. In particular, three patterns are formulated in which group 1 consists of patients with missing data in 0 – 3 variables, group 2 consists of patients with 4 variables with missing data, and group 3 consists of patients with more than four variables with missing data (there were 18 and 23 variables in experiments 1 and 2). For illustrative purposes, analyses using pattern mixtures are only conducted for PS weighted and matched datasets for experiments 1 and 2 in the MIW scenario. Data are first multiply imputed within each group (pattern) separately (using 20 datasets), then propensity score and outcome analyses are also conducted on each of these groups / missingness patterns separately (this reduces the power of any one analysis but the aim is to check consistency of results). The results of estimates of treatment effect for each

pattern are then pooled using meta-analysis, by weighting the estimates using the proportion of patients in the gentamicin plus penicillin group in each pattern (61).

The number of patients in the gentamicin plus penicillin treatment group in each pattern is approximately half of those in the penicillin alone group – both in experiments 1 and 2 (see **Tables 4.1.7 and 4.1.8**).

Table 4.1.7: Experiment 1 distribution of patients by treatment group per pattern

Pattern	Penicillin alone		Penicillin plus Gentamicin		Total	
	N	%	N	%	N	%
1	1408	54.23	818	58.18	2226	55.62
2	584	22.50	288	20.48	872	21.79
3	604	23.27	300	21.34	904	22.59
Total	2596	100.00	1406	100.00	4002	100.00

Table 4.1.8: Experiment 2 distribution of patients by treatment group per pattern

Pattern	Penicillin alone		Penicillin plus Gentamicin		Total	
	N	%	N	%	N	%
1	2067	50.16	1220	53.07	3287	51.20
2	923	22.40	432	18.79	1355	21.11
3	1131	27.44	647	28.14	1778	27.69
Total	4121	100.00	2299	100.00	6420	100.00

In each pattern (both for experiments 1 and 2), the performance of propensity score full matching and weighting methods were compared. As demonstrated in **Figures 4.1.1 and 4.1.2**, both of the methods managed to minimise imbalance of covariate distribution between gentamicin plus penicillin and penicillin monotherapy in all the patterns (the ASMD for all the variables <=10%).

Figure 4.1.1: Experiment 1 balancing of covariates by patterns

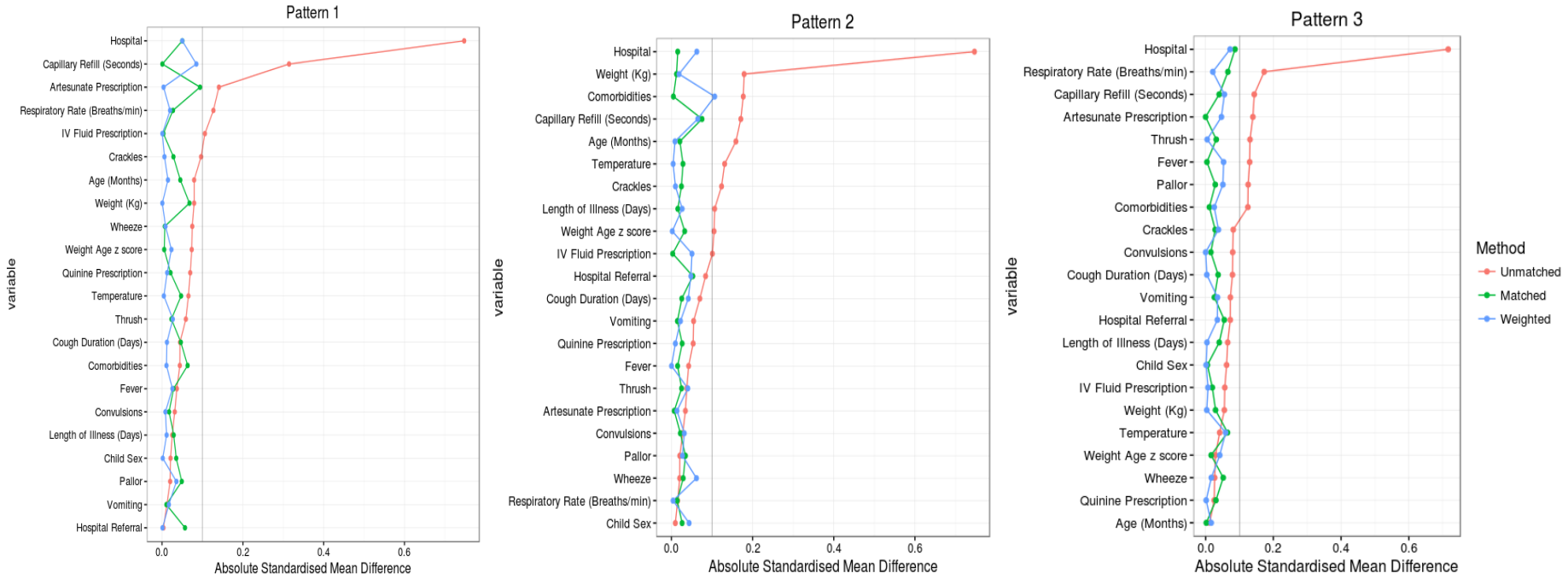
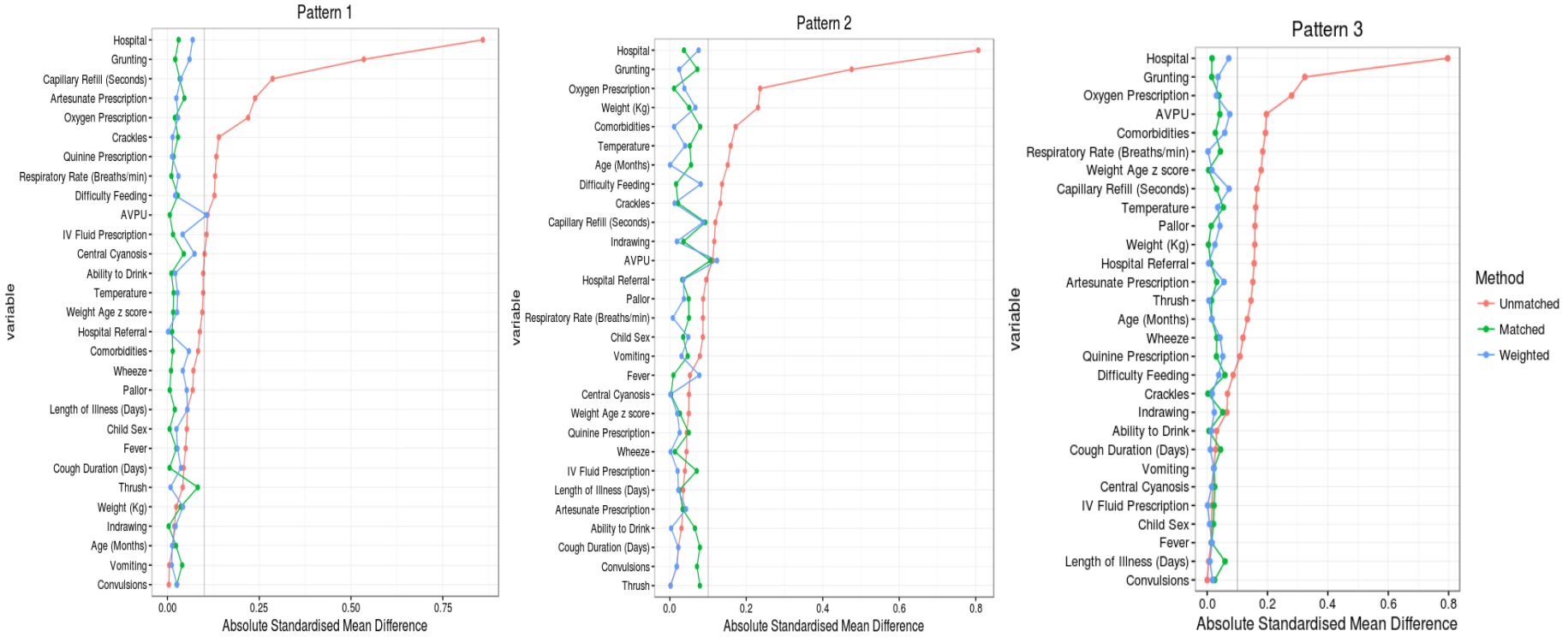


Figure 4.1.2: Experiment 2 balancing of covariates by patterns



The pooled treatment effects across patterns 1 – 3 (**Tables 4.1.9 and 4.1.10**), for experiments 1 and 2, were consistent with those obtained in the primary analysis under a missing at random assumption in Chapter 3. These estimates are larger compared to corresponding estimates derived in primary analysis – which I call MAR estimates – but show no statistically significant differences in effect of the two antibiotic regimens. The loss of power in the analysis is evidenced by larger credible intervals. In case the credible intervals of the MAR and pooled MNAR do not overlap then this may indicate that data are potentially MNAR (93). However, the overlap of estimates obtained from analyses linked to these two mechanisms is substantial suggesting it is reasonable to assume MAR in the data used in this thesis.

Table 4.1.9: Experiment 1 treatment effect estimates

Pattern	Unmatched data	Matched data	Weighted data
1	1.86 [0.90, 4.10]	1.58 [0.83, 3.06]	1.97 [0.91, 4.57]
2	2.61 [0.76, 9.03]	1.39 [0.47, 3.71]	2.36 [0.68, 10.59]
3	1.92 [0.68, 5.05]	1.65 [0.64, 4.06]	1.72 [0.65, 4.76]
Pooled	2.01 [0.82, 5.05]	1.55 [0.70, 3.35]	1.97 [0.80, 5.47]
MAR estimates	1.75 [0.94, 2.77]	1.31 [0.80, 1.92]	1.46 [0.85, 2.44]

***** The MAR row is the result from Chapter 3 and it is displayed to allow comparison with the pooled estimates from pattern specific estimates.

Table 4.1.10: Experiment 2 treatment effect estimates

Pattern	Unmatched data	Matched data	Weighted data
1	2.08 [1.20, 3.63]	1.63 [0.99, 2.61]	1.48 [0.94, 2.36]
2	3.19 [1.07, 9.49]	2.29 [1.04, 5.00]	2.03 [0.82, 5.37]
3	1.58 [0.75, 3.46]	1.60 [0.85, 3.00]	1.54 [0.79, 3.03]
Pooled	2.08 [1.03, 4.31]	1.72 [0.96, 3.06]	1.57 [0.88, 2.94]
MAR estimates	1.68 [1.15, 2.36]	0.92 [0.69, 1.20]	1.04 [0.76, 1.40]

In summary, pattern mixture analysis has demonstrated that missing data patterns do not influence the overall treatment estimates – thus it would be plausible to consider the variables to be missing at random (MAR).

4.2. Exploring the potential for unmeasured confounding to bias treatment effect estimates

Propensity score methods aim to generate matched treated and (active) control patients whose distribution of measured covariates are as similar as possible. However, two patients with similar covariate distribution may still differ in terms of unmeasured variables – and this may introduce bias in estimated treatment effects (73).

Biologically, the effectiveness of gentamicin plus penicillin (when administered in correct doses) is expected to be the same or greater than that of penicillin monotherapy. The addition of gentamicin to penicillin has a possible advantage against some micro-organisms (Gram negative – which include *Legionella pneumophila, Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, and Gram positive for example *Staphylococcus aureus*) potentially causing indrawing pneumonia (120). Thus my hypothesis is that penicillin plus gentamicin might confer an advantage in terms of survival. However, it is important to explore for unexpected results and so the analyses are based on two – tailed tests of significance. Should I find no difference between the treatment arms, then the absence of difference may reflect true lack of a biological advantage in this population or might reflect inadequate power to detect only a small effect. Otherwise should I find better outcomes for penicillin than for penicillin plus gentamicin then biologically this would be an unexpected finding – given the known adverse effects of gentamicin. Such a finding would perhaps raise greater concerns about possible residual bias given the possibility that sicker children might preferentially be treated with penicillin plus gentamicin. The use of PS

methods should help overcome treatment allocation bias but may not eliminate residual bias. When interpreting the findings of the comparative effectiveness, it will therefore be important to examine what effect PS adjustment has both on the statistical significance and the direction of any effect difference. In the particular case where there appears to be an advantage of using penicillin alone then this finding might particularly raise concern over residual bias.

In Chapter 3, I describe experiments comparing mortality in groups receiving penicillin and penicillin plus gentamicin combination therapy. In these experiments, risk ratio values lower than 1 suggest a lower mortality favouring penicillin plus gentamicin over penicillin monotherapy[21] in treatment of indrawing pneumonia. In the Experiment 2 results (on unmatched data) risk ratios greater than 1 suggest that penicillin was significantly more effective than penicillin plus gentamicin. However, the use of propensity score methods (optimal full matching, weighting and sub-classification) minimised imbalance, though to different extents, on observed variables between patients in the two treatment groups. As a result, differences in observed treatment effects lessened (risk ratios approximated 1) and were no longer significant. Although not statistically significant in PS adjusted models the direction of effects in most of the analyses reported in Chapter 3 suggested better outcomes with penicillin monotherapy potentially indicating the presence of hidden bias in the data (see highlighted results in **Tables 4.1.4 and 4.1.5**). This is because biologically effectiveness of penicillin and gentamicin would be expected to be at least that of penicillin monotherapy. I therefore conduct sensitivity analysis (using PS

[21] Penicillin monotherapy was used as the reference category for the treatment variable.

trimming and instrumental variables – described below) to examine the reliability of estimates of the primary results obtained in experiments 1 and 2.

4.2.1. PS trimming

When using routine datasets in comparative analyses it would be ideal to have treatment groups with 100% overlapping distributions of background characteristics. However, complete overlap may not be achieved and therefore Stürmer (2010) (121) suggests exclusion of patients outside the overlapping region (where there is an inability to find appropriately matched patients). One possibility explaining a tendency to see poorer outcomes in children treated with penicillin and gentamicin is that clinicians' treatment assignment is skewed such that patients who appear sicker (having a greater number of clinical signs of more severe illness) are assigned the 'stronger' or broad spectrum treatment. In this situation as mentioned by Stürmer (2010), specific types of treatment allocation may be more likely associated with increased mortality.

The potential for skewed assignment of gentamicin and penicillin to sicker children can be examined by exploring the distribution of PS in the experimental populations. In **Figures 4.2.1 – 4.2.3** the skewness of curves for the two treatment groups are in opposite directions – with that of penicillin plus gentamicin skewed towards higher propensity scores that are linked to presence of a greater number of positively identified clinical characteristics of pneumonia. In theory, the use of propensity scores is supposed to account for this skewed assignment by comparing only outcomes of those with similar propensity scores that are then assumed to suggest they have similar clinical profiles and thus similar risks (and the different matching methods achieve this in slightly

different ways). Trimming takes this approach one step further by excluding patients who are likely outliers (at PS distribution tails) and creates a sample with clinical characteristics that are more homogeneous.

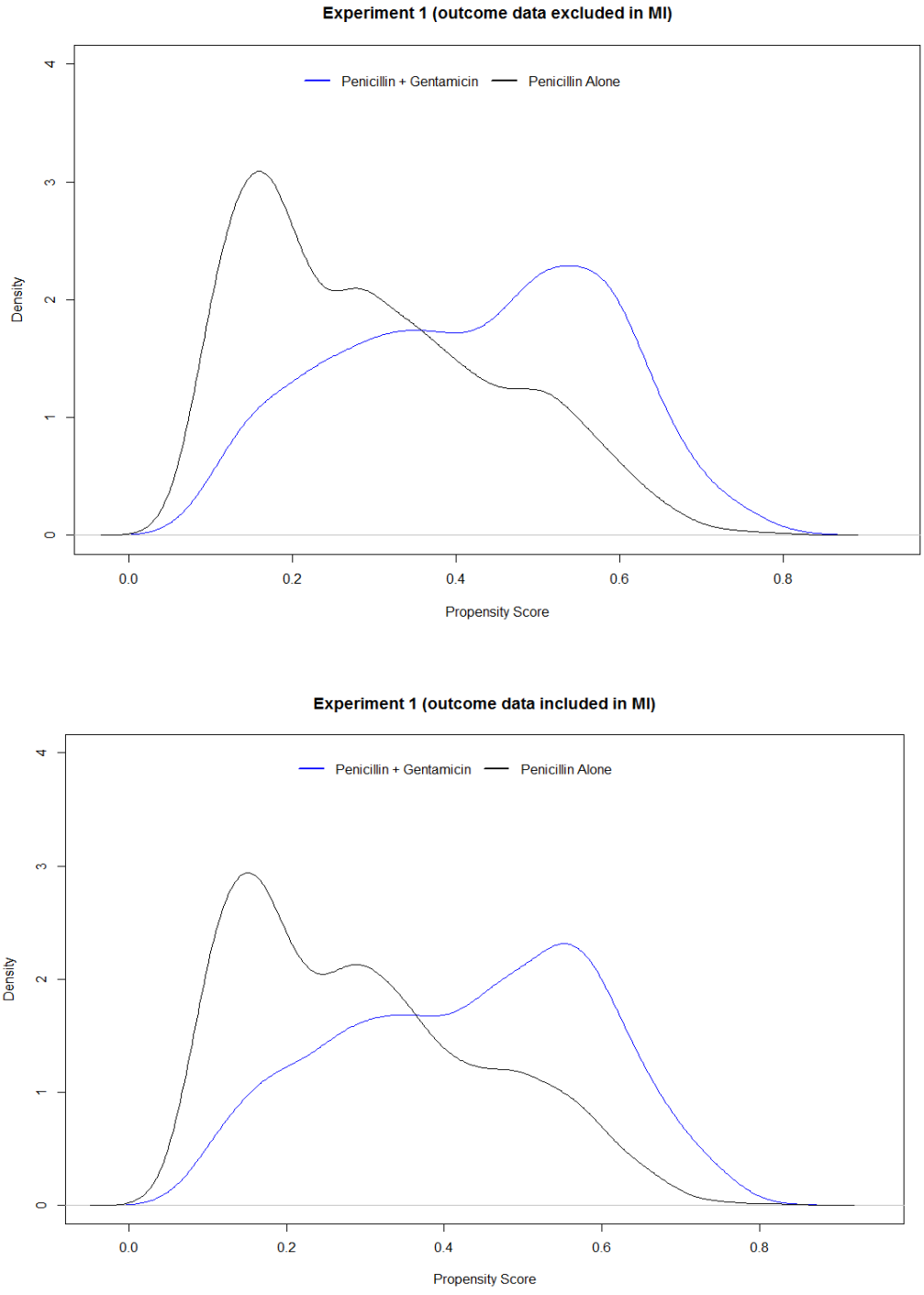


Figure 4.2.1: Distribution of propensity scores for penicillin montherapy and penicillin plus gentamicin treatment groups in experiment 1.

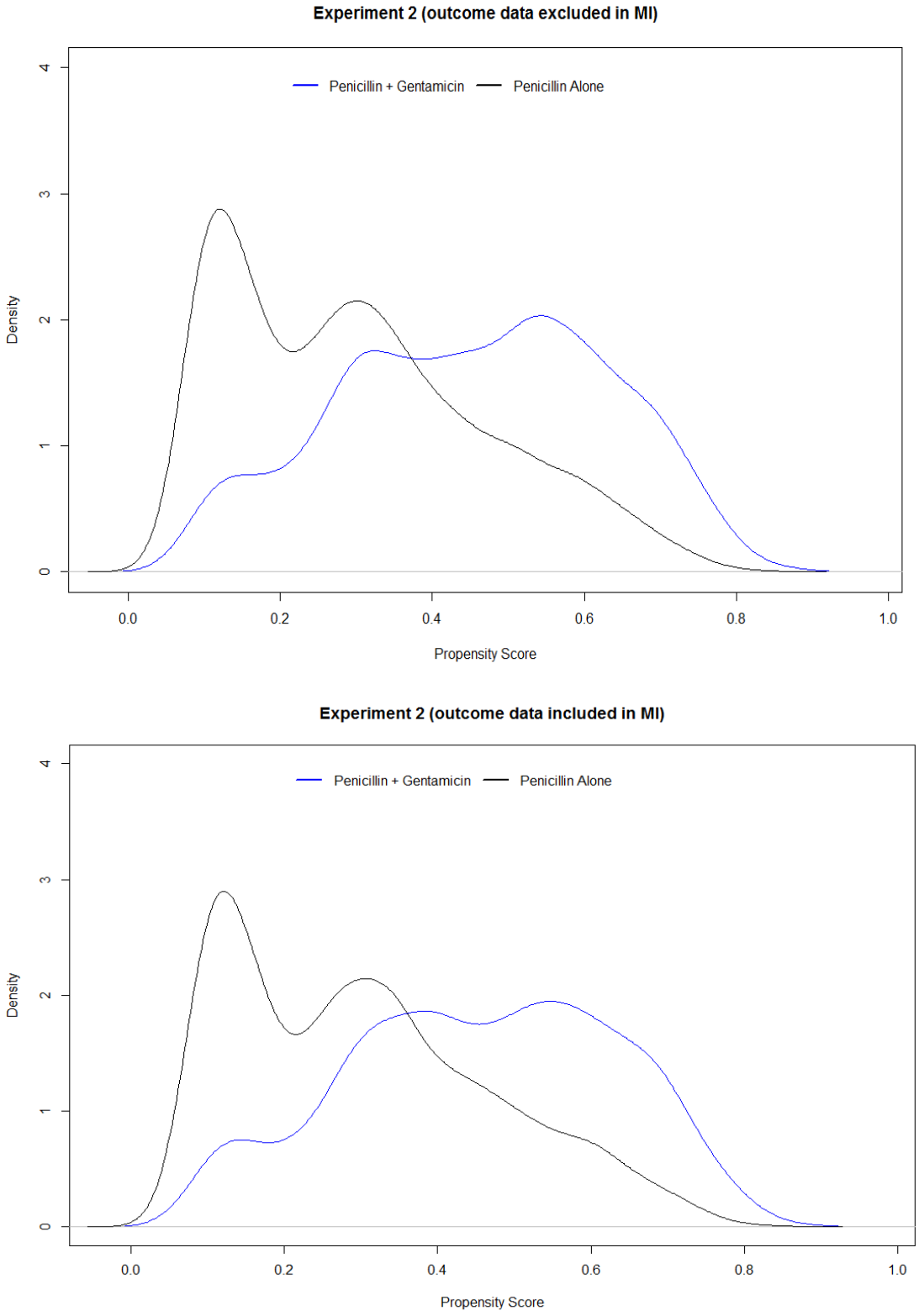


Figure 4.2.2: Distribution of propensity scores for penicillin montherapy and penicillin plus gentamicin treatment groups in experiment 2.

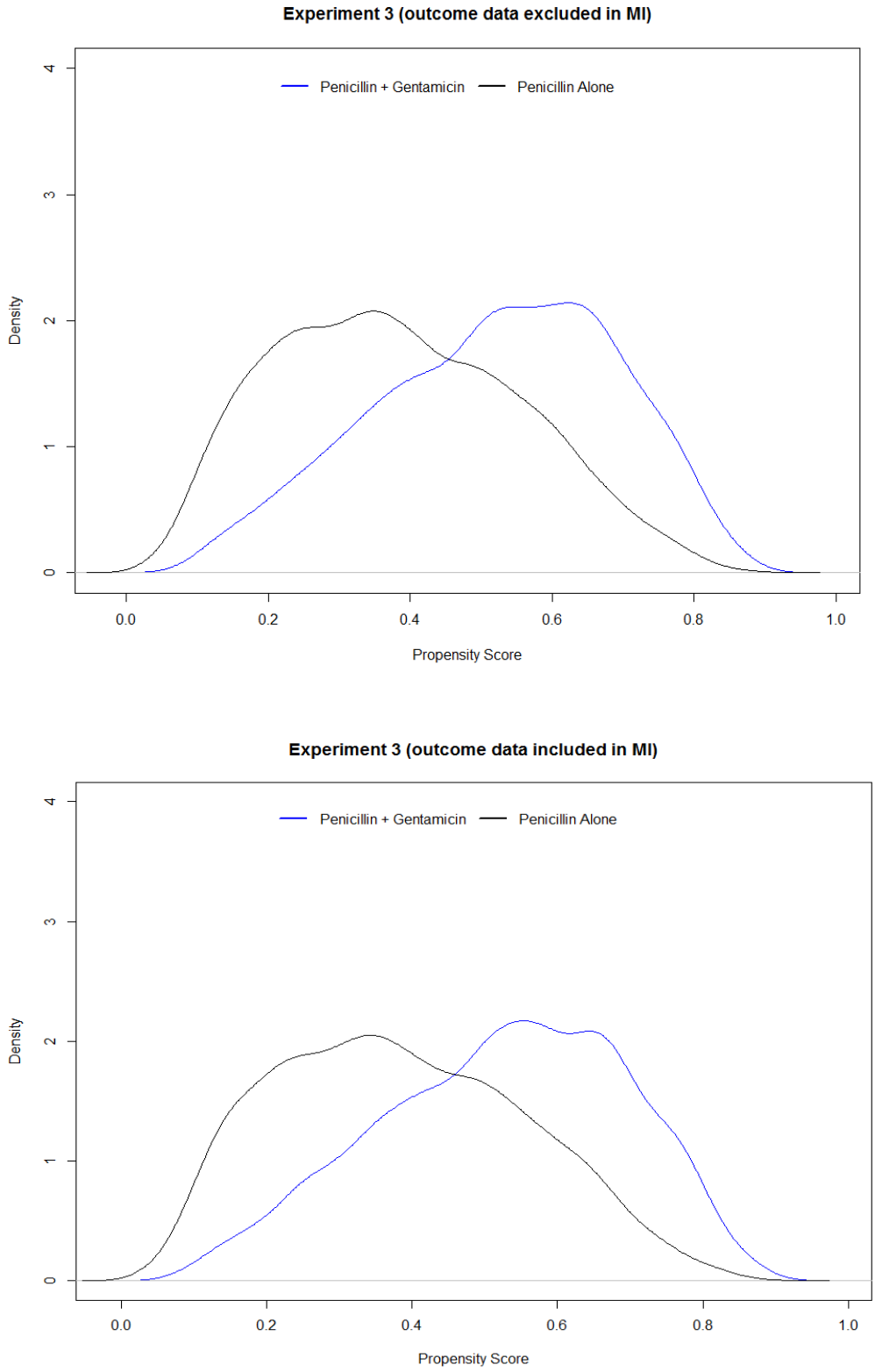


Figure 4.2.3: Distribution of propensity scores for penicillin montherapy and penicillin plus gentamicin treatment groups in experiment 3.

I therefore conducted new analyses, for experiments 1 – 3, for patients using varying conditions to exclude elements of the population (trimming) outside overlapping regions. Specifically, I conduct analyses for populations within the following PS distribution overlapping limits: 0 – 100; 1 – 99; 2.5 – 97.5 and 5 – 95, the latter being the most restrictive. Also I present the unrestricted[22] analysis – use of which is equivalent to the primary results presented in Chapter 3. The narrower the overlapping regions, the more homogenous (in characteristics) the patients become (though the smaller the sample size).

This procedure results in the exclusion of more mortality events in the penicillin plus gentamicin group compared with the penicillin monotherapy group across the trimming percentiles from the 0 – 100 to 5 – 95% (demonstrated using the MIW scenario and results presented in **Table 4.2.1**). This also results in a reduction in the apparent differences in penicillin versus penicillin plus gentamicin treatment groups suggesting less biased estimates (in experiments 1 – 3 as demonstrated in **Tables 4.2.2 – 4.2.4** below) and consequently possibly suggesting residual bias in the earlier analysis. However, it is important to note that the 95% credible intervals around treatment effect estimates in experiments 1 and 2 continue to suggest no difference in clinical effectiveness. Here, I only present the 5 – 95% trimming results in **Tables 4.2.2 – 4.2.4** as the form of sensitivity analysis most likely to reveal possible bias with other results on all the trimming percentile ranges presented in **Appendices D.4 – D.8**.

[22] Unrestricted analysis covers up to areas outside overlapping regions

Table 4.2.1: Proportion of deaths in the trimmed datasets for experiments 1 – 3 in the MIW scenario

	Experiment 1		Experiment 2		Experiment 3	
	Penicillin plus gentamicin	Penicillin alone	Penicillin plus gentamicin	Penicillin alone	Penicillin plus gentamicin	Penicillin alone
Unrestricted	33/1363 (2.42%)	26/2639 (0.99%)	87/2296 (3.79%)	50/4124 (1.21%)	264/4701 (5.62%)	82/5864 (1.40%)
0 – 100	33/1362 (2.42%)	26/2635 (0.99%)	85/2292 (3.78%)	50/4120 (1.21%)	257/4695 (5.47%)	76/5861 (1.30%)
1 – 99	30/1336 (2.25%)	26/2586 (1.01%)	80/2250 (3.41%)	48/4041 (1.19%)	242/4606 (5.25%)	71/5746 (1.24%)
2.5 – 97.5	28/1294 (2.16%)	24/2507 (0.96%)	71/2181 (3.22%)	46/3917 (1.17%)	223/4465 (4.99%)	68/5571 (1.22%)
5 – 95	26/1201 (2.16%)	24/2382 (1.01%)	62/2026 (3.06%)	46/3752 (1.22%)	202/4231 (4.77%)	65/5278 (1.23%)

* The unrestricted analyses represent estimates presented in Chapter 3 and there are more mortality events excluded in penicillin plus gentamicin group compared to penicillin monotherapy as the restriction is increased to 5 – 95% for all the experimental scenarios.

Table 4.2.2: Experiment 1: Treatment effect estimates under different approaches

	Optimal full matching		PS weighting		PS sub - classification	
	Unrestricted	5 – 95	Unrestricted	5 – 95	Unrestricted	5 – 95
MIA	0.83 [0.50, 1.32]	0.70 [0.43, 1.14]	1.18 [0.61, 2.26]	1.06 [0.56, 1.99]	1.31 [0.70, 2.44]	1.24 [0.69, 2.23]
MIAY	0.97 [0.58, 1.62]	0.72 [0.45, 1.16]	1.07 [0.58, 1.99]	0.96 [0.54, 1.71]	1.38 [0.73, 2.61]	1.33 [0.71, 2.49]
MIW	1.28 [0.79, 2.07]	1.22 [0.98, 1.50]	1.46 [0.85, 2.44]	1.39 [0.90, 2.15]	1.46 [0.82, 2.61]	1.34 [0.77, 2.32]
MIWY	1.05 [0.65, 1.71]	1.10 [0.71, 1.67]	1.23 [0.69, 2.20]	1.12 [0.65, 1.93]	1.31 [0.72, 2.38]	1.30 [0.75, 2.25]

* The restriction result in reduced treatment effect estimates closer to the null for all the PS methods used. Results on all the restrictions are presented in the appendix.

Table 4.2.3: Experiment 2: Treatment effect estimates under different approaches

	Optimal full matching		**PS weighting**		**PS sub - classification**	
	Unrestricted	5 – 95	Unrestricted	5 – 95	Unrestricted	5 – 95
MIA	1.12 [0.89, 1.43]	1.06 [0.92, 1.22]	1.02 [0.78, 1.38]	1.01 [0.86, 1.19]	1.49 [0.89, 2.05]	1.52 [0.91, 2.25]
MIAY	1.11 [0.87, 1.36]	1.05 [0.90, 1.23]	0.99 [0.72, 1.27]	1.00 [0.79, 1.26]	1.36 [0.89, 2.08]	1.42 [0.98, 2.16]
MIW	0.92 [0.69, 1.20]	0.82 [0.67, 1.00]	1.04 [0.76, 1.40]	1.05 [0.84, 1.31]	1.62 [0.99, 2.53]	1.46 [0.96, 2.22]
MIWY	0.88 [0.66, 1.13]	0.85 [0.70, 1.03]	1.01 [0.70, 1.26]	1.00 [0.76, 1.31]	1.63 [0.98, 2.51]	1.49 [0.95, 2.39]

Table 4.2.4: Experiment 3: Pooled treatment effect estimates under different approaches

	Unrestricted	5 – 95
MIA	1.60 [1.25, 2.05]	1.49 [1.27, 1.75]
MIAY	1.57 [1.20, 1.92]	1.46 [1.22, 1.74]
MIW	1.51 [0.77, 1.92]	1.40 [0.75, 2.63]
MIWY	1.63 [0.98, 2.51]	1.45 [0.94, 2.23]

4.2.2. Exploring residual confounding through instrumental variable analysis

An IV method aims to find a proxy randomised experiment in a routine or observational dataset (75). A valid IV should satisfy the following three conditions: (i) it should be usable as a variable for randomly and effectively assigning patients into alternative groups (and this is to ensure that the IV is not influenced by any unobserved variables so it helps mimic the case of a randomised controlled trial); (ii) relevance – the choice of IV should be logical and have a direct or confirmed effect on treatment received and; (iii) it should not be directly associated with the outcome except through the treatment (75). According to Baiocchi (2014), assumption one may partly be verified by examining covariate distribution across the levels of an IV variable and assumption two could be examined using likelihood ratio tests. And according to Klungel (2015), the third assumption may not be directly verifiable (122) but could be theoretically justified – though I used mediation models to examine this association (123).

4.2.2.1. Selection and use of IV in the CIN dataset

A few IVs used in health studies have been described in Baiocchi (2014) (75). These include: distance to specialty health service, genes, insurance plan, timing of admission, calendar time and preference based IVs. Of relevance to this

analysis would be timing of admission IVs. Multiple studies including Berkley (2004), Bell (2001), Freemantle (2015), Meacock (2016) and Aldridge (2016) have demonstrated that patients who were admitted during the weekend experienced higher mortality compared to those admitted during the weekdays (124-128) – which may be an indication of poorer quality of care and treatment during the weekend. In other words, it is anticipated that children admitted during the weekdays would have better health outcomes. This, in theory, implies that the type of treatment and care received depend on the day of admission – and which later determines the type of health outcome of the patient. Recent papers for United Kingdom (a high income setting) (129) dispute the weekend effect and indicate that all effects seen are because of improper matching. However, given the generally poor quality of care in LMIC and reasons to suggest outcomes are worse at weekend – timing of admission might still be a useful instrumental variable in lower and middle income countries. In order to assess whether a timing of admission variable forms a natural and random experiment the distributions of treatment and other covariates are examined across the levels of the instrumental variable (weekend/weekday) in experiments 1 and 2. The distribution of patient characteristics between weekend and weekday admissions is approximately similar for analysis populations (**Table 4.2.5**). This partially demonstrates its validity as an independent variable that would fairly assign patients into alternative treatment groups. Also all the absolute standardised mean differences (ASMD) of covariates are less than or equal to 0.1 which indicates that the imbalance in the analysis would be tolerable and not expected to grossly bias treatment effects (based on observed variables). A likelihood ratio

test demonstrated that there was a significant (p – value < 0.0001) association between treatment allocation and the admission timing instrumental variable in both experiments 1 and 2. Further, a mediation analysis showed a significant (experiment 1: p – value = 0.034, experiment 2: p – value < 0.0001) mediating effect of treatment between the outcome (mortality) and the timing to admission instrumental variable.

Table 4.2.5: Imbalance of covariates between weekday and weekend admissions

Variable	Experiment 1			Experiment 2		
	Weekdays (n = 3195)	Weekends (n = 988)	ASMD	Weekdays (n = 4946)	Weekends (n = 1539)	ASMD
Child Sex						
Female	45%	46%	0.03	44%	45%	0.01
Male	55%	54%		56%	55%	
Pallor						
Mild/moderate	4%	5%	0.02	5%	5%	0.00
None	95%	94%		93%	93%	
Severe	1%	2%		2%	2%	
Capillary refill						
1 sec	68%	71%	0.07	66%	68%	0.04
2 sec	30%	27%		31%	29%	
>2 sec	3%	2%		3%	3%	
Fever						
Absent	21%	18%	0.05	19%	16%	0.07
Present	79%	82%		81%	84%	
Convulsions						
Absent	95%	96%	0.02	94%	94%	0.03
Present	5%	4%		6%	6%	
Vomiting						
No	65%	62%	0.06	63%	63%	0.00
Yes	35%	38%		37%	37%	
Referral						
No	82%	86%	0.10	81%	84%	0.09
Yes	18%	14%		19%	16%	
Thrush						
Absent	98%	98%	0.00	98%	98%	0.03
Present	2%	2%		2%	2%	
Comorbidities						
None	84%	83%	0.02	82%	80%	0.03
Malaria	9%	10%		10%	13%	
Diarrhoea	3%	2%		3%	2%	
Malaria and diarrhoea	4%	5%		5%	5%	
Crackles						
Absent	47%	47%	0.01	48%	47%	0.02
Present	53%	53%		52%	53%	
Wheeze						
Absent	85%	84%	0.02	85%	84%	0.02
Present	15%	16%		15%	16%	
IV prescription						
No	97%	96%	0.05	95%	95%	0.01

Yes	3%	4%		5%	5%	
Quinine Prescription						
No	97%	97%	0.02	95%	94%	0.04
Yes	3%	3%		5%	6%	
Artesunate Prescription						
No	92%	92%	0.01	92%	90%	0.05
Yes	8%	8%		8%	10%	
Mean WAZ	0.00	-0.01	0.01	0.01	-0.03	0.03
Mean age (months)	19.59	20.47	0.04	20.29	21.05	0.04
Mean weight (Kg)	9.56	9.61	0.01	9.7	9.89	0.05
Mean resp rate (breaths/min)	52.61	51.65	0.08	51.82	51.34	0.04
Mean temp (degrees C)	37.73	37.79	0.06	37.78	37.85	0.06
Mean cough duration (days)	3.40	3.20	0.07	3.45	3.35	0.04
Mean length of illness (days)	3.70	3.46	0.08	3.73	3.56	0.05

The suitability of admission timing (weekend/weekday) as an instrument is also demonstrated in **Table 4.2.6** – where the proportion of weekend deaths was higher than proportion of deaths on weekday admissions.

Table 4.2.6: Summary of deaths by weekend/weekday admissions

Experiment	**Weekend**	**Weekday**
1	17/988 (1.7%)	45/3195 (1.4%)
2	47/1539 (3.1%)	49/4946 (1.0%)

The next sub – sections examines how to estimate average treatment effect (ATE) and average treatment effect for the treated (ATT) using instrumental variable analysis.

4.2.2.1. Estimating average treatment effect (ATE)[23] using an instrumental variable

Let Y_i denote an outcome variable (i = 1, …, n), T_i – a treatment variable, Z_i – an instrumental variable, X_i – an independent variable, and ϵ_i – an error term associated with the outcome (n being the number of observations). Essentially

[23] There are two possible effects that could be estimated: (a) average treatment effect (ATE) which would be estimated as $E(Y_{i1}|Treatment = A) - E(Y_{i0}|Treatment = B)$ and; (b) average treatment effect for the treated (ATT) estimated as $E(Y_{i1}|Treatment = A) - E(Y_{i0}|Treatment = A)$.

when using an IV, two models are specified separately: (i) a treatment model which includes the IV together with all other confounders as covariates, $T_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 X_i$ (similar to PS model specification though a PS model may exclude IVs), (ii) an outcome model, $Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \epsilon_i$ (which excludes the IV). The traditional way of fitting IV models has been using two stage ordinary least squares (122). One limitation of this approach is that the regression model used in stage one may predict probabilities outside the 0 – 1 range (122). Klungel (2015) discussed various methods to overcome this limitation. These methods include use of two stage logit models and a recursive bivariate probit model among others. This chapter considers the use of recursive[24] bivariate probit (RBP) models for estimation of ATE – as they model probabilities (between 0 and 1) and takes into account the correlation of error terms between the treatment and outcome models. Using earlier notations, the RBP model is therefore written as:

$$T_i^* = \alpha_0 + \alpha_1 Z_i + \alpha_2 X_i + \mathcal{E}_i$$

$$Y_i^* = \beta_0 + \beta_1 T_i + \beta_2 X_i + \epsilon_i$$

Where $T_i^* \; and \; Y_i^*$ are continuous latent variables – and in this case would represent the probability of receiving treatment and dying respectively; $\mathcal{E}_i$ is the error term associated with treatment. The error terms $\mathcal{E}_i \; and \; \epsilon_i$ have a bivariate normal distribution with zero means, unit variances and are independent and identically distributed (131), that is:

[24] The recursive nature of this model allows for the estimation of treatment effects in the presence of unobserved variables (130).

$$\begin{pmatrix}\varepsilon_i \\ \epsilon_i\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \ell \\ \ell & 1\end{bmatrix}\right)$$

ℓ – represents correlation between the two error terms. ℓ is useful in examining the effects of unmeasured variables and as a proof of assumption one – such that if $\ell = 0$ or is close to zero then it may be an indication of no influence of unobserved variable(s) (131).

In the next step, the treatment and outcome (mortality) probit models are fitted, with covariates in the treatment model being the same as those used in the corresponding propensity score models (see **Table 3.1.1** of the protocol in Chapter 3 for the list of variables used) – though with the addition of admission timing variable as an IV. On the other hand, the outcome model uses the same covariates as the treatment model with the exclusion of the admission timing variable both in experiments 1 and 2. Here, the parameter estimates are only presented for the treatment variable. Also the correlation coefficients between the error terms are presented alongside the treatment effect estimates.

Interpreting individual coefficients (like for treatment here) is less straightforward in probit models compared to linear regression and logit models where estimates are individually interpretable (132). This is because change in probability due to a unit change in a predictor is jointly dependent on other predictor values and their starting values. However, there are limited ways through which probit model parameters may be interpreted individually: (i) without considering the magnitude, the direction of effect may be inferred based on whether the parameter estimate is either positive or negative; (ii) if both the

magnitude and direction are of interest (as is the case here), then a set of approximations may be conducted. Amemiya (1981) suggested multiplying the individual estimate from a probit model by 1.6 to obtain the result in terms of a log odds ratio (133). As the estimates obtained using PS methods were expressed in terms of risk ratio, the estimated odds ratios are further converted to risk ratios using the modified relationship documented in (134):

$$RR = \frac{OR}{(1 - p_0) + (p_0 \times OR)}$$

Where RR – is the risk ratio; OR – odds ratio and; p_0 – is the proportion of children who died in the penicillin monotherapy treatment group.

After describing the standard methodology for estimating ATE with an instrumental variable, I now present ATE results in **Table 4.2.7**. The estimated treatment effects in experiments 1 and 2 indicate no difference in effects between penicillin and penicillin plus gentamicin. However, the effect sizes obtained are reversed (though not significant) and lower than the effects obtained using PS methods in Chapter 3. Also the obtained estimates for correlation between errors are moderately lower (< .30) possibly indicating less influence from unobserved variables.

Table 4.2.7: Average treatment effect estimates

Experiment	Outcome data excluded in MI		Outcome data included in MI	
	RR (C.I)	ℓ (p – value)	RR (C.I)	ℓ (p – value)
1	0.91 [0.41, 2.20]	0.13 (0.061)	1.02 [0.46, 2.15]	0.04 (0.092)
2	0.44 [0.34, 1.32]	0.28 (0.057)	0.70 [0.25, 2.01]	0.21 (0.052)

***** The reference category in this analysis was penicillin monotherapy and therefore RR < 1 would favour penicillin plus gentamicin.

* Even though the estimates for the correlation parameter (ℓ) in both experiments are not statistically significant, the magnitudes are greater than zero indicating there might be some influence due to unobserved variables.

4.2.2.2. Estimating average treatment effect for the treated (ATT) using an instrumental variable

In order to estimate the ATT (that is, the average treatment effect for being treated with penicillin plus gentamicin), a two stage modelling approach is adopted. In stage 1, a treatment logit model is fitted (including the IV as a covariate). The probabilities (P_i) for belonging to either gentamicin plus penicillin or penicillin alone are then obtained. In stage 2, an outcome model (log binomial regression) is specified with the treatment indicator probabilities obtained in stage 1 used as covariates. ATT results are presented in **Table 4.2.8**.

Table 4.2.8: Average treatment effect estimates for the treated

Experiment	Outcome data excluded in MI: RR (C.I)	Outcome data included in MI: RR (C.I)
1	0.96 [0.37, 2.31]	0.98 [0.53, 2.22]
2	0.49 [0.19, 1.29]	0.66 [0.25, 1.78]

***** The reference category in this analysis was penicillin monotherapy and therefore RR < 1 would favour penicillin plus gentamicin.

The IV analyses for ATT, both in experiments 1 and 2, based on the 95% C.I also seem to indicate that there is no significant difference in treating indrawing pneumonia using either penicillin monotherapy or a combination of penicillin and gentamicin.

In both experiments and in estimation of ATT and ATE, Wu-Hausman tests (135) were conducted to examine if IV models were fitting better than models fitted on PS weighted datasets and this showed no difference in all the cases. However, it is important to note that the models fitted on the PS adjusted datasets were considered primary as they clearly separated design and outcome analysis phases.

4.2.2.3. Instrumental variable analyses within the five PS subclasses in experiment 3

As opposed to experiments 1 and 2 that examined effectiveness of antibiotic treatments for children only with indrawing pneumonia, experiment 3 included all forms of pneumonia admissions (experiment 1 and 2 populations were a subset of experiment 3). Analysis in Chapter 3 therefore used PS subclasses to denote pneumonia severity. The use of subclasses stratified patients such that those who had lower propensity scores were considered less sick, while those who had higher propensity scores were considered sicker. As demonstrated in **Table 4.1.6** (highlighted results for the MIW scenario are those also presented in Chapter 3), the estimates in subclasses 1 – 3 and even the pooled effect estimate showed a protective effect of penicillin monotherapy indicating the likely presence of residual bias. Here, I examine the performance of combining PS sub-classification with IV analyses within each subclass to help minimise residual bias. This I illustrate using the same MIW scenario which had been presented as the primary analysis. First, variables are multiply imputed (using 20 datasets). Then propensity scores are estimated, patients are classified into 5 PS subclasses, and IV analysis (using the method described above for estimating ATT) is conducted on each dataset. The results are then meta – analytically pooled across the sub – classes within each imputed dataset. Finally, treatment effect results are pooled across the imputed datasets using Rubin's rules (59). The results for these IV analyses for experiment 3 in the MIW scenario are presented in **Table 4.2.9**. The treatment effects in subclasses 1 – 4 show no statistical significance in outcome after treatment with either gentamicin plus

penicillin or penicillin monotherapy. However, gentamicin plus penicillin seems to be more effective in treating children in subclass 5 (the sickest children) than penicillin alone. Overall, we see, as in Chapter 3, a transition from a tendency to better outcomes with penicillin monotherapy towards a benefit of treatment with penicillin plus gentamicin as we move from subclass 1 to 5 (from less severe to more severe pneumonia). The use of IV analysis within the PS subclasses therefore seems to have reduced bias in this case as the pooled estimate derived from IV analyses suggests no significant difference in outcomes in the two treatment groups.

Table 4.2.9: Treatment effect estimates in experiment 3 (IV analysis combined with PS sub - classification)

Subclass	PS sub – classification alone	PS sub – classification with IV analysis
1	2.71 [1.56, 4.71]	2.11 [0.96, 4.62]
2	2.58 [1.34, 4.95]	2.02 [0.84, 4.88]
3	2.07 [1.19, 3.60]	0.82 [0.50, 1.34]
4	1.52 [0.84, 2.75]	1.03 [0.63, 1.68]
5	0.98 [0.60, 1.59]	0.87 [0.77, 0.98]
Pooled	1.75 [1.36, 2.27]	1.05 [0.78, 1.39]

***** The highlighted estimates are derived from the models used in the primary analyses in Chapter 3

4.2.3. Discussion

In this chapter, I have explored the potential for bias/residual (with bias quantified using the formula defined in sub – section 4.1.1 above) confounding as an explanation for the treatment effect estimates reported in Chapter 3. I have done this through sensitivity analyses that: (a) varies the approach to using multiple imputation with PS methods; (b) restricts analysis to populations with more closely overlapping characteristics (PS trimming) and; (c) use instrumental variables. The findings are now discussed.

The three propensity score methods seem to perform slightly differently when estimated after multiple imputation conducted with inclusion and exclusion of outcome data in the chained equations. Optimal full matching and PS sub – classification seem to eliminate the largest amount of bias when used with the across method for estimating PS with outcome data excluded in multiple imputation – a result which is consistent (only for PS matching) with that of Mitra (2016) (81). PS weighting seems to perform better in reducing bias when estimated using the within method for PS estimation with outcome data included in the multiple imputation chained equations – a finding which is also consistent with the results of Leyrat (2017) (116).

According to definitions by Rosenbaum and Rubin (1983) (47), PS should be a balancing score in the sense that it should satisfy the following three assumptions:

- Positivity assumption: each patient should have a non – zero probability of being assigned into either or any of the treatments being compared.
- Stable unit treatment value assumption (SUTVA): the potential outcomes for a patient are not influenced by treatment received by other patients.
- Strongly ignorable treatment assignment (SITA): there are no unmeasured confounders that would influence final treatment effect estimates.

Under the positivity assumption, according to Leyrat (2017), the derived PS should be a direct function of the pre – treatment variables. Any score which is not a direct function of the covariates may be considered less "finer" (116). And according to the definition of Rosenbaum and Rubin (1983), less "finer" scores

may be considered as non – balancing scores (47). The across method of PS estimation averages propensity scores across datasets and thus is not a direct function of the observed variables. This would yield less 'finer' scores – making the averaged PS less of a balancing score (47). Further, Leyrat (2017) demonstrated that the across method of deriving PS was not a consistent estimator of the true PS (see Chapter 6 for formal proof). That is, averaging PSs across imputed datasets would not yield the true PS. This theoretically makes the within method of PS estimation a more desirable approach than the across method. Considering the within method as used in this analysis, the inclusion of outcome data in PS optimal full matching and weighting minimised bias better as compared to exclusion of outcome data. The performance in bias reduction was approximately equivalent in PS sub-classification when outcome data were excluded and included in imputation models (with PS also estimated using the within method). Based on this reason, inclusion of outcome data in multiple imputation with the within PS estimation (in PS optimal full matching, weighting and sub – classification) would potentially give results with both theoretically and practically desirable properties.

The PS density plots (in experiments 1 – 3) demonstrated the potential existence of unobserved confounders in terms of treatment assignment in which those who had higher PS scores (linked to higher prevalence of 'positive' clinical findings) were more often assigned broader spectrum antibiotics. This group of patients also had a higher risk of dying. In PS trimming, elimination of more patients using defined percentiles excluded some of these higher risk patients and resulted in more homogeneous patient populations with treatment effect

estimates moving towards the null. If we assume that gentamicin plus penicillin would be at least equivalent to penicillin, these findings suggest possible residual bias. Further, instrumental variable analyses pointed to similar findings on treatment effects observed in experiments 1 and 2 (with amelioration of a tendency to suggest more favourable outcomes on penicillin alone) though with direction reversed as compared to results obtained in Chapter 3. The use of PS sub – classification and instrumental variable may be considered doubly robust in minimising allocation bias. In experiment 3 in Chapter 3, using PS sub – classification alone suggested benefit of penicillin – which is counter intuitive based on the biology of treatments. However, when PS sub – classification is used together with instrumental variable analysis, this tendency toward favourable outcomes on penicillin is eliminated.

Experiments 1 and 2 (Chapter 3) results showed no significant differences in treatment outcomes and the sensitivity analyses (which may further reduce bias) confirm this finding of no appreciable difference. However, in experiment 3 the sensitivity analyses, especially the use of PS sub – classification and instrumental variable analysis, appear to perform better at overcoming bias suggesting the results in Chapter 3 experiment 3 results should be interpreted with caution.

The method used to quantify bias (defined in sub – section 4.1.1) relies only on observed variables, but this analysis did not try to quantify potential bias due to unobserved variables. Thus the estimated bias in this analysis quantify differences rather than actual bias.

Chapter 5

Examining the effectiveness of Zinc treatment in children admitted with diarrhoea in Kenya's public hospitals: An observational study

---

I outlined the methodological basis and rationale for this work and examined the current evidence for effectiveness of Zinc treatment in detail in Chapter 1 (sub – section 1.1.4.2). In Chapter 3, I presented a detailed protocol that guided pneumonia analysis. I also developed a similar protocol to guide the analyses of Zinc effectiveness I now report. This detailed protocol is provided in **Appendix E.1**. I begin this chapter with a summary of the issue being tackled and the methods used then present the results and discussion.

Chapter summary

Introduction: Kenyan paediatric treatment protocols (adopted from WHO guidelines) recommend the use of Zinc supplement for all children with diarrhoea. However, there is limited evidence of benefit for young children aged 1 – 5 months and those who are well-nourished. I examine effectiveness of Zinc supplementation for children admitted with diarrhoea to Kenya's public hospitals with different nutritional and age categories. This is to determine whether there is support for the current policy where Zinc is prescribed for all children with diarrhoea.

Methods: This analysis used the CIN dataset to explore the effect of Zinc prescription on time to discharge for children aged 1 – 5 and 6 – 59 months and amongst those classified as either severely – moderately malnourished or well-nourished. To overcome the challenges associated with non – random allocation of treatments and missing data in these observational data, I used propensity score (PS) methods and multiple imputation to minimize bias. A flexible competing risk regression model was used to examine possible benefits of Zinc; and to examine the possible influence of unobserved confounders, an instrumental variable analysis was performed.

Findings: The analysis included 1 645 and 11 546 children aged 1 – 5 months and 6 – 59 months respectively. The estimated sub-distribution hazard ratios for being discharged[25] were 1.25 [1.07, 1.46] and 1.17 [1.10, 1.24] in these respective age categories. The pooled effectiveness of Zinc in reducing time to discharge across the two age categories was 1.17 [1.11, 1.24]. This implies that a child receiving Zinc has on average a 17% higher chance of being discharged at any point in time than a child not receiving Zinc. The association with shorter time to discharge was seen in well-nourished (WAZ >-2SD) and under-nourished children (WAZ <= - 2 SD).

Conclusion and relevance: Zinc is associated with shorter time to discharge. In the absence of significant adverse effects, data support the continued use of Zinc for admissions with diarrhoea including those aged 1 – 5 months and in those who are well-nourished, groups for whom evidence of the benefit of Zinc has to date been weak.

[25] Time for discharge measured in days

5.1. Introduction

Diarrhoea is a major cause of morbidity and mortality in lower and middle income countries (LMIC) with the primary treatment being rehydration regimens matched to the severity of the dehydration (7, 27). Co-treatment with Zinc supplements is also recommended (24, 27). This is based on results of randomised controlled trials and subsequent systematic reviews (25, 136-139) suggesting that Zinc supplements reduce diarrhoea duration for children aged six months and above. Kenyan guidelines in keeping with those of the World Health Organisation recommend treatment of diarrhoea with fluids and co-treatment with oral Zinc for 14 days for all children irrespective of illness severity (27).

Despite these recommendations, there remains some debate on the benefits of Zinc since:

i. The prevalence of Zinc deficiency varies by setting (140) and this may contribute to between country variations in Zinc effectiveness reported in a systematic review by Patel (2012) (26).

ii. Trials supporting the use of Zinc have included fewer participants from Africa than other low and middle income settings (25) and it is suggested that African children with diarrhoea have poor health outcomes (7).

iii. There are few data supporting the benefits of Zinc in children younger than six months and few data on the effects of Zinc in children of all ages who are relatively well-nourished (25).

Conducting trials to address all these questions would likely be expensive and time-consuming. Appropriate analyses of observational datasets may help address such questions while also providing data on the effectiveness of Zinc treatment in non-trial populations.

5.1.1. Objectives

The primary focus of this analysis was to examine the effectiveness of Zinc supplementation in reducing time (in days) to discharge for children (aged 1 – 5 and 6 – 59 months) admitted with diarrhoea in Kenyan hospitals. Systematic reviews have included data on mostly trials conducted in hospital settings and did not find an effect on mortality but do suggest a valuable effect on duration of diarrhoea (25). As CIN patient population was hospitalised, I therefore focused on length of stay (to model time to getting discharged alive) likely to be influenced by duration of symptoms. In secondary analyses conducted within each age group, I aimed to examine the effectiveness of Zinc amongst those classified as either severely-moderately malnourished or well nourished. Further, as part of exploratory analyses, time to experiencing inpatient mortality (in those who received Zinc versus those who did not) was examined in the two age categories.

5.2. Methods[26]

5.2.1. Data source

This analysis uses the Kenyan Clinical Information Network (CIN) dataset that provides observational data on all admissions to 13 Kenyan County (plus one sub – county) hospitals (see Chapter 1 sub – section 1.2.2 on CIN introduction).

5.2.2. Statistical analysis

5.2.2.1. Analysis populations

All children included in the analysis had diarrhoea on admission. I excluded children who were aged < 1 month and >= 60 months (as current policy recommends Zinc treatment for children aged 1 – 59 months), had shock or who had impaired consciousness (AVPU score P or U). I primarily examined the effect on time to discharge of whether children were prescribed Zinc in two groups: all children aged 1 – 5 months (group 1) and all children aged 6 – 59 months (group 2). Within each of these populations I examined the effects of Zinc in severely-moderately malnourished and well-nourished children. To define nutritional status, I derived weight for age z scores[27] using WHO reference population data (141). Children with z – scores < -2 were classified as severely-moderately malnourished, while those with z – scores >=-2 were considered well-nourished (141).

[26] Also see step by step analysis methods in analysis protocol presented in appendix E.1.

[27] Weight for Height Z (WHZ) scores and Mid Upper Arm Circumference (MUAC) are preferred to define Severe Acute Malnutrition (SAM) but data often missing in CIN dataset so I used Weight Age Z score (WAZ).

5.2.2.2. Handling of missing data

To maintain the effective sample size, imputation methods were used to account for missing data in the covariates used in the PS models – in which 10 datasets[28] (114) were derived using multiple imputation with chained equations (142). Outcome data were used for imputation in the chained equations.

5.2.2.3. Handling non-random allocation of Zinc supplement

As CIN comprises data from routine care settings Zinc treatment cannot be considered as allocated at random. In fact, not being prescribed Zinc is a result of failure to adhere to national guidelines which may not be random (see the discussion on treatment assignment approaches in Chapter 1 sub-section 1.2.5). Analyses must try and account for any potential bias in treatment allocation. To help create balance in population characteristics amongst those with and without Zinc prescription, I used propensity score (PS) methods. Propensity scores describe the probability of receiving Zinc based on measured characteristics of the population, in this case clinical signs, symptoms, co-treatments and comorbidities (62). Outcomes in children with similar propensity scores (between those who received and those who did not receive Zinc) can then be compared as one means to overcome allocation bias. In these analyses, I compared the ability of two PS approaches to reduce possible bias – optimal full matching and weighting (37, 62). Both are aimed at creating groups of patients that are comparable in terms of the distribution of observed signs and symptoms

[28] The current literature (114) recommends the use of more than 5 imputed datasets and therefore 10 should be sufficient. Amount of missing data and multiple diagnostics have been reported in appendix.

(though they may result in somewhat different groups being compared). In order to select the optimum PS implementation method, standardised mean differences were used as diagnostic checks for covariate balance and overlap (107, 108) between Zinc and non-Zinc groups. Even though both the PS methods would retain all patients in the analysis, the method that resulted in the minimum average absolute standardised mean differences for the majority of the variables was considered the most appropriate (37). For each of the two PS implementation methods, PSs were estimated using the within method (defined in Chapter 2 sub – section 2.1.2) after multiple imputation.

An additive interaction[29] was used in modelling the use of Zinc by nourishment status (143). As such, I derived a variable with four levels representing those who: (i) received Zinc and were well-nourished; (ii) received Zinc and were malnourished; (iii) did not receive Zinc and were well-nourished and; (iv) did not receive Zinc and were malnourished. The reference group used in the analyses, was well-nourished children who received Zinc as they were expected to be most likely to be discharged sooner than the other three subgroups. To ensure the four subgroups had comparable patient characteristics, I followed the step by step approach suggested by Spreeuwenberg (2010) (144). This involved: (i) Fitting PS models using multinomial logistic regression with a probit link function to obtain four propensity scores per patient (which should add to one); (ii) Examining PS distribution overlap. This is important as a patient in a treatment subgroup should have some probability of being assigned/belonging to other subgroups.

[29] An additive interaction is preferable when analysing treatment effectiveness as opposed to multiplicative interaction (143). This is to have an understanding of effectiveness in all the possible subgroups under consideration.

Biased estimates may be obtained in case of PS non-overlap; (iii) Examining covariate balance across the four subgroups. Covariables used in creating propensity scores in all the analyses included key signs and symptoms suggested for diagnosing and assessing severity of diarrhoea and dehydration in the Kenyan paediatric guidelines together with variables considered *a priori* to influence the clinical outcomes of interest such as fluid regimen prescribed and comorbidities. See **Appendix E.2.1** for full description.

5.2.2.4. Modelling of time to being discharged alive

The primary outcome considered in the analyses was time to being discharged alive. Mortality was treated as a competing risk as it would preclude the chance of a patient being discharged alive. To allow for covariates with varying effects across discharge time points, I used a flexible modelling approach suggested by Scheike (2011) (145). Both Kolmogorov-Smirnov and Cramer von Mises test statistics (145) were used to determine whether covariates had varying or constant effects across time points prior to discharge. Patients who absconded, were referred to other hospitals and those who had length of stay greater than 15 days (as Zinc should be administered for two weeks and should influence the outcome of acute diarrhoea) were censored (analyses without censoring were also attempted as part of the sensitivity analysis). For each of the analyses, propensity scores were estimated on each of the ten imputed datasets using multivariable logistic regression. Then multivariable Scheike's flexible regression models were fitted to datasets with and without PS adjustment and while adjusting for all the variables also used in the PS models for primary

analyses. However, in analysis of Zinc effectiveness by nourishment status, only propensity scores together with the four level variable (derived as explained above) were used as covariates in Scheike's model as advised in Spreeuwenberg (2010) (144). Only adjusted (subgroup) treatment effectiveness estimates were pooled across the ten imputed datasets to obtain a single estimate.

As secondary (exploratory) analyses, time to experiencing inpatient mortality was also modelled on the same PS adjusted datasets (considering being discharged alive as a competing risk in Scheike's regression model). This analysis was useful since previous systematic reviews had demonstrated no Zinc effect on mortality (25).

5.2.2.5. Sensitivity analysis using an instrumental variable

As the CIN data are limited to the recorded variables, I used an instrumental variable (weekday/ weekend admission – also discussed in Chapter 4) to examine the potential influence of any unmeasured variables in all the analysis (146). An instrumental variable method aims to find a natural experiment in a routine or observational dataset (75). As indicated in Chapter 4 sub – section 4.2.2, a valid IV should satisfy the following three conditions: (i) it should be usable as a variable for randomly and effectively assigning patients into alternative groups (and this is to ensure that the IV is not influenced by any unobserved variables so it helps mimic the case of a randomised controlled trial); (ii) relevance – as the choice of IV should be logical and have a direct effect on treatment received and; (iii) it should not be directly associated with the outcome but only through the treatment (75). According to Baiocchi (2014), assumption one may partly be

verified by examining covariate distribution across the levels of an IV variable and assumption two could be examined using likelihood ratio tests. And according to Klungel (2015), the third assumption may not be directly verifiable (122) but could be theoretically justified (though I used mediation models (123) to test this association).

To model using instrumental variables in a survival context, Tchetgen (2015) (147) suggested the use of a control function where modelling happens in two steps. Step 1 involves the estimation of residuals. The residuals are estimated from a model where treatment membership is modelled as a function of the IV and covariates influencing treatment assignment. Step 2 involves the use of these residuals as predictors in a second stage survival model – and in this case I used the Scheike's flexible competing risk model.

5.2.2.6. Sensitivity analysis on multiple imputation assumptions

The multiple imputation conducted in the primary analysis assumes data were missing at random (see Chapter 2 sub – section 2.1.2 on definitions of missing data mechanisms). For sensitivity analysis, patients were grouped based on the number of missing variables to potentially denote a missingness not at random situation (53) – see **Appendix E.1** for detailed description. In particular, also as formulated in Chapter 4, three patterns were used in which pattern 1 consisted of patients with missing data in 0 – 3 variables, pattern 2 consisted of patients with 4 variables with missing data, and pattern 3 consisted of patients with more than four variables with missing data. Thereafter, multiple imputation and

propensity score analyses were repeated for each of the patterns for groups 1 and 2. Zinc effectiveness estimates were pooled across the three patterns.

5.3. Results

5.3.1. Inclusion and exclusion

A total of 1 645 (aged 1 – 5 months) and 11 546 (aged 6 – 59 months) children were eligible for analysis in age groups 1 and 2 respectively. And calculations in **Appendix E.1 Figure 2** demonstrated that the sample sizes were sufficient for time to discharge analyses in these two age groups to detect a sub-distribution hazard ratio of 1.2. About 72% (1 – 5 months) and 77% (6 – 59 months) were treated with Zinc supplements (**Figure 5.0.1**) – and this varied little over the time period studied.

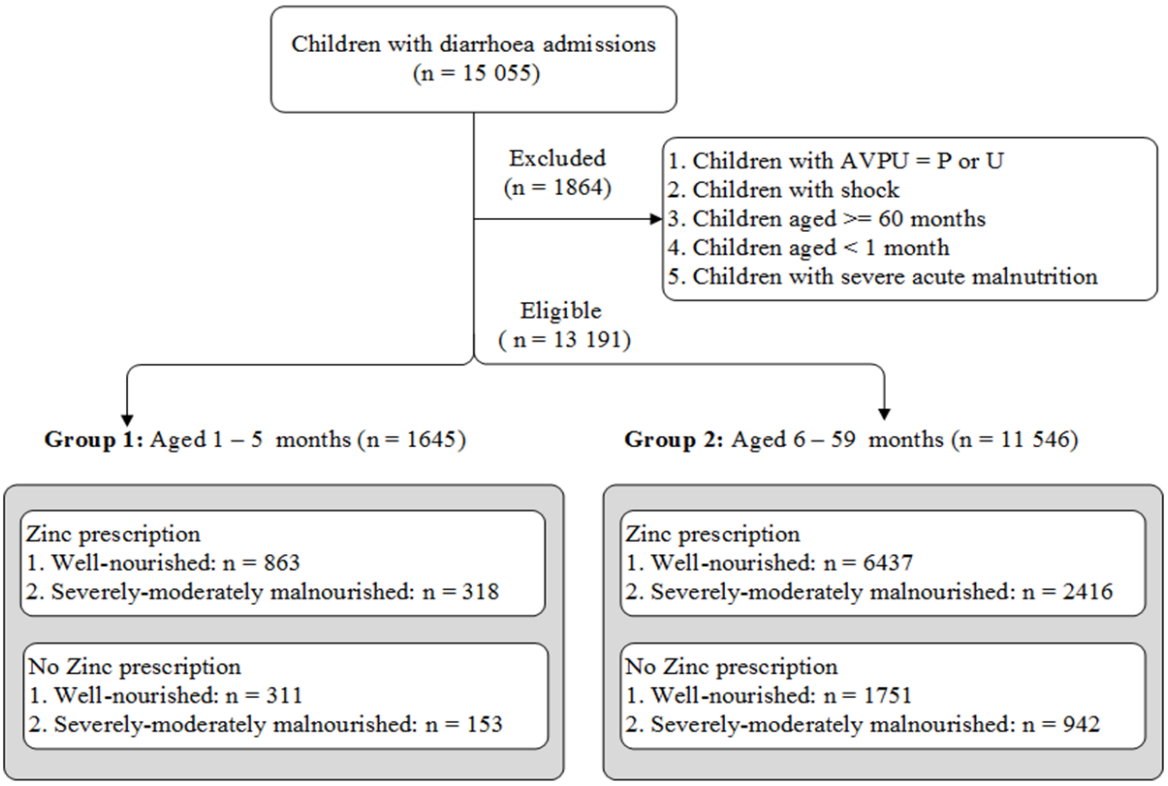


Figure 5.0.1: Eligibility

5.3.2. Comparing performance of PS optimal full matching and weighting in groups 1 (1 – 5 months) and 2(6 – 59 months) respectively

For group 1, the performance of optimal full matching was comparable to that of weighting, while weighting performed better than optimal full matching in group 2 (see **Figures 5.0.2 and 5.0.3**). Thus, in the subsequent sections, outcome analyses for groups 1 and 2 are based on PS weighted datasets for both age groups.

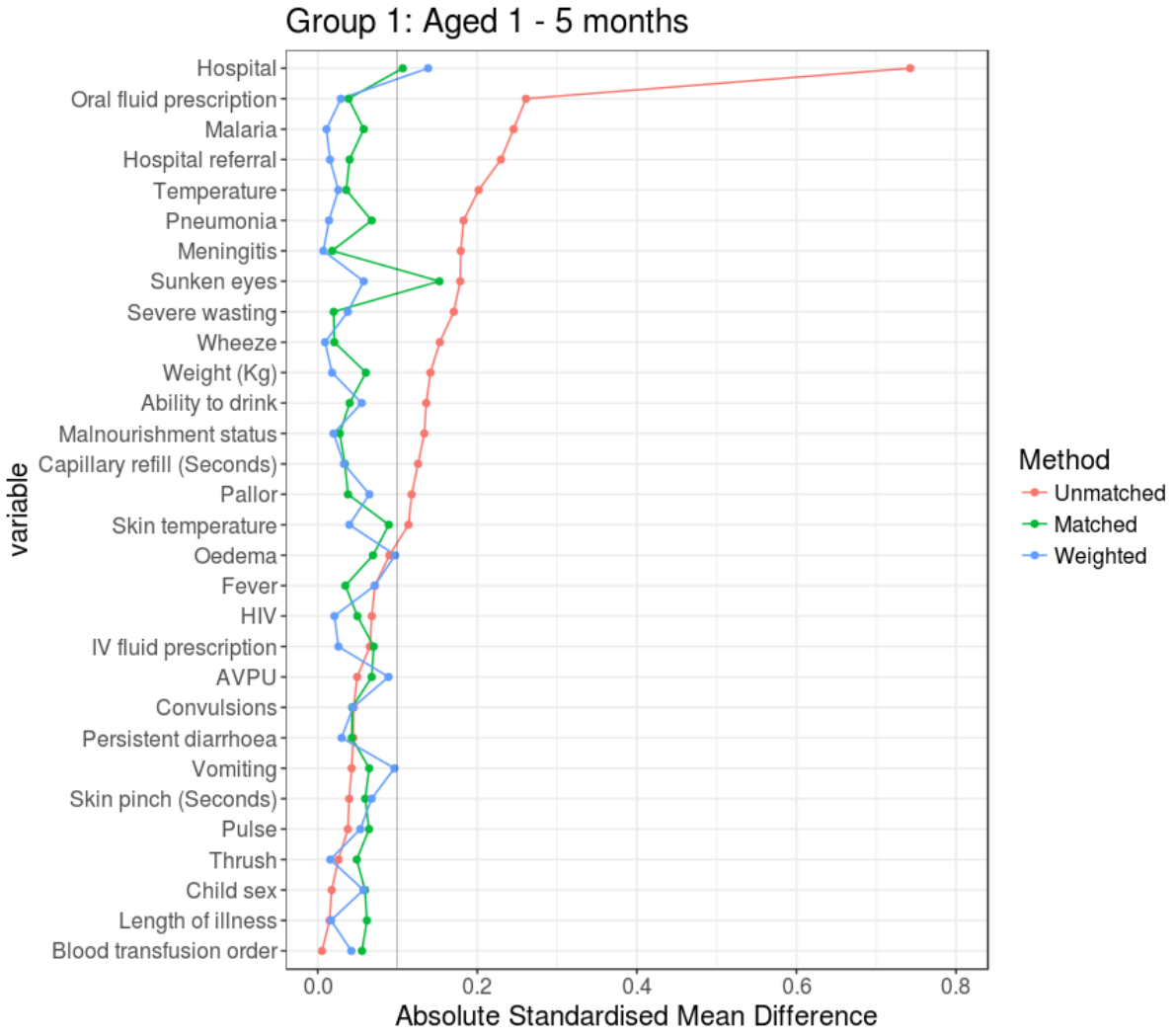


Figure 5.0.2: Comparing performance of the two PS implementation methods in group 1 (1 – 5 months). The y – axis contains all the variables used in the PS models. While x – axis shows absolute standardised mean difference (ASMD) which is a measure of covariate balance between Zinc and non-Zinc groups. An ASMD value of <= 0.1 indicates the method has performed well in creating comparable groups.

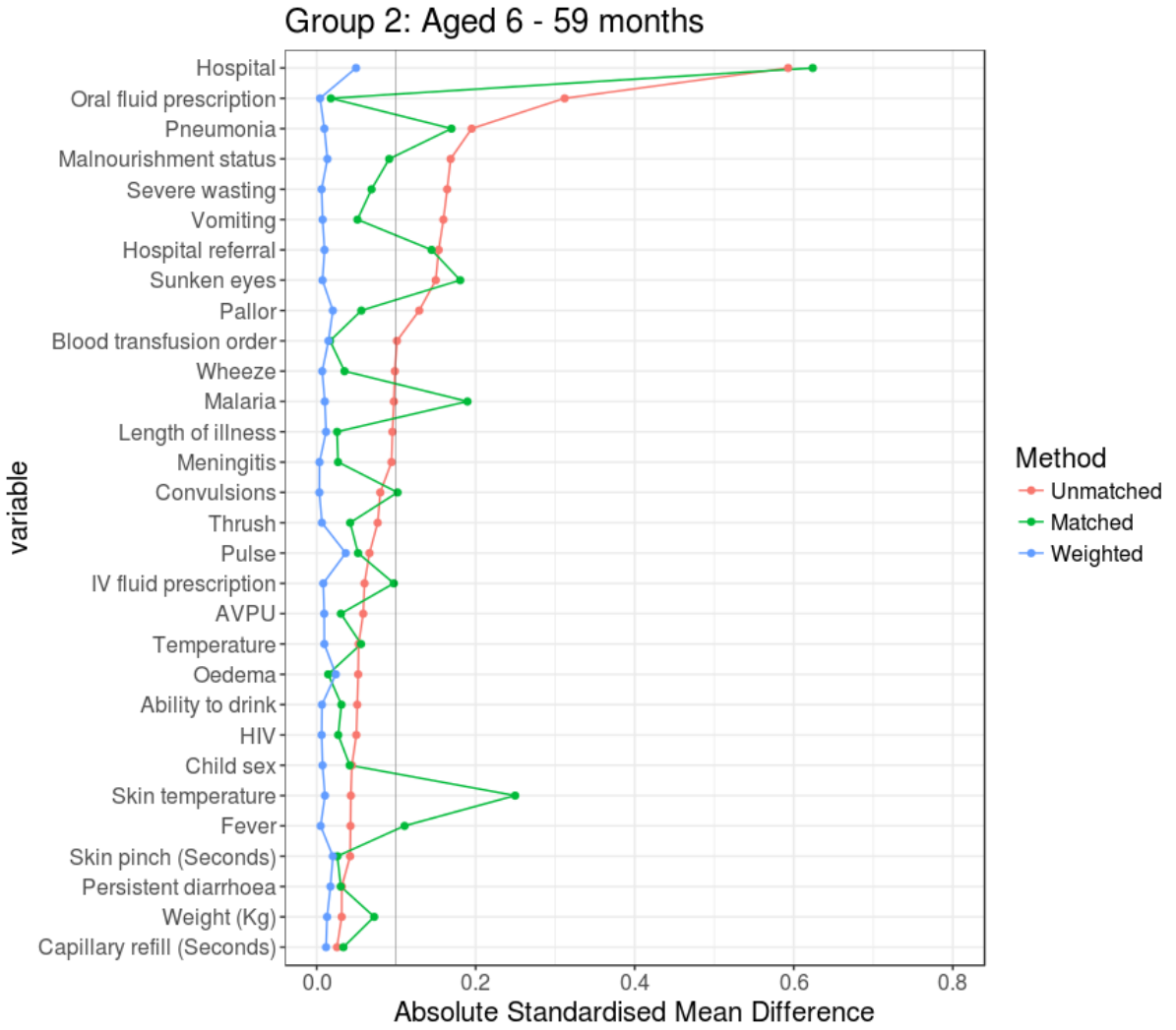


Figure 5.0.3: Comparing performance of the two PS implementation methods in group 2 (6 – 59 months)

5.3.3. Outcome analysis

5.3.3.1. Exploring probability of being discharged alive in age groups 1 and 2

Cumulative incidence functions were explored in the raw datasets (for groups 1 and 2) to describe the probability of being discharged alive over time (**Figure 5.0.4**). Approximately 60% of the children in both age groups were discharged by the fifth day of their stay in the hospital. In both age groups and without PS

adjustment children who received Zinc were discharged sooner than those who did not receive Zinc.

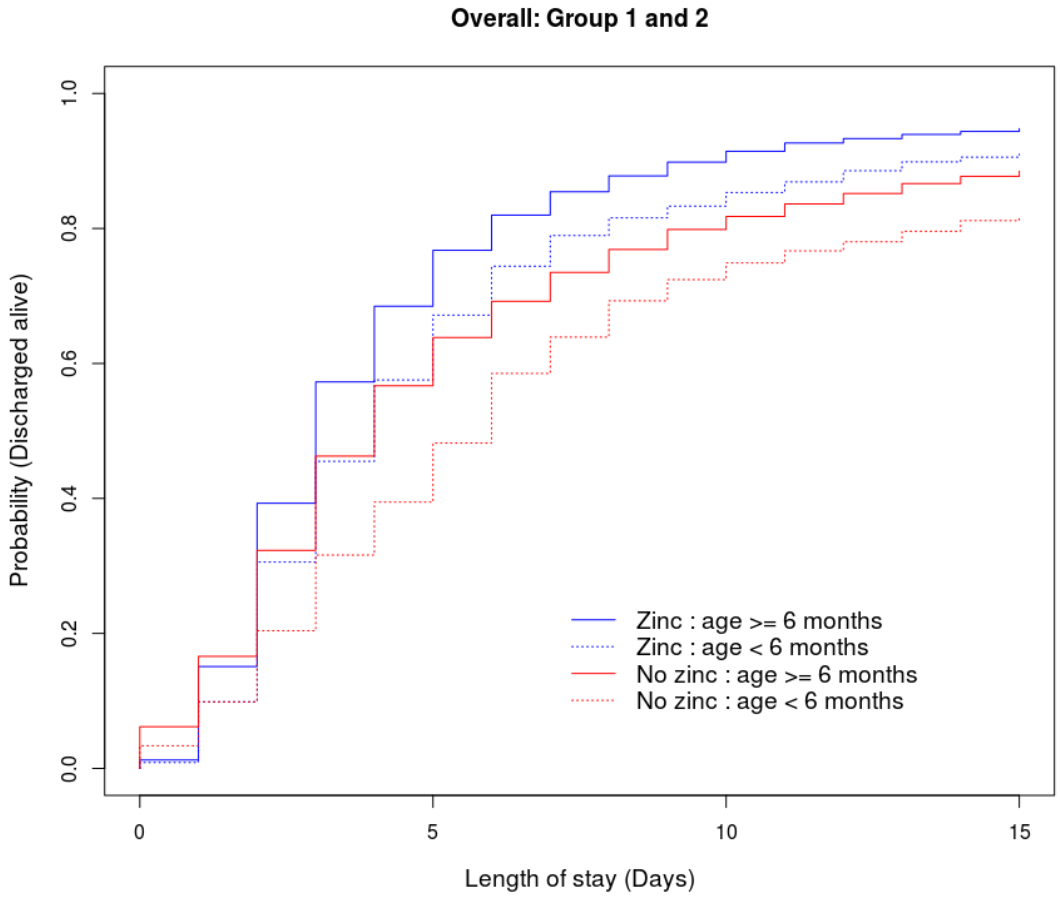


Figure 5.0.4: Probability of getting discharged alive (the cumulative incidence curves were estimated on datasets without PS adjustments)

5.3.3.2. Modelling the probability of getting discharged alive

The estimated sub-distribution hazard ratios (SHR) in multivariable PS weighted models (of being discharged alive in the Zinc group versus the non-Zinc group) were 1.25 [1.07, 1.46] and 1.17 [1.10, 1.24] in groups 1 and 2 respectively. In a pooled analysis (across the two age groups) with weighting based on precision of the estimates, the overall effectiveness of Zinc was 1.17 [1.11, 1.24].

This can be interpreted as a child receiving Zinc having on average a 17% higher chance of being discharged at any point in time than a child not receiving Zinc (**Figure 5.0.5**). Multivariable models without PS weighting produced somewhat higher effect estimates (SHR 1.33 [1.16, 1.53] and 1.30 [1.23, 1.38] for age groups 1 and 2 respectively) suggesting PS weighting may be adjusting for some potential confounding favouring Zinc treatment. Kolmogorov-Smirnov and Cramer Von Mises tests showed that Zinc had constant effects across all the time points prior to discharge (see **Appendix E.2.3)**.

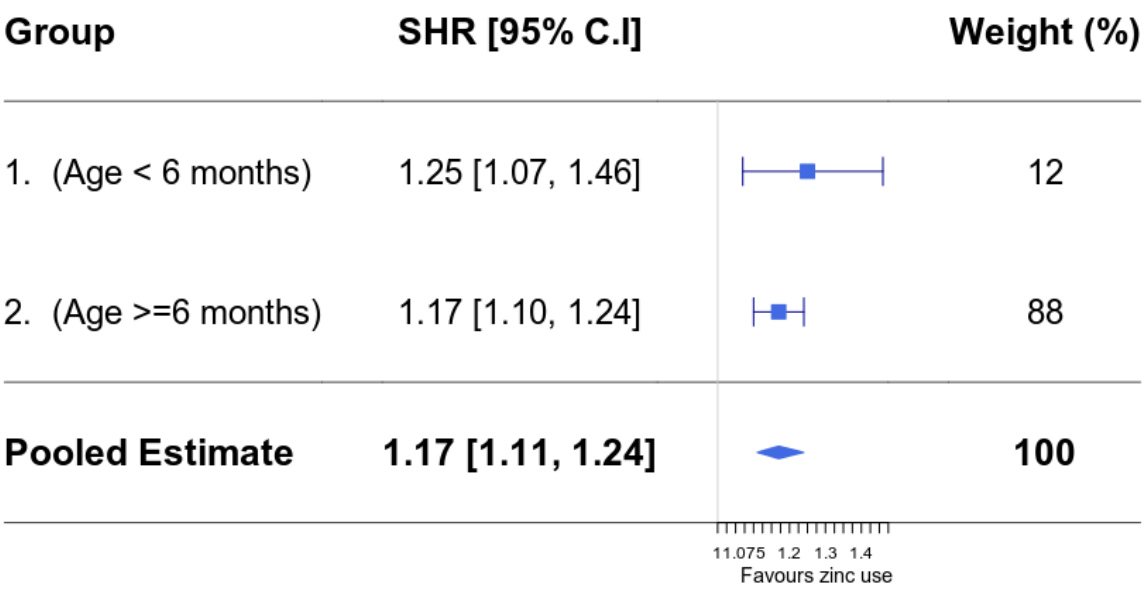


Figure 5.0.5: Estimated sub-distribution hazard ratios in groups 1 and 2. An SHR value greater than 1 favours the use of Zinc.

5.3.3.3. Analysis without censoring those who stayed beyond 15 days

I repeated the analysis of time to discharge without censoring those who stayed beyond 15 days and the estimated SHR were approximately equivalent to what was obtained in **Figure 5.0.5** of the main results (Group 1: 1.38 [1.14, 1.67] and; Group 2: 1.24 [1.16, 1.34]).

5.3.3.4. Zinc effectiveness for well-nourished and malnourished children

Approximately a third of children in both age groups were severely – moderately malnourished (1 – 5 months: 28.6% and 6 – 59 months: 29.1%). PS plots indicated substantial overlap in the PS distributions of the four subgroups, that is, each child had some probability of belonging to any of the four subgroups (see **Figures 5.0.6 and 5.0.7 a – d**).

Figure 5.0.6: Examining common support for children aged 1 – 5 months

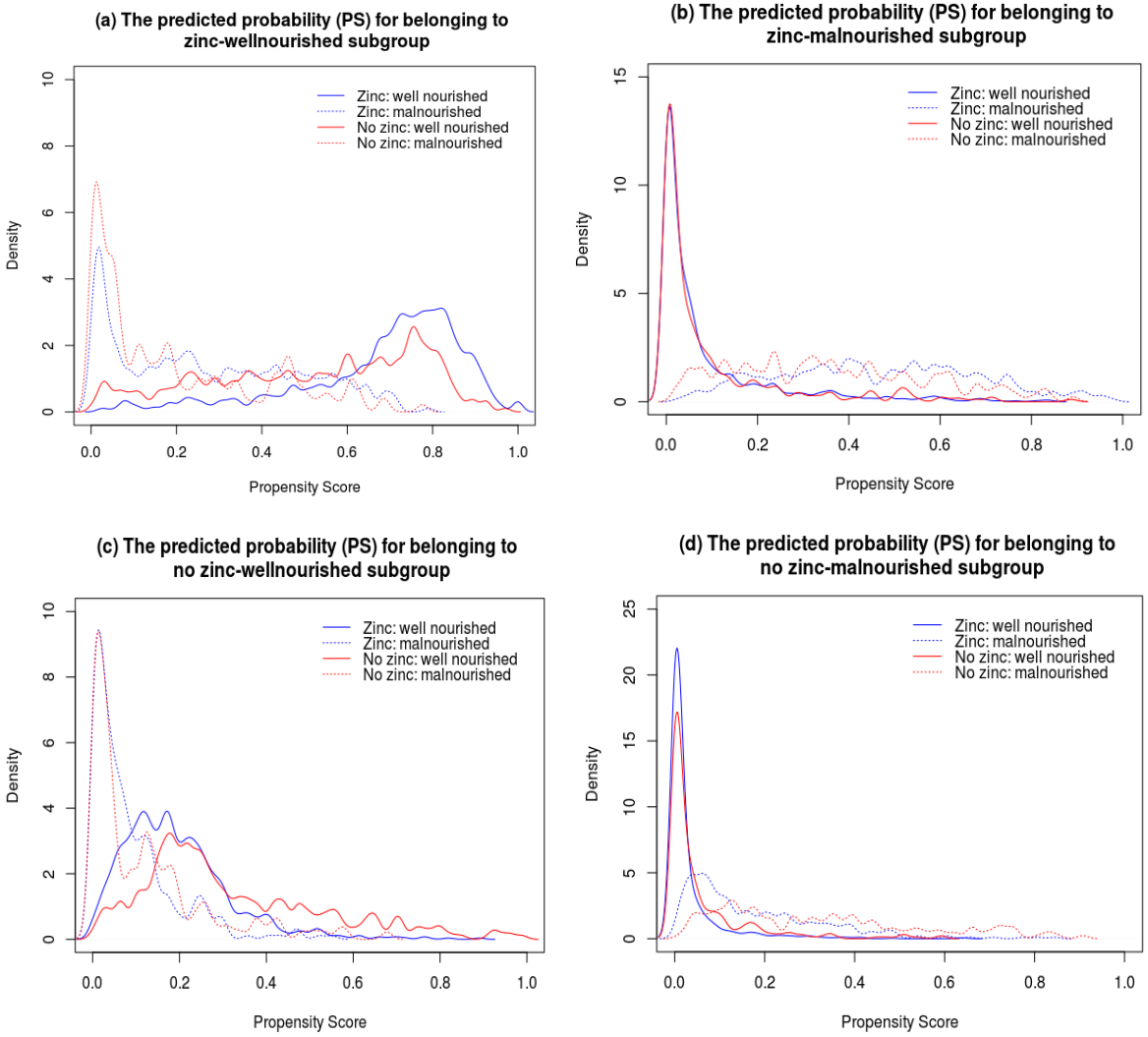
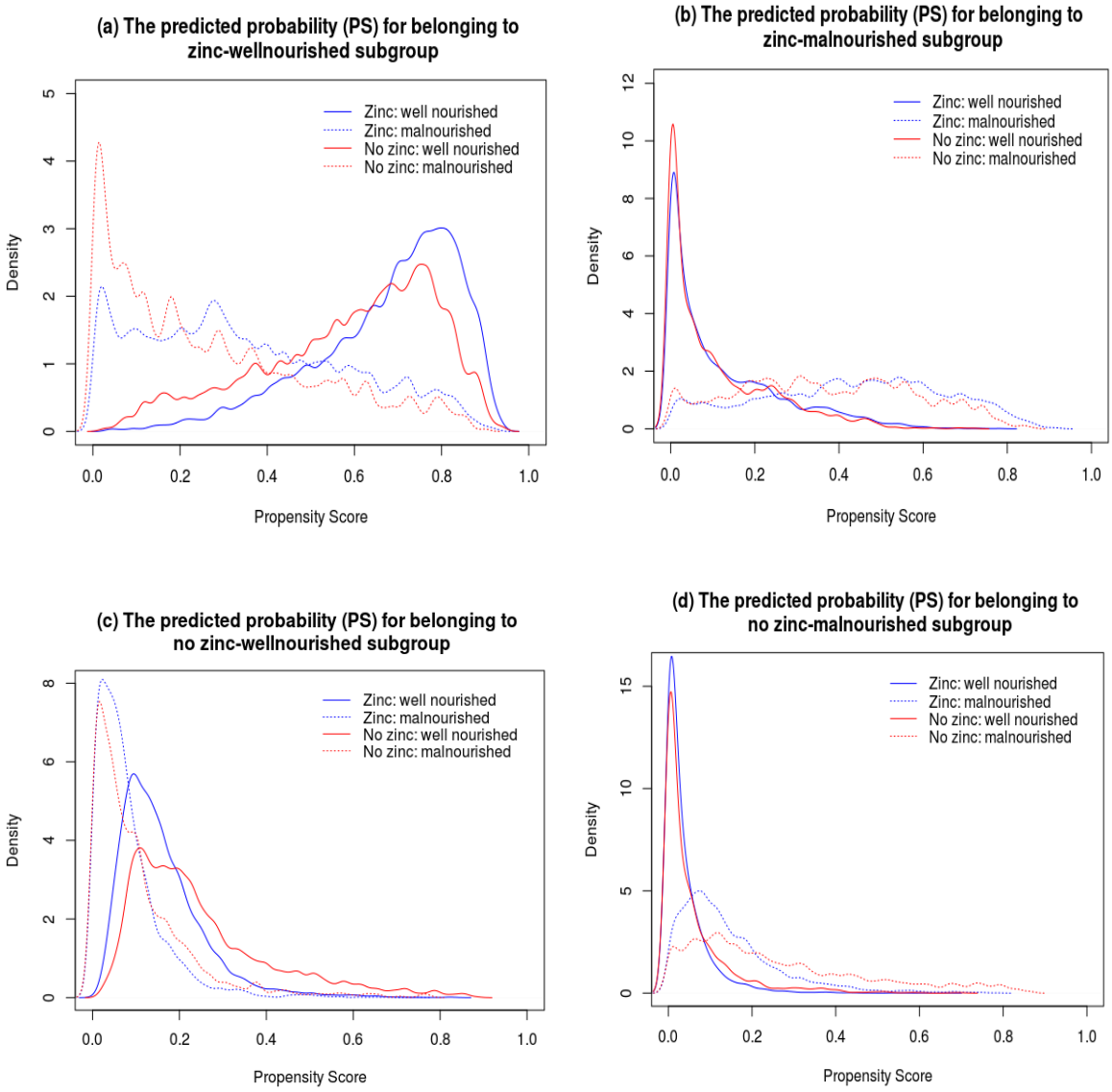


Figure 5.0.7: Examining common support for children aged 6 – 59 months



Also as demonstrated using likelihood ratio tests, the distribution of patient characteristics across the four subgroups were approximately similar after PS adjustments (see **Appendices E.2.4 and E.2.5**) – and this implies that interpreting the outcome results of one subgroup in comparison with another would be reasonable.

In analysis by nourishment status, PS adjusted Scheike's flexible model estimates suggested that those who received Zinc and were well-nourished were more likely to be discharged sooner, followed by: those who did not receive Zinc and were well-nourished, then those who received Zinc but were malnourished and, lastly those who did not receive Zinc but were malnourished (**Table 5.0.1**, also see **Appendix E.2.6** on cumulative incidences). Results interpreted in a similar way were found using PS unadjusted data (see **Table 5.0.1** footnote for the actual results).

Table 5.0.1: PS adjusted SHR

	1 – 5 months	6 – 59 months
Zinc-malnourished	0.82 [0.67, 1.01]	0.85 [0.79, 0.91]
No Zinc-wellnourished	0.88 [0.74, 1.04]	0.90 [0.84, 0.98]
No Zinc-malnourished	0.63 [0.49, 0.82]	0.64 [0.58, 0.70]
Zinc-wellnourished (reference group)	-	-

*Using PS unadjusted data, the following effects were estimated: (i) 1 – 5 months (Zinc-malnourished: 0.52 [0.44, 0.61], No Zinc-wellnourished: 0.74 [0.63, 0.87], No Zinc-malnourished: 0.41 [0.33, 0.51]); (ii) 6 – 59 months (Zinc-malnourished: 0.57 [0.54, 0.60], No Zinc-wellnourished: 0.83 [0.77, 0.90], No Zinc-malnourished: 0.41 [0.38, 0.45]).

5.3.3.5. Sensitivity analysis with weekday/weekend admission as an instrumental variable

To assess whether weekend/weekday admission variable forms a natural and random experiment the distributions of covariates were examined across the levels of the instrumental variable (grouped as weekend/weekday) for children aged 1 – 5 and 6 – 59 months separately. The distribution for each of the patient characteristics was approximately similar with the majority of variables having absolute standardised mean differences of <= 0.1 between groups (see **Appendix E.2.7**). Likelihood ratio tests also showed that weekend/weekday admission variable was significantly associated with Zinc prescription in both age groups.

Further, mediation analysis demonstrated that time to discharge was significantly associated with weekend/weekday admission variable through the treatment (Zinc/No Zinc). The estimated SHRs (comparing Zinc versus no Zinc) using this IV for children aged 1 – 5 and 6 – 59 months were 1.24 [1.18, 1.30] and 1.31 [1.27, 1.35] respectively. These results are interpreted in a similar way as those reported in **Figure 5.0.5**. Consistent results with the primary analysis were also obtained in IV analysis of Zinc effectiveness by nourishment status.

5.3.3.6. Modelling inpatient mortality (exploratory analysis)

Both Kolmogorov-Smirnov and Cramer Von Mises tests showed that the use of Zinc had varying effects on mortality (across the discharge time points) both for children aged 1 – 5 and 6 – 59 months (see **Appendix E.2.8**). Thus, time specific estimates (with corresponding 95% C.I) were modelled using Scheike's regression and the results are presented in **Figures 5.0.8 a and b**, also see **Appendix E.2.9**). For children aged 1 – 5 months, Zinc seemed to have no effect on mortality for those who stayed in the hospital for about a week. However, beyond one week, those who received Zinc seemed to have significantly reduced risk of dying. While Zinc use was associated with reduced mortality for children aged 6 – 59 months across all the discharge time points.

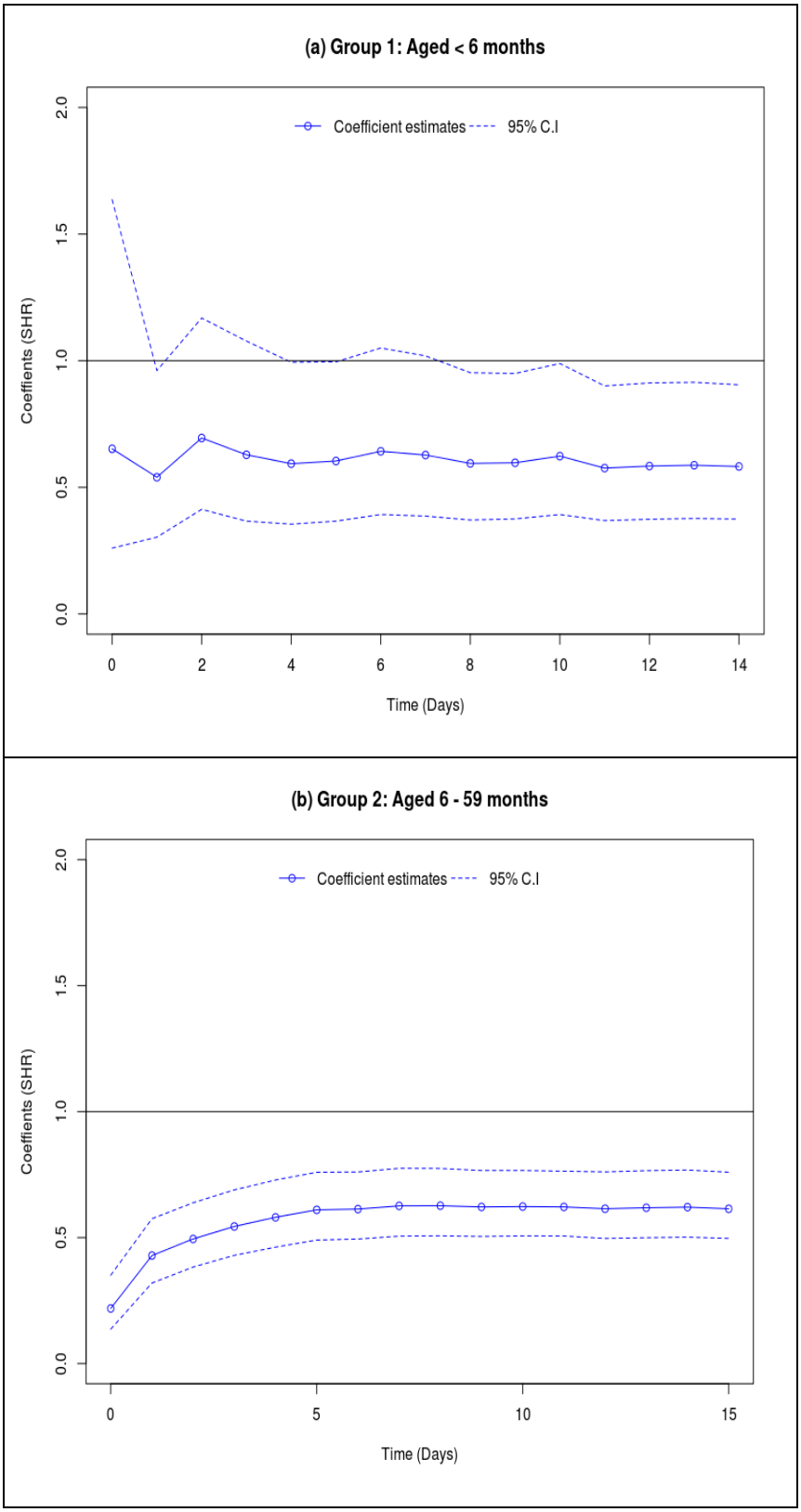


Figure 5.0.8 a and b: Estimated mortality time-varying SHR with corresponding 95% C.I in groups 1 (1 – 5 months) and 2 (6 – 59 months): The continuous blue line shows the estimated SHR coefficients by discharge time point and the non-continuous blue lines show the 95% C.I.

5.3.3.7. Multiple imputation sensitivity analysis

The sample sizes per pattern were as presented in **Table 5.0.2**. Pattern 1 in both groups had the largest sample size, followed by pattern 2 and lastly pattern 3.

Table 5.0.2: Sample size per pattern

Pattern	Group 1	Group 2
1	758	5700
2	346	2140
3	541	3706
Total	1645	11546

PS weighting was conducted (to be consistent with the primary analysis) and this minimised covariate imbalance across the three patterns in group 2 and all the variables had ASMD < =10%. However, about a third and half of the variables in patterns 1 and 2 respectively in group 1 had ASMD > 10%, and six variables had ASMD > 10% in pattern 3 of group 1 (see **Appendices E.2.10 and E.2.11**). The estimated Scheike's model Zinc treatment effectiveness were as presented in **Figures 5.0.9 a and b**. The pattern specific trends were in opposite directions although in group 2 the effect estimate was larger where there was most missingness. However, the pooled effects were consistent with those observed in **Figure 5.0.5**. This indicates that the earlier assumption of data missing at random was plausible.

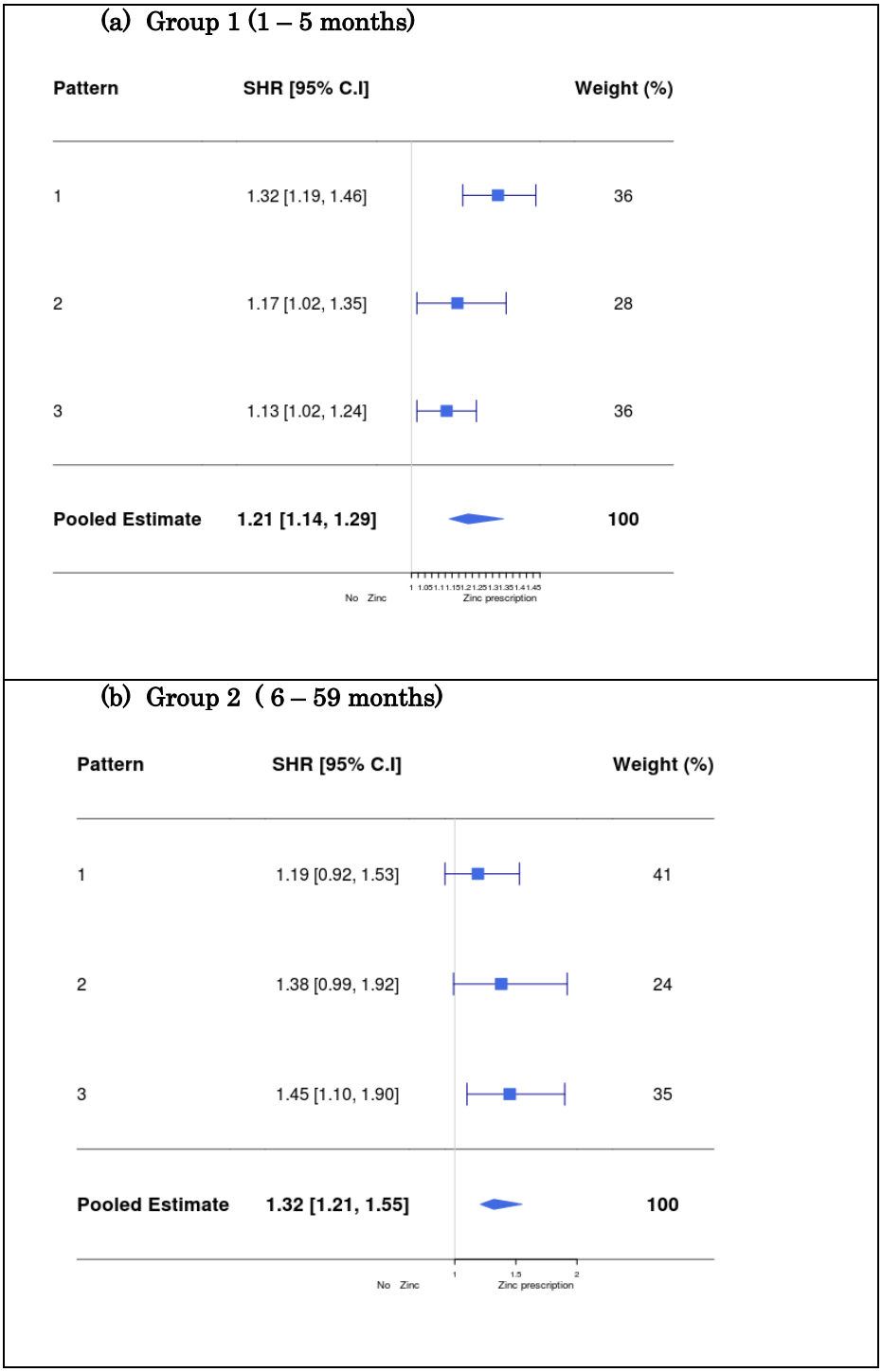


Figure 5.0.9 a and b: The estimated SHR per pattern in groups 1 and 2. Interpretation is based on the pooled effects rather than pattern specific.

5.4. Discussion

Zinc prescription was above 70% and this has changed little over time indicating good policy adoption. This is potentially due to wide dissemination of paediatric protocols that also provide guidance on diarrhoea and dehydration classification and prescription of fluids.

The results indicate that Zinc may be beneficial in reducing time to discharge for children aged 1 – 59 months admitted with diarrhoea. An association with benefit is seen for both well-nourished and severely – moderately malnourished children. This analysis supports and strengthens the current treatment policy used in Kenya which was implemented on the basis of limited evidence supporting the use of Zinc in those aged less than 6 months and in well-nourished children. Two previous trials conducted in Africa (Ethiopia (n = 177) and Nigeria (n = 60)) with children under 6 months (25, 137) showed no effect of Zinc in reducing diarrhoea duration. These analyses offer the advantage of larger sample size but have the limitation of being based on a non-randomised study. Previous studies (25) have reported no serious adverse effects associated with the use of Zinc apart from a risk of increased vomiting (within 10 minutes of administration). I was unable to examine differences in this possible adverse effect but if it occurs it does not seem to be increasing the length of stay for hospitalised children with diarrhoea. The analysis conducted without censoring children who stayed in the hospital beyond 15 days had consistent findings comparable to those in **Figure 5.0.5**. The findings therefore seem to contribute additional evidence on the likely value of Zinc in support of its routine use for hospitalised children with diarrhoea in Kenya.

5.4.1. Strengths and limitations

The use of propensity scores and multiple imputation provided a means to analytically handle the challenges of non-random allocation and missing data by creating samples of patients that are comparable in terms of *observed* signs, symptoms, co-treatments and comorbidities. Analyses of these samples provided biologically plausible results showing the benefits of Zinc in reducing time to discharge to all children below five years. However, as analytic methods used in observational studies may not completely eliminate unobservable bias (148, 149) (that is bias due to unobserved factors that may influence time to discharge), I conducted sensitivity and exploratory analyses with the use of an instrumental variable and time to experiencing mortality respectively. The findings of instrumental variable analyses were consistent with those obtained with the PS methods. Also this analysis demonstrated that Zinc was associated with reduced mortality – and this was consistent with the findings of a community trial in Bangladesh, which showed reduced chances of mortality by 50% (150). It may be biologically less plausible that Zinc would result in any effect on very early mortality (day 1 or even day 2), but large effects are seen in this early phase. This may be an indication of residual bias reflecting decision not to give Zinc to children who are the most severely ill. These findings on mortality, are also in contrast with those of RCTs included in the systematic reviews (25, 151), which showed no effect of Zinc on mortality.

It is also possible that shorter length of stay may not only be due to administration of Zinc treatment but also as a result of improved clinical

symptoms or based on clinician judgement of attaining clinical stability.

5.5. Conclusion

The findings that Zinc is associated with shorter length of stay is consistent with meta-analyses of RCTs but extends this observation to less studied populations. RCTs have, however, not shown an advantageous effect of Zinc on mortality in acute diarrhoea. The protective association of Zinc against mortality we see does raise the possibility that there is residual (unobservable) bias in my observational analyses possibly due to clinicians allocating this treatment to less sick children (which also raises the possible influence of residual bias on effects seen on time to being discharged). Despite this caveat, and as Zinc is well tolerated and cheap, these results support the continued use of routine Zinc supplementation in all children aged 1 – 59 months hospitalised with diarrhoea in Kenya and probably elsewhere in Africa. Also it is easier to have a single rule for one population than different rules for multiple sub – populations.

Chapter 6

Discussion

In this thesis, I have examined the applicability of propensity score methods in observational datasets for comparative effectiveness evaluations. First, I conducted a systematic review to examine how these methods are applied in practice. Then the systematic review informed the methodology applied in the two clinical questions, first to examine alternative antibiotic regimens in treatment of indrawing (intermediate severity) pneumonia and secondly to explore the effectiveness of Zinc in children with diarrhoea admitted to hospitals in Kenya (these are discussed further in section 6.2). For the pneumonia analyses I focused on mortality as the main outcome while in the case of Zinc the primary outcome was length of stay (as a proxy for the severity and duration of the diarrhoea) with effects on mortality examined in exploratory analyses. I used the Clinical Information Network dataset to address both questions. I was part of the team that established the data collection and management strategies for the CIN, a database that was created as an example of those that may become available in Africa as health information systems improve, particularly if electronic medical records become more widely adopted.

As well as addressing two important clinical questions, for which there is limited evidence supporting current policy, my thesis work makes a specific contribution in exploring the methodology of the use of propensity score (PS) methods. This

includes the strategy for estimating PS and the use of multiple imputation in the process (with outcome data included and excluded in chained equations). Here I discuss some of the implications of my thesis work and its findings in sub – sections 6.1 – 6.5 before considering some of the strengths and limitations of my work in sub – section 6.6. I discuss areas for further research in sub – section 6.7, recommendations in sub – section 6.8 and conclude in sub -section 6.9.

6.1. Use of routine datasets

Large routine databases like the Clinical Practice Research Data Link (CPRD) (152) and those linked to the Kaiser Permanente (KP) healthcare system (153) have been built in the UK and USA respectively. The emergence of such datasets has seen an increase in observational comparative effectiveness studies over time (45). Comparative effectiveness of treatments using these datasets have focussed on cardiology (examples – Suissa (2016), Gamble (2016), and Hennessy (2004)) (154-156), diabetology (examples – McEwan (2014) and Nyeland (2013)) (157, 158), pain (examples – Allen (2016) and Allen (2015)) (159, 160) among other conditions. On the other hand, there are few examples of large, good quality routine clinical databases in Lower and Middle Income Countries (LMIC). One LMIC example is the Analysing Longitudinal Population-based HIV / AIDS data on Africa (ALPHA) network (161) – which was established in 2005 and is ongoing in sites in Uganda, Kenya, Malawi, Zimbabwe, South Africa and Tanzania. The network aims to broaden HIV epidemiology evidence and strengthen HIV analytical capacity and cooperation between participating sites. To date it has accumulated data on more than 1.5 million HIV patients. There

are few such systems that capture data from LMIC paediatric hospital care. In one example, paediatric diagnosis and outcome data have been collected over six years in 16 provincial hospitals in Papua New Guinea (162) with a total of about 100, 000 patient records. The emergence of electronic medical record systems (EMRs) holds great promise as a means to generate much larger routine clinical datasets in LMIC – but implementing them remains a major challenge (including in Kenya) as it has been in many countries and typically high levels of funding are required (163).

In 2013, we established the Clinical Information Network (described in detail in Chapter 1 sub – section 1.2.2) as a model to retrospectively collect high quality data capturing the assessment, diagnoses and treatments of clinical teams providing routine care on paediatric wards in Kenya's public hospitals and the outcomes of this hospital care. These hospitals typically have one paediatrician leading services predominantly provided by junior clinical teams. Data systems and standardised clinical forms were specifically implemented in all hospitals at the start of this work to optimise the quality of routine data. Patient data in these hospitals are collected post discharge by trained data clerks guided by well-defined standard operating procedures, under supervision by the hospital medical records department and the research team. Clinicians admitting patients fill standardised Paediatric Admission Record forms (Appendix A.2) that have been shown to improve documentation of clinical symptoms and signs (164). Together with discharge forms, treatment sheets and laboratory reports these are all part of the patient files that are the primary data source. Feedback to

hospitals as part of the CIN activities has helped improve the quality of CIN's clinical data (165, 166). The CIN platform therefore provided a unique opportunity to examine the potential value of routine paediatric datasets in comparative effectiveness evaluation from a LMIC.

Comparative effectiveness research (CER) is a potentially useful strategy for a number of reasons: (i) there may be limited RCT data and / or debate over the ability to generalise findings from existing RCTs; (ii) RCTs are more often of the explanatory type with inclusion / exclusion criteria that limit generalisability – while CER could potentially assess effects across a large range of patients that present to typical facilities and; (iii) CER might therefore provide insight into effectiveness by providing data on the actual outcomes of care after a guideline has been decided at policy level.

While RCT design has become much more rigorous over the years, the same is not always true for observational designs. It is well known that drawing inference from non-RCT studies starts with higher risk of bias. This is despite the efforts to improve the reporting of observational studies through the use of checklists like STROBE (60). In fact, observational studies require more careful attention to their design and analysis as they are faced with challenges that might compromise their internal validity (38).

To try to limit these challenges, it is important to replicate the strengths of good RCTs discussed in Chapter 1 sub – section 1.1.3 by paying especially careful attention to the design of an observational study. Guiding principles to achieving this have been discussed in Chapter 1 sub – sections 1.1.6.1 – 1.1.6.5.

Most important when contrasting alternative treatments is overcoming the analytic challenges of non – random allocation of treatments and missing data as routine datasets are not designed for research. In my work I took advantage of failure of clinicians to adhere to recommended treatment guidelines (for Kenya), a process that may not be random. This could result in routine datasets that yield treatment groups (patients for whom there is / is not adherence to guidelines) that are not comparable in terms of clinical characteristics. The use of propensity scores has been proposed as a means to 'balance' patient groups to make them comparable in terms of important (in this case clinical) characteristics (167). Since routine datasets are limited to observed variables, it is important to conduct sensitivity analyses to examine if any results observed could be influenced by unobserved variables.

This thesis used propensity score trimming – excluding patients from the tails of the PS distribution then re-estimating the treatment effects; and instrumental variable analysis in this regard. The use of such sensitivity analyses methods was rare (and where found they were not rigorously employed) in the articles reviewed in Chapter 2, though articles mostly acknowledged the possibility of their results being biased due to unobserved confounders (see a discussion in sub – section 2.2.3 of Chapter 2). The careful attention I paid to the design of the CER studies I report and linked sensitivity analyses are therefore unusual in recently published literature. I discuss the particular strengths and limitations of using these sensitivity analysis methods in more detail below (sub – section 6.6).

6.2. Clinical questions addressed

This thesis focused on two important clinical questions for contemporary pneumonia and diarrhoea case management, illnesses that remain two leading causes of morbidity and mortality in Kenyan children (5). First examining the comparative effectiveness of penicillin plus gentamicin versus penicillin alone in treating indrawing pneumonia. This question was important as Kenyan pneumonia treatment guidelines were changed in March 2016 – with recommendations made that children with indrawing pneumonia are treated with amoxicillin in outpatient settings. However, children with indrawing pneumonia have continued to experience non – trivial mortality which has been a concern of Kenyan practitioners and those from other settings (19). Therefore, it was important to test the hypothesis that the use of a broader spectrum antibiotic regimen (penicillin plus gentamicin) might result in improved clinical outcomes for this group of patients. The comparative effectiveness of these two antibiotic regimens (narrow and broad spectrum) was examined in three patient populations referred to as experiments. Experiment 1 constituted patients for whom pneumonia severity was imputed using the clinical signs recorded by the clinician at admission assigning a classification in accordance with the treatment guideline recommendations; experiment 2 constituted patients whose pneumonia severity was determined/assigned by the clinicians – including only those they classified as indrawing (intermediate severity) pneumonia (this experiment included 80% of the experiment 1 patient population as might be expected). Experiment 3 focused on all forms of pneumonia severity, including the most severely ill group who, according to national guidelines should only have been

prescribed the broad spectrum regimen. Stratification of analysis into three experiments was important as experiment 1 mimicked the population inclusion criteria of a typical prospective clinical trial; experiment 2 examined treatment effectiveness in situations where clinician intuition was likely to play a role in severity classification and treatment rather than strict adherence to the guidelines; and experiment 3 was necessary as no trial has examined effectiveness of narrow and broad spectrum therapies in all forms of pneumonia severity admitted to hospitals.

The second clinical question focused on examining effectiveness of Zinc in children admitted with diarrhoea and specifically those aged 1 – 5 and 6 – 59 months, and in those classified as either well-nourished or moderate-severely malnourished. These questions were derived from a recent systematic review by Lazzerrini (2016) (25) which included 33 trials (n = 10 841) in which only two were conducted in Africa (Ethiopia (n = 177) and Nigeria (n = 60) – and both examined Zinc efficacy in children less than six months of age. The review concluded there were inadequate data to support the use of Zinc in children aged 1 – 5 months and in those who are well-nourished. Thus, this analysis was important to determine whether the use of Zinc in all children aged 1 – 59 months is appropriate as these are the recommendations in the Kenyan paediatric treatment guidelines (and in the guidelines of many African countries) made on the basis of limited evidence in these specific sub-populations.

The Zinc analysis also presented a case which considered the use of time to event in observational comparative effectiveness evaluation in which PS methods were

used to adjust for baseline differences. This outcome was based on the findings of the systematic review – that demonstrated an effect of Zinc to reduce diarrhoea duration but no effect on mortality. I believe these analyses may be the first to use a competing risks approach to test an intervention effect in the context of CER. Here the analysis examined the probability of being discharged alive from the hospital. However, in case of mortality, the possibility of this event would be precluded. Thus this methodology should have accounted for possible 'survivor bias' in that averting early mortality (a poor outcome) could result in longer average length of stay.

I discuss the findings of these clinical questions in sub – section 6.5 but before this I consider some of the methodological issues I encountered.

6.3. Methodological strategies used in published studies – addressing deficiencies

In Chapter 2, I updated the existing reviews on the use of propensity scores in the last 7.5 years specifically focusing on observational comparative effectiveness research. Most importantly I examined how PS models are estimated in the presence of missing data in the most recent literature.

The majority of observational comparative effectiveness studies used PS matching methods that reduced the effective sample size. The commonest method was 1:1 pair matching of treatment and (active) control groups. Even with correct reporting of the methodological aspects of PS 1:1 pair or 1:k matching, Gary (2016) (91) indicates that this method results in pruning of data which may increase imbalance across groups potentially resulting in increased bias in the analysis. As an improvement on commonly used methods I compared

three PS methods (on the imputed datasets) that do not reduce the effective sample size – optimal full matching, weighting and sub – classification. The PS methodological approaches I employed are much less commonly reported in published literature.

This literature also rarely includes justification for variables selected for inclusion in the PS models, checking of common support before proceeding with the analysis and the conduct of sensitivity analysis methods to explore for unmeasured confounders. Thus my work addressed one of the key concerns expressed by Rubin (2008) (39) who argued that it is important to examine for common support to ensure that propensity score densities are sufficiently overlapping across treatment groups and that one should not proceed with analysis if densities do not overlap satisfactorily. Other PS methodological aspects used in this thesis (consistent with those reported in previous articles) included assessment of covariate balance between treatment groups and specification of the PS estimation method (logistic regression), treatment estimand (ATT for both clinical questions), and number of strata used for PS sub – classification (five strata).

I found that missing data are ignored or poorly addressed when estimating propensity scores even in the most recent literature. In a majority of articles reviewed in chapter 2 complete case analyses were conducted – which may result in biased estimates if data are not missing completely at random (MCAR, that is the missingness is unrelated to the study/ interventions). It is possible that a majority of researchers do not understand or ignore the process that generated

the observational datasets they are using for analysis. Variables in routine datasets are likely to have different processes that generated them and by extension different missingness mechanisms and efforts should be made to understand these.

I made an assumption of data missing at random (MAR not MCAR) and then proceeded to conduct multiple imputation and thereafter estimated PS on imputed datasets using the within method (discussed in chapter 2 sub – section 2.1.2) for the pneumonia analysis in Chapter 3. Completeness was used as one measure of the quality of data for different variables (also used for multiple imputation). The CIN dataset captured a majority of the key variables that were useful in the design phase of my experiments and in very rare cases proxies were used for variables that had poor documentation – for instance pulse oximetry values were missing for the majority of patients and hence a record of clinician ordering of oxygen was used as a proxy for clinically determined hypoxia. Multiple imputation was important at the design phase as particular key variables needed to be observed in order to generate patient populations that formed experiment 1 of my pneumonia analyses (otherwise the effective sample size would be reduced by approximately 40% with complete cases). To ensure that the most plausible values were imputed, auxiliary variables were added to the imputation models as predictors. Auxiliary variables were not part of guidelines but might (on the basis of clinical logic informed by discussions with clinicians) still influence treatment assignment – for instance the occurrence of a convulsion in children with pneumonia.

In the next sub – section, I provide an overview of how I attempted to optimise PS methods (optimal full matching, weighting and sub – classification) and implement these together with multiple imputation to answer clinical questions on pneumonia and diarrhoea treatment.

6.4. Optimising the methodological approach

In Chapter 3, differences in the covariable balance between treatment groups (for optimal full matching, weighting and sub – classification) were examined using the absolute standardised mean differences (ASMD) as the most recommended approach (167). This is in contrast to the more widely used approach in the literature of using statistical testing I report in Chapter 2. In experiment 1, PS weighting performed better than optimal full matching and sub – classification; in experiment 2, the performance of optimal full matching and weighting was comparable – with sub classification performing the poorest in bias reduction. Only sub – classification was used in experiment 3 as this experiment aimed to investigate comparative effectiveness in all cases of pneumonia in which the propensity scores were used as a proxy for disease severity. Thus patients with lower propensity scores were considered less ill while those with higher propensity scores were considered more ill.

As analysis in chapter 3 was restricted to the observed variables with outcome data excluded from the imputation models, Chapter 4 examined various sensitivity analyses (only on the pneumonia data analysed in Chapter 3). This included use of imputation strategies including outcome data in the imputation models and application of this approach to the two methods (across and within)

for estimating PS after multiple imputation. Further, it examined if the results obtained in chapter 3 would be influenced by unobserved variables. This involved the use of instrumental variable analysis and PS trimming. In addition to these, time to experiencing inpatient mortality was also analysed to investigate the influence of unobserved confounders in the Zinc analysis.

In Chapter 4, PS trimming created more homogeneous patient populations in which treatment effect differences seemed to move towards the null as increasingly stringent restrictions were placed on population eligibility in terms of PS. Effect estimates were consistent across those obtained using PS weighting (without PS trimming) and when using instrumental variable analysis. Instrumental variables have been used mostly in the field of economics but recently have gained popularity in the medical field – see for example (168-170) where these have been applied. Estimates similar to those derived using PS methods were also obtained using the instrumental variable analyses in the analysis of Zinc data.

In Chapter 4, I also examined the possible influence of inclusion and exclusion of outcome data in the imputation models with PS estimated using the across and within approaches. There was some suggestion the two approaches performed differently in terms of creating comparable treatment populations with reduced bias for PS full matching, weighting and sub – classification. PS weighting eliminated the largest amount of bias when outcome was included in the imputation and PS were estimated using the within method. However, PS sub-classification and full matching reduced the largest amount of bias when

implemented with the across method with outcome data excluded from the multiple imputation chained equations.

Even though the across method of estimating PS after multiple imputation resulted in populations with minimised bias, Leyrat (2017) (116) has demonstrated that the across method yields PS estimates that are not statistically consistent. See the proof presented in **Appendix F.1**. Going by this proof by Leyrat (2017), then it means it is only the within method that should be considered a theoretically and practically valid PS estimation method. In chapter 4, the inclusion of outcome data in the imputation models (with PS estimated using the within method) had the potential of eliminating some bias compared to when outcome data were excluded in the three PS implementation methods – full matching, weighting and sub - classification. Therefore, my findings suggest it would be recommended, as a standard approach, to include outcome data in the multiple imputation chained equations with PS estimated using the within method when using any of the PS implementation methods. Outcome data should then be hidden after performing multiple imputation before the investigator proceeds with the design phase of their experiment. In my work in this thesis, PS matching and weighting have been demonstrated to perform better than sub – classification – though PS weighting and matching have performed differently on different datasets. Thus it is important that PS weighting and matching must be tried and the method that produces the smallest ASMD (with respect to observed variables) would be most preferred.

Based on sensitivity analyses in Chapter 4, the second clinical question (on the use of Zinc for diarrhoea admissions) compared only PS optimal full matching and weighting (as they performed better than sub – classification). PS weighting performed better in terms of minimising covariate imbalance for both patient groups (1 – 5 and 6 – 59 months). The analyses here used multiple imputation including the outcome data with PS estimated using the within method. The work I have presented leads me to suggest a set of methodological principles I present in summary form in **Box 6.4.1** for future work using observational datasets to conduct CER.

Box 6.4.1: A summary of the methodological steps that are proposed when using observational datasets for CER

- Create preliminary experimental design including sample size estimation.
- Identify key and auxiliary variables.
- Examine data missingness per variable.
- Make an effort to understand the possible missingness mechanisms in the data.
- Use imputation including outcomes but then blind the investigator to these outcomes.
- Test the performance of PS full matching and weighting. In both, use the within method for estimating PS after multiple imputation.
- Examine covariable balance using ASMD.
- Examine viability of experimental analysis and if satisfactory then proceed.
- Conduct multiple sensitivity analysis using methods including instrumental variables and trimming.

6.5. Key clinical findings

In Chapter 3 where I address the treatment of pneumonia, in analyses not employing propensity score methods, effect estimates (particularly for experiments 1 and 2) suggested clinical superiority of penicillin over the use of penicillin and gentamicin. And as indicated in Chapters 1 and 3, the biological

plausibility of this finding is questionable. The use of ASMD demonstrated that almost half of the variables in pneumonia experiments 1 – 3 were associated with ASMD > 10% when comparing the alternative treatment groups in the PS unadjusted analysis. Examining patient distribution in the PS unadjusted datasets (**Appendix F.2**), patients in the penicillin plus gentamicin group seemed to have more positive clinical signs compared to those who received penicillin monotherapy. This points to the possible bias that could influence findings if the datasets were to be analysed without PS adjustments. With the use of PS methods, all the variables had ASMD < 10%. This then resulted in treatment effect estimates indicating no statistical difference between the use of penicillin monotherapy and penicillin plus gentamicin in treatment of indrawing pneumonia in experiments 1 and 2. This demonstrated that the use of PS methods had minimised imbalance on observed variables thus reducing the effect of allocation bias in the analysis. This in general indicates that PS may be useful in minimising imbalance in other comparative effectiveness studies rather than using regression methods alone without PS adjustments.

Clinically, and by extension, my results suggest that there would likely be no advantage in the use of penicillin plus gentamicin over amoxicillin as equivalence had been demonstrated between amoxicillin and penicillin in treatment of children admitted with indrawing pneumonia in Kenyan hospitals (31). The findings of this analysis therefore indicate that the continued use of (injection) penicillin or high dose oral amoxicillin will likely give similar treatment outcomes as treatment using gentamicin and penicillin (both

injections) for this specific group of patients. This is perhaps in contrast to clinicians' expectations that using a broader spectrum antibiotic regimen in this population with non-trivial risk of death could reduce mortality (the basis of doing these analyses). A corollary of these findings may be that in those children who die either that the cause of death is not attributable to a treatable bacterial infection or, if a treatable bacterial infection is present, that a different class of antibiotic entirely is required.

The higher mortality in the group of patients with indrawing pneumonia that I studied than is reported in clinical trials conducted on this population (0.8%) is consistent with a study by Agweyu (2017) (171). This study compared pneumonia patient outcomes between trial and non – trial participants in the same settings and demonstrated that mortality in non – trial settings was somewhat higher than in trial settings (but not significantly so). As there are continued concerns of mortality in pneumonia patient groups characterised by indrawing in non – trial settings (110, 171), it may therefore be necessary to investigate the effectiveness of other broader spectrum antibiotics such as third generation cephalosporins and azithromycin to help in reducing mortality. Or potentially the aetiology of pneumonia in this group of patients may be non – bacterial. This explanation may be supported by the Pneumonia Etiology Research for Child Health (PERCH) study conducted in multiple sites (including Baltimore, Mali, Kenya, Gambia, Zambia, Bangladesh and Thailand) with preliminary findings in Kenya showing that more than half of the patients studied had viral pneumonia (172). The PERCH study was initiated to help examine the causes of pneumonia

in order to improve on the diagnosis and treatment. This was motivated by the fact that pneumonia could potentially have multiple causes including bacteria, viruses and fungi. And as diagnosis and treatment of pneumonia in LMIC is based on the WHO guidelines – in which only signs are used (without thorough diagnostic work up) in diagnosis and classification - it seems likely that the causative organisms and mechanisms may be very heterogeneous in children. Thus, in a case where the cause is non – bacterial, reliance on treatment with antibiotics may not improve health outcomes and instead other interventions may be needed. These may include oxygen systems, continuous positive airway pressure (CPAP) (173), or better diagnosis of respiratory distress which may be caused by other illnesses but be confused with pneumonia (174, 175).

6.6. Strengths and limitations

Conducting comparative effectiveness analyses using observational datasets can offer the advantage of larger sample sizes at lower cost than randomised controlled clinical trials (39). The CIN data used for analyses in this thesis had a larger sample size, with data derived from multiple sites, compared to many of the individually randomised studies that have examined treatments for pneumonia and diarrhoea. The smallest experiment (Experiment 1) in pneumonia analysis in this thesis had at least twice the sample size of the largest RCT used for examining efficacy of antibiotics in treatment of indrawing pneumonia (see sub – section 1.1.4 of Chapter 1 for a summary of these studies – mostly conducted in single sites). Also the Zinc effectiveness analysis had a sample size 75 times that of the largest Zinc trial that had been conducted in

Africa. Observational datasets also include patients that may not qualify for enrolment in a typical explanatory randomised controlled trial – and therefore perhaps provide more true to life estimates of treatment effects similar to those observed in highly pragmatic trials (176). However, as most observational datasets are not meant for research, they have the key analytic challenge of non – random allocation of treatments. In my work despite extensive efforts to use rigorous methods to overcome this challenge, suggestions that bias persisted remained. First, the sensitivity analysis conducted in the antibiotic treatment of pneumonia gave the impression of bias elimination but in all analyses the direction of effect tended to favour the penicillin group over the penicillin plus gentamicin group, perhaps best illustrated by the drift towards the null with the use of increasingly stringent PS trimming. Second, the use of instrumental variable analysis provided consistent results (as also observed in the analysis of alternative antibiotic treatment) as those derived using the propensity score methods in examining effectiveness of Zinc treatment – which resulted in an impression of no influence of unobserved confounders. The analysis of time to experiencing inpatient mortality, however, suggested persistent hidden bias in treatment allocation in which the Zinc treatment was potentially assigned to children who were less likely to die. This indicates that finding the most valid instrument to completely eliminate unobservable bias may be a challenge in analyses using routine datasets.

In as much as propensity score methods may try to eliminate covariate distribution imbalance between treatment groups, it may be difficult to eliminate

the influence due to unobserved confounders. For instance, observational datasets do not capture the aspect of gut feeling in the process of diagnosis and treatment. Gut feeling has been shown to play a significant role in diagnosis and treatment assignment processes (55, 57). For example, as also discussed in Chapter 1 sub – section 1.2.5.2, an observational study by Van den Bruel (2010) demonstrated that gut feeling (which also presumably represented a summary of formally recorded clinical signs) had greater diagnostic value than signs/symptoms alone in determining the risk of serious infection (55). This was also the case in a study by Zachariasse (2017) (177) in which clinical impression by nurses in the initial assessment (in addition to other clinical signs) was a good predictor of admission. The use of small panels of clinical features that make up guidelines (which do not allow the use of gut feeling) cannot comprehensively define a patient's condition. And thus allocation to different treatments could be influenced by many explicitly and implicitly observed factors, with the potential result of continued susceptibility to bias when comparing groups allocated alternative treatments as part of routine care (even if patients have equivalent PS values based on observed variables).

There is also the more practical challenge of what clinicians are likely or willing to document especially the junior clinicians. In an effort to capture as much detail as possible, junior clinicians in the CIN are encouraged to use aides like the paediatric admission record forms (PAR – see appendix A.2), which has previously resulted in improved documentation as demonstrated in Gachau (2017) and Tuti (2017) (166, 178), to capture as much data as possible. However,

we may not be quite certain about the accuracy of documentation by the clinicians as there would likely be inter-observer differences in assessment of clinical signs which might result in misclassifications – which may bias results in one direction if systematic or may simply increase noise in the system if random. Another challenge in using observational data (that also applies to highly pragmatic trials) is that it is difficult to be certain what treatments are actually given over the course of an admission. In the data I used the focus was only on admission clinical features and outcomes, although data are available on treatment changes (as illustrated in Chapter 3 **Figure 3.2.1**).

6.7. Further research

There are continued concerns of high mortality in children with indrawing pneumonia (171) with analyses in this thesis supporting this observation and pointing to the lack of advantage of using broader spectrum antibiotics in a population that current WHO recommendations indicate can be treated as outpatients. What may therefore be useful is to develop simple prognostic model(s) that will help in determining those children who are at higher risk who should be admitted to hospital amongst the group of pneumonia children not meeting current (post-2013) criteria for a diagnosis of severe pneumonia. This may mean revising guidelines that can still provide an important mechanism to classify illness severity for use by the junior clinicians or non-specialist health workers who provide the majority of care in LMIC. A study by Tuti (2017) (178) provides an example of some possible additional characteristics that could be examined to inform admission decisions for this group of patients. However, an

important further step would ideally be to test any revisions to guideline recommendations prospectively to establish both feasibility and benefit. The PERCH studies while useful in identifying common pathogens have not really helped advance the approaches to treatment to reduce mortality – other than indicating that vaccines especially against Respiratory Syncytial Virus (RSV) may be of benefit. So it would be valuable to: (i) do detailed case reviews of all events to examine if there were modifiable factors in the whole care process that if addressed might have averted mortality or; (ii) examine the effectiveness of other interventions such as better supportive care, better monitoring for deterioration, and potentially broader spectrum treatments among others. Lastly perhaps it might be useful do post-mortem studies in children with non-severe pneumonia who die to try and establish whether there was an alternative cause of death.

The safety of Zinc has been demonstrated in prior trials and may be effective for treatment of children with diarrhoea admissions in Kenya based on analyses presented in this thesis. However, the possible residual bias results in uncertainty about the true effectiveness. And although the analysis attempted to avoid the need for a RCT, such trials are probably needed as the best source of evidence providing high certainty of effect estimates.

6.8. Recommendations

It is important to note that observational data may still be used in situations where randomisation may be infeasible or unethical. For example, it may be unethical to expose patients to treatments practically known to be inferior (179).

The analysis of such data may be through the use of PS methods – as these PS methods seemed to better reduce imbalance between groups than the use of regression alone without PS adjustments (which resulted in more biased estimates) in the two questions tackled. In general cases, observational datasets are obviously still useful as a means to determine effectiveness and thus as a means to establish the consequences of a policy rather than perhaps being the primary source to determine policy. For example, in this thesis the high mortality observed for children with indrawing pneumonia argues for the reassessment of a policy based on efficacy trials. Without such findings assessing outcomes in routine settings policies may not be questioned.

In cases where randomisation is permissible, pragmatic randomised trials would probably be the best means to examine treatment effectiveness. The Pragmatic Explanatory Continuum Indicator Summary 2 (PRECIS-2) tool (180) explains nine characteristics of a robust pragmatic trial. See **Box 6.8.2** for a summary of these characteristics.

Box 6.8.2: Summary characteristics of a pragmatic trial (PRECIS – 2)

- Eligibility— The patients eligible should be similar to those who would receive treatment in the actual care settings.
- Recruitment— patient recruitment should require no extra effort as what would be used in usual care to attend to patients.
- Setting—The setting should be as similar as possible to those where usual care happens. For instance, multiple clinics to enroll patients with diverse characteristics.
- Organisation—The resources and expertise needed to deliver treatments should be close to what is used in actual practice.
- Flexibility (delivery)— The delivery of treatments should be as flexible as would be delivered in routine care.
- Flexibility (adherence)—The monitoring of patients and adherence to treatments should be similar to that in usual care.
- Follow-up—The intensity of outcome measurement and follow up should be close to that practised in routine care.
- Primary outcome— The outcome should be relevant to the patients.
- Primary analysis— All the necessary data should be available for the outcome analysis.

These randomised trials should be cheaply conducted with a large number of patients recruited and ideally employing outcomes that are easily measurable. One variety of pragmatic trial is the registry trial (181). In this approach randomised trials may be embedded in existing prospective patient registries – and such platforms may save on costs and time especially if they can be reused for many trials. The key concern in such trials is the quality of data obtained from the registry platforms. The Clinical Trials Transformation Initiative (CTTI) (182) consulted experts on how best clinical trials would be executed on registry platforms – and one of the key areas of guidance they developed was to help determine whether data generated from the registries were relevant, robust and reliable. Further guidance on data quality and how to conduct these trials has also been provided in the PRECIS – 2 tool (180). These trials should possess all

the ethics characteristics of the explanatory trials – in which consent should be sought from the participants and research ethics review boards, though this is debated as it is a major source of costs (180) and some suggest that waivers may be granted if there is a true equipoise since these trials typically test existing treatments already in routine use. An example of a trial that was executed using a registry platform is the TASTE trial by Fröbert (2013) (183) which examined the effect of thrombus aspiration in patients with ST-Segment Elevation Myocardial Infarction. This trial enrolled a total of 7244 patients from the Swedish Coronary Angiography and Angioplasty Registry with the primary outcome being all-cause mortality at 30 days. As previous RCT meta - analyses in this field showed conflicting findings, treatment policy frameworks may be informed by this large pragmatic trial. Another example that was attempted in the CIN is the trial by Ayieko (2018, submitted) which examined if clinicians adhered to the new clinical guidelines of treating patient with pneumonia (severity classification based on the new pneumonia guidelines). These trials may be extended in the future in CIN.

6.9. Conclusion

The analysis in this thesis demonstrated that treatment effectiveness evaluation using routine data risks bias potentially due to unobserved confounders. And analyses of such datasets using propensity scores may only be useful when randomised controlled trials are not feasible. The use of PS has a greater potential of minimising imbalances as compared to the use of regression methods alone. However, it is probably better to pursue pragmatic trials (where

randomisation is feasible) than rely on routine datasets for comparative effectiveness evaluations.

8. References

1. Levels & Trends in child mortality [Internet].: UNICEF; 2014 []. Available from: http://www.data.unicef.org/fckimages/uploads/1410869227_Child_Mortality_Report_2014.pdf.

2. Committing to Child Survival: A Promise Renewed [Internet].; 2014 []. Available from: http://files.unicef.org/publications/files/APR_2014_web_15Sept14.pdf.

3. MDG4: reduce child mortality [Internet].: WHO; 2015 []. Available from: http://www.who.int/topics/millennium_development_goals/child_mortality/en/.

4. Global Health Observatory data repository [Internet].; 2016 []. Available from: http://apps.who.int/gho/data/node.main.686?lang=en.

5. Kenya Demographic and Health Survey [Internet].; 2014 []. Available from: http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf.

6. Kenya reports on situation of children and women on eve of Day of the African Child 2010 celebrations [Internet].; 2010 []. Available from: https://www.unicef.org/kenya/health_5736.html.

7. Pneumonia and Diarrhoea Progress Report 2014 [Internet].: John Hopkins Bloomberg School of Public Health; 2014 []. Available from: http://www.jhsph.edu/research/centers-and-institutes/ivac/resources/IVAC-2014-Pneumonia-Diarrhea-Progress-Report.pdf;.

8. Wandera EA, Mohammad S, Ouko JO, Yatitch J, Taniguchi K, Ichinose Y. Variation in rotavirus vaccine coverage by sub-counties in Kenya . Tropical Medicine and Health. 2017;45(9).

9. Keats EC, Ngugi A, Macharia W, Akseer N, Khaemba EN, Bhatti Z, et al. Progress and priorities for reproductive, maternal, newborn, and child health in Kenya: a Countdown to 2015 country case study. The Lancet Global Health. 2017 08/01;5(8):e782-95.

10. English M, Scott A. What Is The Future For Global Case Management Guidelines for Common Childhood Diseases? . 2008;5(12):e241.

11. Ye C, Beyene J, Browne G, Thabane L. Estimating treatment effects in randomised controlled trials with non-compliance: a simulation study. BMJ Open Research Methods. 2014;4(6):e005362.

12. Kausto J, Solovieva S, Virta LJ, Viikari-Juntura E. Partial sick leave associated with disability pension:propensity score approach in a register-based cohort study. BMJ Open Public Health. 2012;2(6):e001752.

13. Viera AJ, Bangdiwala SI. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. Family Medicine. 2007;39(2):132-7.

14. Silverman S, Miller P, Sebba A, Weitz M, Wan X, Alam J, et al. The direct assessment of nonvertebral fractures in community experience (DANCE) study: 2-year nonvertebral fragility fracture results. Osteoporos Int. 2013;24(8):2309-17.

15. Berkley JA, Lowe BS, Mwangi I, Williams T, Bauni E, Mwarumba S, et al. Bacteremia among Children Admitted to a Rural Hospital in Kenya. The new england journal of medicine. 2005;352(1):39-47.

16. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?". The Lancet. 2005;365(9453):82-93.

17. Hanney SR, Castle-Clarke S, Grant J, Guthrie S, Henshall C, Mestre-Ferrandiz J, et al. How long does biomedical research take? Studying the time taken between biomedical and health research and its translation into products,policy, and practice. Health Research Policy and Systems. 2015;13(1).

18. Olayemi E, Asare EV, Benneh-Akwasi Kuma AA. Guidelines in lower-middle income countries. Br J Haematol. 2017;177(6):846-54.

19. Mulholland K, Carlin JB, Duke T, Weber M. Challenges of trials of antibiotics for pneumonia in low-income countries. The Lancet Respiratory Medicine. 2014;2(12):952-4.

20. Lassi ZS, Das JK, Haider SW, Salam RA, Qazi SA, Bhutta ZA. Systematic review on antibiotic therapy for pneumonia in children between 2 and 59 months of age Archives of Disease in Childhood. 2014;99:687-93.

21. Lodha R, Kabra SK, Pandey RM. Antibiotics for community-acquired pneumonia in children 2013. Cochrane Database Syst Rev. 2013(6).

22. Hospital Care for Children [Internet].; 2005 []. Available from: http://apps.who.int/iris/bitstream/10665/43206/1/9241546700.pdf.

23. Gilani Z, Kwong YD, Levine OS, Deloria-Knoll M, Scott JAG, O'Brien KL, et al. A Literature Review and Survey of Childhood Pneumonia Etiology Studies: 2000–2010. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 54(Suppl 2). 2012:S102-8.

24. Hospital Care for Children [Internet].; 2013 []. Available from: http://apps.who.int/iris/bitstream/10665/81170/1/9789241548373_eng.pdf.

25. Lazzerini M, Wanzira H. Oral zinc for treating diarrhoea in children. Cochrane Database of Systematic Reviews. 2016(12).

26. Patel A, Mamtani M, Dibley MJ, Badhoniya N, Kulkarni H. Therapeutic Value of Zinc Supplementation in Acute and Persistent Diarrhea: A Systematic Review . PLoS ONE. 2010;5(4).

27. Basic paediatric protocols [Internet].: Kenyan Ministry of Health; 2016 []. Available from: https://www.tropicalmedicine.ox.ac.uk/_asset/file/basic-paediatric-protocols-2016.pdf.

28. Basic Paediatric Protocols: November 2013 Edition [Internet].: Ministry of Health; 2013 []. Available from: http://www.idoc-africa.org/index.php/86-clinical-guide/ken-guide/134-basic-paediatric-protocols-november-2013-edition.

29. Simoes EAF, Cherian T, Chow J, Shahid-Salles SA, Laxminarayan R, John TJ. Acute Respiratory Infections in Children. In: Disease Control Priorities in Developing Countries. 2nd ed. Washington (DC): World Bank: ; 2006. p. 483-97.

30. Pulmonary Consolidation [Internet]. Wikipedia; 2014 []. Available from: https://en.wikipedia.org/wiki/Pulmonary_consolidation.

31. Agweyu A, Gathara D, Oliwa J, Muinga N, Edwards T, Allen E, et al. Oral amoxicillin versus benzyl penicillin for severe pneumonia among kenyan children: a pragmatic randomised controlled noninferiority trial. Clin Infect Dis. 2015;60(8):1216-24.

32. Enarson PM, Gie RP, Mwansambo CC, Maganga ER, Lombard CJ, Enarson DA, et al. Reducing Deaths from Severe Pneumonia in Children in Malawi by Improving Delivery of Pneumonia Case Management. PLoS ONE. 2014;9(7):e102955.

33. Hospital Care for Children [Internet].; 2013 []. Available from: http://apps.who.int/iris/bitstream/10665/81170/1/9789241548373_eng.pdf.

34. Mullei K, Wafula F, Goodman C. A Case Study of Integrated Management of Childhood Illness (IMCI) Implementation in Kenya. Kenya Medical Research Institute; 2008.

35. Song JW, Chung KC. Observational Studies: Cohort and Case-Control Studies. Plastic and Reconstructive Surgery. 2010;126(6):2234-42.

36. Lindsay R. Beyond clinical trials: The importance of large databases in evaluating differences in the effectiveness of biphosphonate in postmenopausal osteoporosis. Bone. 2007;40(5):S32-5.

37. Stuart EA. Matching methods for causal inference:A review and a look forward. Statist. Sci. 2010;25(1):1-21.

38. Rosenbaum PR. Design of Observational Studies. Springer Series in Statistics; 2009.

39. Rubin D.B. For Objective Causal Inference, Design Trumps Analysis. The Annals of Applied Statistics. 2008;2(3):808-40.

40. Draft recommendation for management of children with severe febrile illness and impared circulation without signs of severly impaired circulation: Guideline Panel Meeting [Internet].: Health Services, Implementation Research and Clinical Excellence Collaboration; 2013 []. Available from: http://www.idoc-africa.org/images/documents/guidelines/Panel%20recommendation%20on%20fluid%20resuscitation%20I%20final.pdf.

41. Draft recommendations for care of umbilical cord in newborns: Guideline Panel Meeting. [Internet].: Health Services, Implementation Research and Clinical Excellence Collaboration; 2013 []. Available from: http://www.idoc-africa.org/images/documents/guidelines/Panel%20recommendations%20on%20cord%20care%20final.pdf.

42. Draft recommendations for hospital management of children (less than 5 years) with sickle cell disease: Guideline Panel Meeting 2013. [Internet].: Health Services, Implementation Research and Clinical Excellence Collaboration; 2013 []. Available from: http://www.idoc-africa.org/images/documents/guidelines/Panel%20recommendations%20on%20sickle%20cell%20anaemia%20final.pdf.

43. Agweyu A, Kibore M, Digolo L, Kosgei C, Maina V, Mugane S, et al. Prevalence and correlates of treatment failure among Kenyan children hospitalised with severe community-acquired pneumonia: a prospective study of the clinical effectiveness of WHO pneumonia case management guidelines. Tropical Medicine & International Health. 2014;19(11):1310-20.

44. Rosenbaum PR, Silber JH. Matching and thick description in an observational study of mortality after surgery. Biostatistics. 2001;2(2):217-32.

45. Thoemmes FJ, Kim ES. A Systematic Review of Propensity Score Methods in the Social Sciences . Multivariate Behavioral Research. 2011;46(1):90-118.

46. Zakrison TL, Austin PC, McCredie VA. A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines. Eur J Trauma Emerg Surg. 2017.

47. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. JSTOR. 1983;70:41-55.

48. Wamai R. The Kenya Health System—analysis of the situation and enduring challenges . Japan Medical Association. 2009;52:134-40.

49. Kenya - Kenya Service Availability and Readiness Assessment Mapping (SARAM) report, 2013 [Internet].; 2013 []. Available from: http://apps.who.int/healthinfo/systems/datacatalog/index.php/catalog/42.

50. Kenya Health Sector Referral Implementation Guidelines [Internet].; 2014 []. Available from: https://www.medbox.org/kenya/kenya-health-sector-referral-implementation-guidelines/preview?q=.

51. Kihuba E, Gathara D, Mwinga S, Mulaku M, Kosgei R, Mogoa W, et al. Assessing the ability of health information systems in hospitals to support evidence-informed decisions in Kenya. Global Health Action. 2014;7.

52. Muinga N, Ayieko P, Opondo C, Ntoburi S, Todd J, Allen E, et al. Using Health worker opinion to assess changes in structural components of quality in a cluster randomised trial. BMC Health Services Research. 2014.

53. Gathara D, Malla L, Ayieko P, Karuri S, Allen E, Irimu G, et al. Variation in and risk factors for paediatric inpatient all-cause mortality in a low income setting: Data from an emerging clinical information network. 2017;17(1):99.

54. English M, Esamai E, Wasunna A, Were F, Ogutu B, Wamae A, et al. Delivery of paediatric care at the first-referral level in Kenya. The Lancet. 2004;364:1622-9.

55. Stopler E, Van Roeyn P, Van de Wiel M, Van Bokhoven M, Houben P, Van der Weijde T, et al. Consensus on gut feelings in general practice. BMC Family Practice. 2009.

56. Stopler E, Van Roeyn P, Van de Wiel M, Van Bokhoven M, Houben P, Van der Weijde T, et al. The diagnostic role of gut feelings in general practice a focus group study of the concept and its determinants. BMC Family Practice. 2009;10:17.

57. Van den Bruel A, Thompson M, Buntinx F. Clinicians' gut feeling about serious infections in children: observational study. BMJ Open. 2012;345:e6144.

58. Shann F, Barker J, Poore P. Clinical signs that predict death in children with severe pneumonia. Pediatr Infect Dis. 1989;8(12):852-5.

59. Rubin DB. An overview of Multiple Imputation. In Proceedings of the Survey Research Section, American Statistical Association. 1988:79-84.

60. Jan P Vandenbroucke JP, Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. PLoS Medicine. 2007;4(10):e297.

61. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat Methods Med Res. 2014 01 Oct 2014;23(5):440-59.

62. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011;46(3):399-424.

63. Heckman J, Ichimura H, Todd P. Matching as an econometric evaluation estimator. The Review of Economic Studies. 1998;65:261-94.

64. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stats Med. 2013;33(6):1057-69.

65. Gu X, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances and algorthms. Journal of Computational and Graphical Statistics. 1993;2:405-20.

66. Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity Score Matching and Subclassification in observational Studies with Multilevel Treatments. Journal of the International Biometric Society. 2016;72(4):1055-65.

67. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software. 2011;42(8):1-28.

68. Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity matching. American Journal of Epidemiology. 2013.

69. Jasjeet SS. The Neyman Rubin Model for Causal Inference and Estimation via Matching Methods. The Oxford Handbook of Political Methodology. 2007.

70. Pirracchio R, Carone M, Resche Rigon M, Caruana E, Mebazaa A, Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. Stat Methods Med Res. 2013;25(5):1938-54.

71. Szafara KL, Kruse RL, Mehr DR, Ribbe MW, van der Steen JT. Mortality Following Nursing Home–Acquired Lower Respiratory Infection: LRI Severity, Antibiotic Treatment, and Water Intake. Journal of the American Medical Directors Association. 2012 May 2012;13(4):376-83.

72. Austin PC, Stuart EA. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. Statistical methods in medical research. 2015;26(6):2505-25.

73. Rosenbaum PR. Sensitivity Analysis in Observational Studies. Encyclopedia of Statistics in Behavioral Science. 2005;4:1809-14.

74. DiPrete T, Gangl M. Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments . Sociological Methodology. 2004;34:271-310.

75. Baiocchi M, Cheng J, Small DS. Tutorial in Biostatistics: Instrumental Variable Methods for Causal Inference. Statistics in Medicine. 2014;33(13):2297-340.

76. An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data [Internet].; 2010 []. Available from: http://www.personal.psu.edu/ljk20/rbounds%20vignette.pdf.

77. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med. 2014 15 Jun 2014;33(13):2297-340.

78. Toh S, Garcia Rodriguez LA, Hernan MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. Pharmacoepidemiol Drug Saf. 2012 May 2012;21(SUPPL.2):13-20.

79. Lee KJ, Simpson JA. Introduction to multiple imputation for dealing with missing data. Respirology. 2014 February 2014;19(2):162-7.

80. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: A weighting approach. Stat Methods Med Res. 2007 June 2007;16(3):259-75.

81. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Statistical Methods in Medical Research. 2016;25(1):188-204.

82. Nakai M, Ke W. Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. 2011;5(1):1-13.

83. Yao XI W, X., Speicher PJ, Hwang ES, Cheng P, Harpole DH, Berry MF, et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. J Natl Cancer Inst. 2017;109(8).

84. Kang H. The prevention and handling of the missing data. Springerplus. 2013;64(5):402-6.

85. Fielding S, Maclennan G, Cook JA, Ramsay CR. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. Trials. 2008;9:51.

86. Bell LM, Fiero M, Horton NJ, Hsu C. Handling missing data in RCTs; a review of the top medical journals. BMC Medical Research Methodology. 2014;14:118.

87. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson DR. A review of the reporting and handling of missing data in cohort studies with repeated assessment of repeated measures. BMC Medical Research Methodology. 2012;12:96.

88. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology. 2012;23(5):729-32.

89. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276-82.

90. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement . J Clin Epidemiol. 2009;62(10):1006-12.

91. Why Propensity Scores Should Not Be Used for Matching [Internet].; 2015 [cited August 3, 2017]. Available from: http://j.mp/2ovYGsW.

92. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat.Med. 2007;27:2037-49.

93. Molenberghs G, Beunckens C, Sotto C. Every missing not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society. 2008;70(2):371-88.

94. Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, et al. Analysing incomplete longitudinal clinical trial data. Biostatistics. 2004;5(3):445-64.

95. Sazawal S, Black R. Pneumonia Case Management Trials Group. Lancet Infect Dis. 2003;3(9):547-56.

96. Ayieko P, Ogero M, Makone B, Julius T, Mbevi G, Nyachiro W, et al. Characteristics of admissions and variations in the use of basic investigations, treatments and outcomes in Kenyan hospitals within a new Clinical Information Network. Archives of Disease in Childhood. 2015;101(3):223-9.

97. Agweyu A, Opiyo N, English M. Experience developing national evidence-based clinical guidelines for childhood pneumonia in a low-income setting - making the GRADE? BMC Pediatric. 2012;12:1.

98. Ayieko P, English M. Case Management of Childhood Pneumonia in Developing Countries. Pediatr Infect Dis J. 2007;26:432-40.

99. Graham SM, English M, Hazir T, Enarson P, Duke T. Challenges to improving case management of childhood pneumonia at health facilities in resource limited settings. Bulletin of the WHO. 2015;86(5):349-55.

100. Atkinson M, Lakhanpaul M, Smyth A, Vyas H, Weston V, Sithole J, et al. Comparison of oral amoxicillin and intravenous benzyl penicillin for community acquired pneumonia in children (PIVOT trial): a multicentre pragmatic randomised controlled equivalence trial. Thorax. 2007;61(12):1102-6.

101. Hazir T, Fox LM, Nisar YB, Fox MP, Ashraf YP, MacLeod WB, et al. Ambulatory short-course high-dose oral amoxicillin for treatment of severe pneumonia in children: a randomised equivalency trial. Lancet. 2008;371:49-56.

102. Addo-Yobo E, Chisaka N, Hassan M, Hibberd P, Lozano JM, Jeena P, et al. Oral amoxicillin versus injectable penicillin for severe pneumonia in children aged 3 to 59 months: a randomised multicentre equivalency study. Lancet. 2004;364(9440):1141-8.

103. Addo-Yobo E, Anh DD, El-Sayed HF, Fox LM, Fox MP, MacLeod W, et al. Outpatient treatment of children with severe pneumonia with oral amoxicillin in four countries: the MASS study. Tropical Medicine and International Health. 2011;16(8):995-1006.

104. Onyango D, Kikuvi G, Amukoye E, Omolo J. Risk factors of severe pneumonia among children aged 2-59 months in western Kenya: a case control study. The Pan African Medical Journal. 2012(13):45.

105. Wittes J. Sample size calculations for randomised controlled trials. Epidemiologic Reviews. 2002;24(1).

106. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;20(1):40-9.

107. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiol Drug Saf. 2008 2008;17(12):1218-25.

108. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009 10 Nov 2009;28(25):3083-107.

109. Fine J, Gray R. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.

110. UNICEF DATA: Monioring the situation of women and children [Internet].; 2017 []. Available from: https://data.unicef.org/topic/child-health/pneumonia/.

111. Koh JW, Wong JJ, Sultana R, Wong PP, Mok YH, Lee JH. Risk factors for mortality in children with pneumonia admitted to the pediatric intensive care unit. Pediatric pulmonology. 2017;52(8):1076-84.

112. Malla L, Perera R, McFadden E, English M. Comparative effectiveness of injectable penicillin versus a combination of penicillin and gentamicin in children with pneumonia characterised by indrawing in Kenya: A protocol for an observational study. BMJ Open. 2017;7(9):e016784.

113. Danaei G, Rodríguez LA, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. Stat.Methods Med.Res. 2011;22(1):70-96.

114. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99.

115. Hasali MAA, Ibrahim MIM, Sulaiman SAS, Ahmad Z, Hasali JBA. A clinical and economic study of community-acquired pneumonia between single versus combination therapy. Pharm World Sci. 2005;27(3):249-53.

116. Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res. 2017 06/02; 2017/11:0962280217713032.

117. Hill J. Reducing Bias in Treatment Effect Estimation in Observational Studies Suffering from Missing Data. ISERP working paper 04-01. 2004.

118. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat Methods Med Res. 2014 01 Oct 2014;23(5):440-59.

119. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 British Medical Journal Publishing Group;338.

120. Bacterial pneumonia [Internet].; 2012 []. Available from: https://en.wikipedia.org/wiki/Bacterial_pneumonia.

121. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment Effects in the presence of Unmeasured Confounding: Dealing with observations in the Tails of the Propensity Score Distribution - A Simulation Study. American Journal of Epidemiology. 2010;172(7):843-54.

122. Klungel OH, Jamal U.M., De Boer A., Belitser SV, Groenwold RH, Roes KC. Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods. Pharm Anal Acta. 2015;6:353.

123. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation Analysis. Annu Rev Psychol. 2007;58:593.

124. Berkley JA, Brent A, Mwangi I, English M, Maitland K, Marsh K, et al. Mortality among Kenyan children admitted to a rural district hospital on weekends as compared with weekdays . Pediatrics. 2004;114:1737-8.

125. Bell CM, Redelmeier DA. Mortality among patients admitted to hospitals on weekends as compared with weekdays. N Engl J Med. 2001;345(663):668.

126. Meacock R, Anselmi L, Kristensen S, Doran T, Sutton M. Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission. J Health Serv Res Policy. 2017 01/01; 2017/11;22(1):12-9.

127. Freemantle N, Ray D, McNulty D, Rosser D, Bennett S, Keogh BE, et al. Increased mortality associated with weekend hospital admission: a case for expanded seven day services?

. BMJ. 2015;4596:1-6.

128. Aldridge C, Bion J, Boyal A, Chen Y, Clancy M, Evans T, et al. Weekend specialist intensity and admission mortality in acute hospital trusts in England: a cross-sectional study. The Lancet 2017/11;388(10040):178-86.

129. Bray BD, Steventon A. What have we learnt after 15 years of research into the 'weekend effect'? BMJ Qual Saf. 2017 BMJ Publishing Group Ltd;26(8):607-10.

130. Marra G, Papageorgiou G, Radice R. Estimation of a semiparametric recursive bivariate probit model with nonparametric mixing. Aust N Z J Stat. 2014;55(3):321-42.

131. Ieva F, Marra G, Paganoni AM, Radice R. A Semiparametric Bivariate Probit Model for Joint Modeling of Outcomes in STEMI Patients. Computational and Mathematical Methods in Medicine. 2014;2014:7.

132. Probit regression [Internet].; 2017 [cited 22-03-2017]. Available from: http://stats.idre.ucla.edu/stata/output/probit-regression/.

133. Amemiya T. Qualitative Response Models: A Survey. Journal of Economic Literature. 1981;XIX:1483-536.

134. Zhang J, Yu KF. What's the Relative Risk? A method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. JAMA - Journal of the American Medical Association. 1998;280(19):1690-1.

135. Janot A, Vandanjon PO, Gautier M. A Durbin-Wu-Hausman test for industrial robots identification. ; 2013.

136. Lazzerini M, Ronfani L. Oral zinc for treating diarrhoea in children . Cochrane Database of Systematic Reviews. 2012;6.

137. Fischer Walker CL, Bhutta ZA, Bhandari N, Teka T, Shahid F, Taneja S, et al. Zinc Supplementation for the Treatment of Diarrhea in Infants in Pakistan, India and Ethiopia. Journal of pediatric gastroenterology and nutrition. 2006;43(3):357-63.

138. Roy SK, Tomkins AM, Akramuzzaman SM, Behrens RH, Haider R, Mahalanabis D. Randomised controlled trial of zinc supplementation in malnourished Bangladeshi children with acute diarrhoea. Arch Dis Child. 1997;77:196-200.

139. Sazawal S, Black RE, Bhan MK, Bhandari N, Sinha A, Jalla S. Zinc supplementation in young children with acute diarrhea in India. N Eng J Med. 1995;333:839-44.

140. Wessells KR, Brown KH. Estimating the Global Prevalence of Zinc Deficiency: Results Based on Zinc Availability in National Food Supplies and the Prevalence of Stunting. ed. PLoS ONE. 2012;7(11).

141. Child growth standards [Internet].; 2011 []. Available from: http://www.who.int/childgrowth/software/en/.

142. Exuzides A, Colby C. An application of imputation techniques to improve data availability from electronic medical records. Value in Health. 2010. November 2010;13(7):A368.

143. VanderWeele TJ, Knol MJ. A Tutorial on Interaction. Epidemiol. Methods. 2014;3(1):33-72.

144. Spreeuwenberg MD, Bartak A, Croon MA, Hagenaars JA, Busschbach JJ, Andrea H, et al. The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. Med. Care. 2010;48:166-74.

145. Scheike TH, Zhang M-. Analyzing Competing Risk Data Using the R timereg Package. Journal of Statistical Software. 2011;38(2):i02.

146. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med. 2014 15 Jun 2014;33(13):2297-340.

147. TchetgenTchetgen EJ, Walter S, Vansteelandt S, Martinussen T, Glymour M. Instrumental variable estimation in a survival context. , *26*(3), 402–410. . Epidemiology (Cambridge, Mass.). 2015;26(3):402-10.

148. Cook T.D., Steiner PM. Case Matching and the Reduction of Selection Bias in Quasi-Experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data Analysis. Psychol Methods. 2010 March 2010;15(1):56-68.

149. Suh HS, Hay JW, Johnson KA, Doctor JN. Comparative effectiveness of statin plus fibrate combination therapy and statin monotherapy in patients with type 2 diabetes: Use of propensity-score and instrumental variable methods to adjust for treatment-selection bias. Pharmacoepidemiol Drug Saf. 2012 May 2012;21(5):470-84.

150. Larson CP, Roy SK, Khan AI, Rahman AS, Qadri F. Zinc Treatment to Under-five Children: Applications to Improve Child Survival and Reduce Burden of Disease. J HEALTH POPUL NUTR. 2008;26(3):356-65.

151. Mayo-Wilson E, Imdad A, Junior J, Dean S, Bhutta ZA. Preventive zinc supplementation for children, and the effect of additional iron: a systematic review and meta-analysis. BMJ Open. 2014;4(6):e004647.

152. Clinical Practice Research Datalink (CPRD) [Internet].; 2015 [cited 04/01/2016]. Available from: http://www.cprd.com/home/.

153. Strandberg-Larsen M, Schiøtz ML, Frølich A. Kaiser Permanente revisited - can European health care systems learn? . Eurohealth. 2007;13(4):24-6.

154. Suissa S, Dell'Aniello S, Ernst P. Long-Acting Bronchodilator Initiation in COPD and the Risk of Adverse Cardiopulmonary Events. Chest 2018/01;151(1):60-7.

155. Gamble J, Thomas JM, Twells LK, Midodzi WK, Majumdar SR. Comparative effectiveness of incretin-based therapies and the risk of death and cardiovascular events in 38,233 metformin monotherapy users. Medicine. 2016 05/29;95(26):e3995.

156. Hennessy S, Bilker WB, Knauss JS, Kimmel SE, Margolis DJ, Morrison MF, et al. Comparative cardiac safety of low-dose thioridazine and low-dose haloperidol. Br J Clin Pharmacol. 2003 11/17;58(1):81-7.

157. McEwan P, Evans ML, Nyeland ME, Skovgaard R, Richards A, Bergan EQ, et al. Impact of UK guidelines on clinical prescribing in patients with Type 2 diabetes: A comparative effectiveness analysis of liraglutide vs sitagliptin in the UK. Diabetic Medicine. 2014;31:174.

158. Nyeland ME, Ploug UJ, Skovgaard R, Richards A, Bergan EQ, Zammit DC, et al. Comparative Effectiveness of Liraglutide Versus Sitagliptin in Type 2 Diabetes in the United Kingdom: A Retrospective Study in Primary Care. Value in Health 2018/01;16(7):A431.

159. Allen C, Meeraus W, Donegan K. Comparative Risk Of All-Cause Mortality In Older Patients Prescribed Codeine Or Tramadol For Non-Malignant Pain: Retrospective Cohort Study. Pharmacoepidemiol Drug Saf. 2016;9:25.

160. Allen C, Meeraus W, Donegan K. The comparative risk of all-cause mortality in older patients prescribed opioids for non-malignant pain: A retrospective observational cohort study . Drug Safety. 2015;38(10):962-3.

161. Slaymaker E, McLean E, Wringe A, Calvert C, Marston M, Reniers G, et al. The Network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA): Data on mortality, by HIV status and stage on the HIV care continuum, among the general population in seven longitudinal studies between 1989 and 2014 version 1; referees: 2 approved, 1 approved with reservations]. Gates Open Research. 2017;1(4).

162. Duke T, Yano E, Hutchinson A, Hwaihwanje I, Aipit J, Tovilu M, et al. Large-scale data reporting of paediatric morbidity and mortality in developing countries: it can be done. Arch Dis Child. 2015 BMJ Publishing Group Ltd and Royal College of Paediatrics and Child Health.

163. Hersh W. Health care information technology: Progress and barriers. JAMA. 2004 11/10;292(18):2273-4.

164. Ayieko P, Ntoburi S, Wagai J, Opondo C, Opiyo N, Migiro S, et al. A Multifaceted Intervention to Implement Guidelines and Improve Admission Paediatric Care in Kenyan District Hospitals: A Cluster Randomised Trial. PLoS Medicine. 2011;8(4).

165. Tuti T, Bitok M, Malla L, Paton C, Muinga N, Gathara D, et al. Improving documentation of clinical care within a clinical information network: an essential initial step in efforts to understand and improve care in Kenyan hospitals. BMJ Global Health. 2016;1(1):e000028.

166. Gachau S, Ayieko P, Gathara D, Mwaniki P, Ogero M, Akech S, et al. Does audit and feedback improve the adoption of recommended practices? Evidence from a longitudinal observational study of an emerging clinical network in Kenya. BMJ Global Health. 2017 12/01;2(4).

167. Stuart EA, DuGoff E, Abrams M, Salkever D. Estimating Causal Effects in Observational Studies Using Electronic Health Data: Challenges and (some) Solutions. eGEMS. 2013.

168. Sheetz KH, Dimick JB. Comparative effectiveness of laparoscopic vs. Open colectomy: An instrumental variable analysis. J Surg Res. 2014. February 2014;186(2):534.

169. Alawadi Z, Phatak U, Hu C, Bailey C, Kao L, You Y, et al. Comparative effectiveness of primary tumor resection in metastatic colon cancer: An instrumental variable analysis. Annals of Surgical Oncology.Conference: 68th Annual Cancer Symposium of the Society of Surgical Oncology Houston, TX United States.Conference Start: 20150325 Conference End: 20150328.Conference Publication: (var.pagings). 2015 February 2015;22(1 SUPPL. 1):S9.

170. Suh HS, Hay JW, Johnson KA, Doctor JN. Comparative effectiveness of statin plus fibrate combination therapy and statin monotherapy in patients with type 2 diabetes: Use of propensity-score and instrumental variable methods to adjust for treatment-selection bias. Pharmacoepidemiol Drug Saf. 2012 May 2012;21(5):470-84.

171. Agweyu A, Oliwa J, Gathara D, Muinga N, Allen E, Lilford RJ, et al. Comparable outcomes among trial and nontrial participants in a clinical trial of antibiotics for childhood pneumonia: a retrospective cohort study. J Clin Epidemiol 2018/01;94:1-7.

172. Hammitt LL, Kazungu S, Morpeth SC, Gibson DG, Mvera B, Brent AJ, et al. A Preliminary Study of Pneumonia Etiology Among Hospitalized Children in

Kenya. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 2011 12/22;54:S190-9.

173. Walk J, Dinga P, Banda C, Msiska T, Chitsamba E, Chiwayula N, et al. Non-invasive ventilation with bubble CPAP is feasible and improves respiratory physiology in hospitalised Malawian children with acute respiratory failure. Paediatrics and international child health. 2014 11/30;36(1):28-33.

174. Pirracchio R, Resche Rigon M, Mebazaa A, Zannad F, Alla F, Chevret S. Continuous positive airway pressure (CPAP) may not reduce short-term mortality in cardiogenic pulmonary edema: A propensity-based analysis. J Card Fail. 2013 February 2013;19(2):108-16.

175. Garcia-Rio F, Alonso-Fernandez A, Armada E, Mediano O, Lores V, Rojo B, et al. CPAP effect on recurrent episodes in patients with sleep apnea and myocardial infarction. Int J Cardiol. 2013 30 Sep 2013;168(2):1328-35.

176. Thadhani R. Formal trials versus observational studies. Mehta A, Beck M, Sunder-Plassmann G, editors. Fabry Disease: Perspectives from 5 Years of FOS. Oxford: Oxford PharmaGenesis; 2006.

177. Zachariasse JM, van dL, Seiger N, de Vos-Kerkhof E, Oostenbrink R, Moll HA. The role of nursesâ€™ clinical impression in the first assessment of children at the emergency department. Arch Dis Child. 2017 11/01;102(11):1052.

178. Tuti T, Agweyu A, Mwaniki P, Peek N, English M. An exploration of mortality risk factors in non-severe pneumonia in children using clinical data from Kenya. BMC Medicine. 2017 11/13;15(1):201.

179. Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? BMJ. 1998 01/17;316(7126):201.

180. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2018/01;62(5):464-75.

181. Lauer MS, D'Agostino RB. The Randomized Registry Trial â€" The Next Disruptive Technology in Clinical Research? N Engl J Med. 2013;369(17):1579-81.

182. Grignolo A. The Clinical Trials Transformation Initiative (CTTI). Annali dell'Istituto superiore di sanita. 2011;47:14-8.

183. Fröbert O, Lagerqvist B, Olivecrona G, Omerovic E, Gudnason T, Maeng M, et al. Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. N Engl J Med. 2013 10/24; 2018/01;369(17):1587-97.

184. Cowles MK, Carlin BP. Markov Monte Carlo Convergence Diagnostics: A Comparative Review. American Statistical Association. 1996;91(434):883-904.

185. Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. American Statistical Association. 1998;7(4):434-55.

186. Maternal and Child heath: Kenya [Internet].; 2011 []. Available from: http://www.who.int/pmnch/media/membernews/2011/20121216_kenyaparliament.pdf.

187. Munos MK, Fischer W, Christa L, Black RE. The effect of oral rehydration solution and recommended home fluids on diarrhoea mortality. Int J Epidemiol. 2010(39):75-87.

188. Zinc supplementation in the management of diarrhoea [Internet].; 2017 []. Available from: http://www.who.int/elena/titles/zinc_diarrhoea/en/.

189. Charles PL, Roy SK, Khan AI, Ahmed SR, Firdausi QJ. Zinc Treatment to Under-five Children: Applications to Improve Child Survival andReduce Burden of Disease. Health Popul Nutr. 2008;26(3):356-65.

190. Santosham M, Chandran A, Fitzwater S, Fischer-Walker C, Baqui AH, Black R. Progress and barriers for the control of diarrhoeal disease . The Lancet. 2010;376(9734):63-7.

191. Sample Size Calculations for Survival Analysis [Internet].; 2017 []. Available from: http://www.icssc.org/Documents/AdvBiosGoa/Tab%2026.00_SurvSS.pdf.

192. Chiba Y. Bias analysis of the instrumental variable estimator as an estimator of the average causal effect. Contemporary Clinical Trials. 2010 January 2010;31(1):12-7.

193. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat Methods Med Res. 2014 01 Oct 2014;23(5):440-59.

9. Appendices

Appendix A – Chapter 1

Appendix A.1: Summary of pneumonia antibiotic treatments

Study	Country sites	Severity[30]	Study arms	Number enrolled	Outcome	Effect size
Asghar (2008)	Zambia, Yemen, Pakistan, Mexico, India, Ecuador, and Bangladesh	Very severe	ampicillin vs chloramphenicol	958	Mortality	1.65 [0.99, 2.77]
Duke (2002)	Papua New Guinea	Very severe	benzyl penicillin plus gentamicin vs chloramphenicol	1116	Mortality	1.25 [0.76, 2.07]
Bansal (2006)	India	Very severe	crystalline penicillin plus gentamicin (followed by amoxicillin) vs IV and oral amoxicillin-clavulanate	71	Treatment failure	0.86 [0.05, 14.39]
Ribeiro (2011)	Brazil	Very severe	IV oxacillin vs co-amoxyclavulanic	104		0.98 [0.33, 2.92]
Cetinkaya (2004)	Turkey	Severe	IV penicillin plus chloramphenicol vs IV ceftriaxone	97		1.36 [0.47, 3.93]
Straus (1998)	Pakistan	Severe	co-trimoxazole vs amoxicillin	595		
Addo-Yobo (2004)	Colombia, Ghana, India, Mexico, Pakistan, South Africa, Vietnam, and Zambia	Severe	oral amoxicillin versus injectable penicillin	1702	Treatment failure	0.97 [0.77, 1.22]
Hazir (2008)	Pakistan	Severe	ampicillin (followed by oral amoxicillin) vs amoxicillin	2037	Treatment failure	0.86 [0.63, 1.19]
Atkinson (2007)	England	Severe	oral amoxicillin versus benzyl penicillin	246	Treatment failure	1.15 [0.58, 2.30]
Campbell (1988)	Gambia	Severe	procaine penicillin plus (benzyl penicillin) followed by ampicillin vs co-trimoxazole	143	Treatment failure	0.98 [0.12, 1.97]
Peltola (2001)	Gambia	Severe	procaine penicillin or cefuroxime (4 days) vs procaine penicillin or cefuroxime (7days)	154		_
Shann (1985)	Papua New Guinea	Severe	chloramphenicol vs chloramphenicol plus penicillin	748	Treatment failure	0.49 [0.48, 1.09]
Mulholland (1995)	Gambia	(very) severe	trimethoprim-sulfamethoxazole vs chloramphenicol	144	Treatment failure	1.03 [0.45, 2.33]
Nogeova (1997)	Slovakia	(very) severe	Ceftibuten vs cefuroxime	140	Treatment failure	6.81 [1.46, 31.70]
Klein (1995)	South Africa	LRTI	Cefpodoxime vs co-amoxyclavulanate	348	Cure rate	0.69 [0.18, 2.60]
CATCHUP (2002)	Pakistan	Non – severe	Amoxicillin vs co-trimoxazole	1471	Treatment failure	0.85 [0.68, 1.07]
Awasthi (2008)	India	Non – severe	Amoxicillin vs co-trimoxazole	2009	Treatment failure	1.45 [1.13, 1.85]
Bradley (2007)	Argentina,Brazil, Chile, Costa Rica, Mexico, Panama, and United	Non – severe	Levofloxacin vs Amoxicillin and clavulanic acid	709	Cure rate	1.05 [0.46, 2.42]

[30] Severity classification is based on the 2005 WHO guidelines: http://apps.who.int/iris/bitstream/10665/43206/1/9241546700.pdf

	States					
Harris (1998)	USA	Non – severe	Azithromycin vs. Augmentin	456	Treatment failure	1.78 [0.20, 16.27]
Wubbel (1999)	USA	Non – severe	Azithromycin vs. Augmentin	174	Treatment failure	1.26 [0.08, 20.86]
Keeley (1990)	Zimbabwe	Non – severe	Cotrimoxazole vs IM Procaine penicillin	617	Treatment failure	1.72 [0.41, 7.27]
Deivanayagam (1996)	Madras	Non – severe	Ampicillin vs penicillin and Chloramphenicol	115	Cure rate	0.48 [0.15, 1.51]
Camargos (1997)	Brazil	Non – severe	Benzathine penicillin vs procaine penicillin	176	Cure rate	0.60 [0.19, 1.86]
Rasmussen (2005)	Pakistan	Non – severe	4 mg/kg trimethoprim plus 20 mg/kg sulfamethoxazole vs 8 mg/kg trimethoprim plus 40 mg/kg sulfamethoxazole	1143	Treatment failure	1.10 [0.87, 1.37]
Fonseca (2003)	Brazil	Non - severe	Amoxicillin (15 mg/kg) vs Amoxicillin (25 mg/kg)	66	Treatment concentration	_
Dashner (1981)		Non - severe	Amoxicillin 50 mg/kg/day BID vs Amoxicillin 50 mg/kg/day TID	34		_
Cook (1996)	UK	Non - severe	Amoxicillin/ clavulanate BID vs Amoxicillin / clavulanate TID	437	Cure rate (difference)	0.032 [-0.043, 0.11]
Kartasasmita (2007)	Indonesia and Bangladesh	Non - severe	Cotrimoxazole (5 days) vs Cotrimoxazole (3 days)	2000	Cure rate (difference)	0.03 [-0.05, 0.18]
Lupisan (2002)	Philippines	Non - severe	Cotrimoxazole (5 days) vs Cotrimoxazole (3 days)	445		
MASCOT (2002)	Pakistan	Non - severe	Amoxicillin (5 days) vs Amoxicillin (3 days)	2000	Treatment failure (difference)	0.007 [-0.018, 0.032]
ISCAP (2004)	India	Non - severe	Amoxicillin (5 days) vs Amoxicillin (3 days)	2188	Treatment failure (difference)	0.004 [-0.021, 0.03]
Ficnar (1997)	Croatia	Non - severe	Azithromycin (5 days) vs Azithromycin (3 days)	371	Cure rate	0.004 [-0.013, 0.056
Tsarouhas (1998)	USA	Non - severe	Amoxicillin vs penicillin	154	Treatment failure	0.75 [0.17, 3.25]
Sidal (1994)	Turkey	Non - severe	Co – trimoxazole vs penicillin	151	Cure rate	0.57 [0.16, 2.01]
Jibril (1989)	Nigeria	Non - severe	Amoxicillin vs co-Amoxyclavulanic acid	238	Cure rate	0.20 [0.08, 0.53]
Hazir (2010)	Pakistan	Non - severe	Amoxicillin vs placebo	900	Treatment failure	0.77 [0.56, 1.04]
Block (1995)	USA	Non - severe	Clarithromycin (15 mg/kg) vs Clarithromycin (40 mg/kg)	234	Cure rate	1.61 [0.84, 3.08]
Aurangzeb (2003)	Pakistan	Non - severe	Amoxicillin vs Cefuroxime vs. Clarithromycin	171	Cure rate	2.05 [0.18, 23.51]
Roord (1996)	Netherlands	Non - severe	Azithromycin (5 days) vs Azithromycin (3 days)	89	Cure rate	1.07 [0.43, 2.66]
Kogan (2003)	Chile	Non - severe	Azithromycin (3 days) vs Amoxicillin (7 days)	110	Cure rate	17.14 [0.90, 325.93]
Ferwerda (2001)	Netherlands	Non - severe	Azithromycin (3 days) vs Co amox-clav (10 days)	110		_

Appendix A.2: Paediatric Admission Record Form

Paediatric Admission Record – Paediatric Ward

Name		IP No.		Ward			
Residence Sub-location				DOB	dd/mm/yyyy		
Adm Date	dd/mm/yyyy	Sex	M□ / F□	Age	yrs	Mths	days
Referred to hospital?	Y□ N□ *if yes from facility (name):*			Re-admission to <u>this</u> hospital?	Y□ N □		
Presenting Complaints							

History & Examination

Time seen	am □ pm □	Vaccines	*OPV/ Penta* X	*Pneum* X	*Rota V* X	*BCG* Y□ N□	*Measles* Y□ N□
Weight	Kg	Height / Length	Cm	WHZ=		MUAC	mm

Length of illness	Days		*History of presenting complaint*
Fever – No. of days =	Y□	N□	
Cough– No. of days =	Y□	N□	
…if cough yes: Cough > 2 wks	Y□	N□	
Difficulty breathing	Y□	N□	
Diarrhoea No. of days =	Y□	N□	
…if diarrhoea yes: Diarrhoea > 14d	Y□	N□	
Diarrhoea bloody	Y□	N□	*Review of Systems:*
Vomiting, No / 24hrs =	Y□	N□	Respiratory including ENT
Vomits everything	Y□	N□	
Difficulty feeding	Y□	N□	
Convulsions: No. =	Y□	N□	

Partial / focal fits?	Y□	N□	Cardiovascular Gastro-intestinal / Genitourinary

Birth History / Development / Neurological system

Nutritional history (record immunisations above)

Treatment History

Past Medical & Social History: HIV status: Father_____ Mother_____ Children_____

Examination

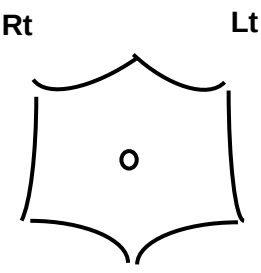
Vital Signs	Temp (°C)		Resp Rate	Bpm	HR	/min	O2 Sat	%	BP	/ mmHg

General Examination

Thrush **Y□ N□** Lymph N > 1cm **Y□ N□**

Wrist / Rib signs Rickets **Y□ N□**

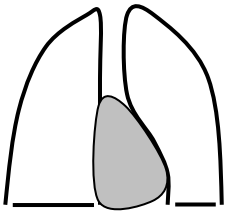
Jaundice	0	+	+++
Oedema	□None □Foot □Knee □Face		
Visible severe wasting		**Y□**	**N□**
A Stridor		**Y□**	**N□**

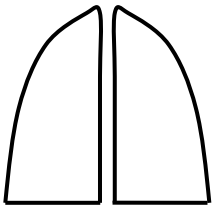


If newborn: Umbilicus Normal □ Pus □ Pus & red skin □

B	Central Cyanosis	**Y□**	**N□**		
	Indrawing	**Y□**	**N□**		
	Grunting	**Y□**	**N□**		
	Acidotic breathing	**Y□**	**N□**		
	Wheeze	**Y□**	**N□**		
	Crackles	**Y□**	**N□**		
C & Dehydr'n	Peripheral Pulse	□ Normal	□ Weak		
	Cap Refill	Secs	*X = not possible*		
	Pallor / Anaemia	**0**	**+**	**+++**	
	Skin warm at:	**□Hand □Elbow □Shoulder**			
	Sunken eyes	**Y□**	**N□**		
	Skin pinch (sec)	**0**	**1**	**≥ 2**	
D	**AVPU**	**A**	**V**	**P**	**U**
	Can drink / breastfeed?	**Y□**	**N□**		
	Stiff neck	**Y□**	**N□**		
	Bulging fontanelle	**Y□**	**N□**		
Infant < 2m	**Irritable**	**Y□**	**N□**		
	Reduced movement / tone	**Y□**	**N□**		

Front **R** **L**

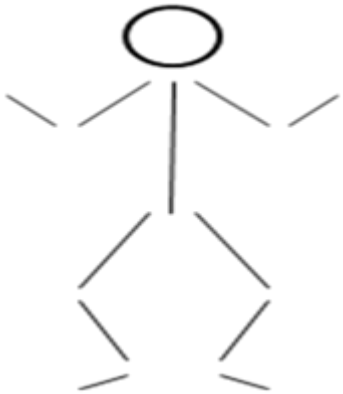




Back

Murmurs

Bones & Joints

ENT exam	**Peripheral Nervous System / Central Nervous System – Head Circumf =**
Rt Ear Lt Ear Nose Throat	**Rt** **Lt** 

Summary of presentation & problems & required investigations – record other tests and all results in medical record

Malaria	□ Blood slide □ Rapid Test	Glucose	□ Stick test □ Laboratory
Haematology	□ Hb □ HCT □ Full haemogram	Chemistry	□Na + K □U&C □Ca □Alb □ LFT
Microbiology	□ Lumbar Punct □ Blood Cult	HIV	□ Rapid test □ PCR
X-Ray	□ CXR □Wrist Other =	***Urine***	□ Urinalysis □ Micro & culture
Other 1		***Other 2***	

Admission Diagnoses – Select ONE *primary diagnosis (tick box indicating "1") and ANY secondary diagnoses (tick box indicating "2"), then indicate level of severity or type of disease if required*

Malaria	1□ 2□	□ Severe □ Non-severe	*Anaemia*	1 2 □ □	□ Sev □Non-sev
Pneumonia	1□ 2□	□ Very Sev □ Severe □ Non-sev	*Asthma*	1 2 □ □	□VSev □Sev □NonSev
Diarrhoea	1□ 2□	□ Non-bloody □ Bloody (dysentery)	*Meningitis*		1□ 2□
Dehydration	1□ 2□	□ Shock □ Severe □ Some	***Rickets***		1□ 2□
HIV	1□ 2□	□Pos □PMTCT+ve □Neg □Refused	***Neonatal sepsis***		1□ 2□
Malnutrition	1□ 2□	□ Severe (only record severe here)	***Prematurity / LBW***		1□ 2□
Other 1	1□ 2□				
Other 2	1□ 2□				

Treatment Supportive care & Observations – indicate what care is needed and sign please

Keep warm	□	Oxygen	□	Clinician Name & Sig
iv / oral fluids plan	□	Blood transfusion	□	
Vitamin A	□	**Nutrition / Feeds plan**	□	
Review status	□ Medical <6hrs □ Priority Nursing Observations			

Appendix B – Chapter 2

References for reviewed articles

Abilleira S., Ribera A., Cardona P., Rubiera M., LopezCancio E., Amaro S., et al. (2017). Outcomes after direct thrombectomy or combined intravenous and endovascular treatment are not different. *Stroke, 48*(2), 375-378.

Agiro A., Sylwestrzak G., Shah C., Power T., & Devries, A. (2014). Examining the association between utilization management and downstream cardiovascular imaging. *Health Services Research, 49*(5), 1616-1637

Alkhouli M., Zack C.J., Zhao H., Shafi I., & Bashir, R. (2015). Comparative outcomes of catheter-directed thrombolysis plus anticoagulation versus anticoagulation alone in the treatment of inferior vena caval thrombosis. *Circulation: Cardiovascular Interventions, 8*(2) (pagination), Arte Number: e001882. ate of Pubaton: 01 Feb 2015.

Amar D., Zhang H., Pedoto A., Desiderio D.P., Shi W., & Tan, K. S. (2017). Protective lung ventilation and morbidity after pulmonary resection: A propensity score-matched analysis. *Anesthesia and Analgesia, 125*(1), 190-199.

Andell P., Karlsson S., Mohammad M.A., Gotberg M., James S., Jensen J., et al. (2017). Intravascular ultrasound guidance is associated with better outcome in patients undergoing unprotected left main coronary artery stenting compared with angiography guidance alone. *Circulation: Cardiovascular Interventions, 10*(5) (pagination), Arte Number: e004813. ate of Pubaton: 01 May 2017.

Anstrom K.J., Brennan J.M., Eisenstein E.L., Federspiel J.J., Dai D., Peterson E.D., et al. (2014). Examination of the treatment selection process in a multicenter observational study. *Circulation: Cardiovascular Quality and Outcomes, 7*(5), 764-769. Antonelli A., Minervini A., Mari A., Bertolo R.,

Bianchi G., Lapini A., et al. (2015). TriMatch comparison of the efficacy of FloSeal versus TachoSil versus no hemostatic agents for partial nephrectomy: Results from a large multicenter dataset. *International Journal of Urology, 22*(1), 47-52.

Asada T., Aoki Y., Sugiyama T., Yamamoto M., Ishii T., Kitsuta Y., et al. (2016). Organ system network disruption in nonsurvivors of critically ill patients. *Critical Care Medicine, 44*(1), 83-90.

Austevoll I.M., Gjestad R., Brox J.I., Solberg T.K., Storheim K., Rekeland F., et al. (2017). The effectiveness of decompression alone compared with additional fusion for lumbar spinal stenosis with degenerative spondylolisthesis: A pragmatic comparative non-inferiority observational study from the norwegian registry for spine surgery. *European Spine Journal, 26*(2), 404-413.

Bachmann M., Bachmann C.J., John K., HeinzelGutenbrunner M., Remschmidt H., & Mattejat, F. (2010). The effectiveness of child and adolescent psychiatric treatments in a naturalistic outpatient setting. *World Psychiatry, 9*(2), 111-117.

Badalato G.M., Kates M., Wisnivesky J.P., Choudhury A.R., & McKiernan, J. M. (2012). Survival after partial and radical nephrectomy for the treatment of stage T1bN0M0 renal cell carcinoma (RCC) in the USA: A propensity scoring approach. *BJU International, 109*(10), 1457-1462. Bajwa E.K., Malhotra C.K., Thompson B.T., Christiani D.C., & Gong, M. N. (2012). Statin therapy as prevention against development of acute respiratory distress syndrome: An observational study. *Critical Care Medicine, 40*(5), 1470-1477.

Ballard D.J., Ogola G., Fleming N.S., Stauffer B.D., Leonard B.M., Khetan R., et al. (2010). Impact of a standardized heart failure order set on mortality, readmission, and quality and costs of care. *International Journal for Quality in Health Care, 22*(6) (pp 437-444), Arte Number: mzq051. ate of Pubaton: eember 2010.

Barbin L., Rousseau C., Jousset N., Casey R., Debouverie M., Vukusic S., et al. (2016). Comparative efficacy of fingolimod vs natalizumab. *Neurology, 86*(8), 771-778.

Batterink J., Lin J., AuYeung S.H.M., & Cessford, T. (2015). Effectiveness of sodium polystyrene sulfonate for short-term treatment of hyperkalemia. *Canadian Journal of Hospital Pharmacy, 68*(4), 296-303.

Bayer O., Schwarzkopf D., Doenst T., Cook D., Kabisch B., Schelenz C., et al. (2013). Perioperative fluid therapy with tetrastarch and gelatin in cardiac surgery - A prospective sequential analysis. *Critical Care Medicine, 41*(11), 2532-2542.

Biersteker H.A.R., Andriessen T.M.J.C., Horn J., Franschman G., Van Der Naalt J., Hoedemaekers C.W.E., et al. (2012). Factors influencing intracranial pressure monitoring guideline compliance and outcome after severe traumatic brain injury. *Critical Care Medicine, 40*(6), 1914-1922. Bige N., Hejblum G., Baudel J.L., Carron A., Chevalier S., Pichereau C., et al. (2015). Homeless patients in the ICU: An observational propensity-matched cohort study. *Critical Care Medicine, 43*(6), 1246-1254.

Bin Abdulhak A.A., Kennedy K.F., Gupta S., Giocondo M., Ramza B., & Wimmer, A. P. (2015). Effect of pre-procedural interrupted apixaban on heparin anticoagulation during catheter ablation for atrial fibrillation: A prospective observational study. *Journal of Interventional Cardiac Electrophysiology, 44*(2), 91-96.

Bulka C.M., Terekhov M.A., Martin B.J., Dmochowski R.R., Hayes R.M., & Ehrenfeld, J. M. (2016). Nondepolarizing neuromuscular blocking agents, reversal, and risk of postoperative pneumonia. *Anesthesiology, 125*(4), 647-655.

Cadilhac D.A., Kim J., Lannin N.A., Levi C.R., Dewey H.M., Hill K., et al. (2016). Better outcomes for hospitalized patients with TIA when in stroke units: An observational study. *Neurology, 86*(22), 2042-2048.

Cao L., Young N., Liu H., Silvestry S., Sun W., Zhao N., et al. (2012). Preoperative aspirin use and outcomes in cardiac surgery patients. *Annals of Surgery, 255*(2), 399-404.

Carlile M., Nicewander D., Yablon S.A., Brown A., Brunner R., Burke D., et al. (2010). Prophylaxis for venous thromboembolism during rehabilitation for traumatic brain injury: A multicenter observational study. *Journal of Trauma - Injury, Infection and Critical Care, 68*(4), 916-923. CartinCeba R., Warner D.O., Hays J.T., & Afessa, B. (2011). Nicotine replacement therapy in critically ill patients: A prospective observational cohort study. *Critical Care Medicine, 39*(7), 1635-1640.

Chao P.W., Chu H., Chen Y.T., Shih Y.N., Kuo S.C., Li S.Y., et al. (2016). Long-term outcomes in critically ill septic patients who survived cardiopulmonary resuscitation. *Critical Care Medicine, 44*(6), 1067-1074.

Chenouard A., Roze J.C., Hanf M., Macher J., Liet J.M., Gournay V., et al. (2015). Evaluation of the relationship between plasma transfusion and nosocomial infection after cardiac surgery in children younger than 1 year. *Pediatric Critical Care Medicine, 16*(2), 139-145.

Chester Wasko M., Dasgupta A., Ilse Sears G., Fries J.F., & Ward, M. M. (2016). Prednisone use and risk of mortality in patients with rheumatoid arthritis: Moderation by use of disease-modifying antirheumatic drugs. *Arthritis Care and Research, 68*(5), 706-710.

Cho Y., Badve S.V., Hawley C.M., McDonald S.P., Brown F.G., Boudville N., et al. (2014). Peritoneal dialysis outcomes after temporary haemodialysis transfer for peritonitis. *Nephrology Dialysis Transplantation, 29*(10), 1940-1947.

Chuang C.M., Chou Y.J., Yen M.S., Chao K.C., Twu N.F., Wu H.H., et al. (2012). The role of secondary cytoreductive surgery in patients with recurrent epithelial ovarian, tubal, and peritoneal cancers: A comparative effectiveness analysis. *Oncologist, 17*(6) (pp 847-855).

Chung R., Houghtaling P.L., Tchou M., Niebauer M.J., Lindsay B.D., Tchou P.J., et al. (2014). Left ventricular hypertrophy and antiarrhythmic drugs in atrial fibrillation: Impact on mortality. *PACE - Pacing and Clinical Electrophysiology, 37*(10), 1338-1348.

Clouston S.A.P., Kuh D., Herd P., Elliott J., Richards M., & Hofer, S. M. (2012). Benefits of educational attainment on adult fluid cognition: International evidence from three birth cohorts. *International Journal of Epidemiology, 41*(6) (pp 1729-1736), Arte Number: ys148. ate of Pubaton: eember 2012.

Corsonello A., Maggio M., Fusco S., Adamo B., Amantea D., Pedone C., et al. (2014). Proton pump inhibitors and functional decline in older adults discharged from acute care hospitals. *Journal of the American Geriatrics Society, 62*(6), 1110-1115.

DirajlalFargo S., Alam K., Sattar A., Kulkarni M., Funderburg N., Wilson W.H., et al. (2017). Comprehensive assessment of the arginine pathway and its relationship to inflammation in HIV. *Aids, 31*(4), 533-537.

Dong Y.H., Bykov K., Choudhry N.K., Donneyong M.M., Huybrechts K.F., Levin R., et al. (2017). Clinical outcomes of concomitant use of warfarin and selective serotonin reuptake inhibitors: A multidatabase observational cohort study. *Journal of Clinical Psychopharmacology, 37*(2), 200-209.

Doussau A., Perez P., Puntous M., Calderon J., Jeanne M., Germain C., et al. (2014). Fresh-frozen plasma transfusion did not reduce 30-day mortality in patients undergoing cardiopulmonary bypass cardiac surgery with excessive bleeding: The PLASMACARD multicenter cohort study. *Transfusion, 54*(4), 1114-1124.

Dugoff E.H., Bekelman J.E., Stuart E.A., Armstrong K., & Pollack, C. E. (2014). Surgical quality is more than volume: The association between changing urologists and complications for patients with localized prostate cancer. *Health Services Research, 49*(4), 1165-1183.

Ekser B., Mangus R.S., Fridell W., Kubal C.A., Nagai S., Kinsella S.B., et al. (2017). A novel approach in combined liver and kidney transplantation with long-term outcomes. *Annals of Surgery, 265*(5), 1000-1008.

FernandezRuiz M., Arias M., Campistol J.M., Navarro D., GomezHuertas E., GomezMarquez G., et al. (2015). Cytomegalovirus prevention strategies in seropositive kidney transplant recipients: An insight into current clinical practice. *Transplant International, 28*(9), 1042-1054.

Fouad A.M., Waheed A., Gamal A., Amer S.A., Abdellah R.F., & Shebl, F. M. (2017). Effect of chronic diseases on work productivity: A propensity score analysis. *Journal of Occupational and Environmental Medicine, 59*(5), 480-485.

Frajzyngier V., Ruminjo J., Asiimwe F., Barry T.H., Bello A., Danladi D., et al. (2012). Factors influencing choice of surgical route of repair of genitourinary fistula, and the influence of route of repair on surgical outcomes: Findings from a prospective cohort study. *BJOG: An International Journal of Obstetrics and Gynaecology, 119*(11), 1344-1353.

Friedman B., Barbash G.I., Glied S.A., & Steiner, C. A. (2016). Hospital revisits within 30 days after conventional and robotically assisted hysterectomy. *Medical Care, 54*(3), 311-318.

Fu A.Z., Johnston S.S., Ghannam A., Tsai K., Cappell K., Fowler R., et al. (2016). Association between hospitalization for heart failure and dipeptidyl peptidase 4 inhibitors in patients with type 2 diabetes: An observational study. *Diabetes Care, 39*(5), 726-734.

Gallina P., Copetti M., Pilotto A., Marcato F., Mello A.M., Simonato M., et al. (2016). Warfarin treatment and all-cause mortality in community-dwelling older adults with atrial fibrillation: A retrospective observational study. *Journal of the American Geriatrics Society, 64*(7), 1416-1424.

Garlipp B., Ptok H., Benedix F., Otto R., Popp F., Ridwelski K., et al. (2016). Adjuvant treatment for resected rectal cancer: Impact of standard and intensified postoperative chemotherapy on disease-free survival in patients undergoing preoperative chemoradiation-a propensity score-matched analysis of an observational database. *Langenbeck's Archives of Surgery, 401*(8), 1179-1190.

Glenn D., Golinelli D., Rose R.D., RoyByrne P., Stein M.B., Sullivan G., et al. (2013). Who gets the most out of cognitive behavioral therapy for anxiety disorders? the role of treatment dose and patient engagement. *Journal of Consulting and Clinical Psychology, 81*(4), 639-649.

Gozalo P., Leland N.E., Christian T.J., Mor V., & Teno, J. M. (2015). Volume matters: Returning home after hip fracture. *Journal of the American Geriatrics Society, 63*(10), 2043-2051.

Graham J., Tomcavage J., Salek D., Sciandra J., Davis D.E., & Stewart, W. F. (2012). Postdischarge monitoring using interactive voice response system reduces 30-day readmission rates in a case-managed medicare population. *Medical Care, 50*(1), 50-57.

Gray E., Pasta D.J., Norris S., & O'Leary, A. (2017). Effectiveness of triple therapy with direct-acting antivirals for hepatitis C genotype 1 infection: Application of propensity score matching in a national HCV treatment registry. *BMC Health Services Research, 17*(1), 288.

Greiss H., Berkowitsch A., Wojcik M., Zaltsberg S., Pajitnev D., Deubner N., et al. (2015). The impact of left atrial surface area and the second generation cryoballoon on clinical outcome of atrial fibrillation cryoablation. *PACE - Pacing and Clinical Electrophysiology, 38*(7), 815-824.

Gronich N., Deftereos S.N., Lavi I., Persidis A.S., Abernethy D.R., & Rennert, G. (2015). Hypothyroidism is a risk factor for new-onset diabetes: A cohort study. *Diabetes Care, 38*(9), 1657-1664.

Gross A.E., Van Schooneveld T.C., Olsen K.M., Rupp M.E., Bui T.H., Forsung E., et al. (2014). Epidemiology and predictors of multidrug-resistant community-acquired and health care-associated pneumonia. *Antimicrobial Agents and Chemotherapy, 58*(9), 5262-5268.

Gupta P., Rettiganti M., Fisher P.L., Chang A.C., Rice T.B., & Wetzel, R. C. (2016). Association of freestanding children's hospitals with outcomes in children with critical illness*. *Critical Care Medicine, 44*(12), 2131-2138.

Gupta P., Richardson T., Hall M., Bertoch D., Hebbar K.B., Fortenberry J.D., et al. (2016). Effect of inhaled nitric oxide on outcomes in children with acute lung injury: Propensity matched analysis from a linked database. *Critical Care Medicine, 44*(10), 1901-1909.

Hamady Z.Z.R., Lodge J.P.A., Welsh F.K., Toogood G.J., White A., John T., et al. (2014). One-millimeter cancer-free margin is curative for colorectal liver metastases: A propensity score case-match approach. *Annals of Surgery, 259*(3), 543-548.

Harrold L.R., Reed G.W., Solomon D.H., Curtis J.R., Liu M., Greenberg J.D., et al. (2016). Comparative effectiveness of abatacept versus tocilizumab in rheumatoid arthritis patients with prior TNFi exposure in the US corrona registry. *Arthritis Research and Therapy, 18*(1) (pagination), Arte Number: 280. ate of Pubaton: 01 e 2016.

Her A.Y., Cho K.I., Singh G.B., Garg S., Kim Y.H., Koo B.K., et al. (2017). A comparison of peri-procedural myocardial infarction between paclitaxel-coated balloon and drug-eluting stent on de novo coronary lesions. *Yonsei Medical Journal, 58*(1), 99-104.

Hoffman R.M., Koyama T., Fan K.H., Albertsen P.C., Barry M.J., Goodman M., et al. (2013). Mortality after radical prostatectomy or external beam radiotherapy for localized prostate cancer. *Journal of the National Cancer Institute, 105*(10), 711-718.

Honda M., Hiki N., Kinoshita T., Yabusaki H., Abe T., Nunobe S., et al. (2016). Long-term outcomes of laparoscopic versus open surgery for clinical stage i gastric cancer: The LOC-1 study. *Annals of Surgery, 264*(2), 214-222.

Hong J.L., Funk M.J., Buse J.B., Henderson L.M., Lund J.L., Pate V., et al. (2017). Comparative effect of initiating metformin versus sulfonylureas on breast cancer risk in older women. *Epidemiology, 28*(3), 446-454.

Horneff G., Klein A., Klotsche J., Minden K., Huppertz H.I., WellerHeinemann F., et al. (2016). Comparison of treatment response, remission rate and drug adherence in polyarticular juvenile idiopathic arthritis patients treated with etanercept, adalimumab or tocilizumab. *Arthritis Research and Therapy, 18*(1) (pagination), Arte Number: 272. ate of Pubaton: 24 No 2016.

Ibinson J.W., Ezaru C.S., Cormican D.S., & Mangione, M. P. (2014). GlideScope use improves intubation success rates: An observational study using propensity score matching. *BMC Anesthesiology, 14*(1) (pagination), Arte Number: 101. ate of Pubaton: Noember 05, 2014.

Inoue T., & Fushimi, K. (2013). Stroke care units versus general medical wards for acute management of stroke in japan. *Stroke, 44*(11), 3142-3147.

Iqbal M.B., Arujuna A., Ilsley C., Archbold A., Crake T., Firoozi S., et al. (2014). Radial versus femoral access is associated with reduced complications and mortality in patients with non-ST-segment-elevation myocardial infarction: An observational cohort study of 10 095 patients. *Circulation: Cardiovascular Interventions, 7*(4), 456-464.

Iqbal M.B., Ilsley C., Kabir T., Smith R., Lane R., Mason M., et al. (2014). Culprit vessel versus multivessel intervention at the time of primary percutaneous coronary intervention in patients with ST-segment-elevation myocardial infarction and multivessel disease: Real-world analysis of 3984 patients in london. *Circulation: Cardiovascular Quality and Outcomes, 7*(6), 936-943.

Karkouti K., Callum J., Crowther M.A., McCluskey S.A., Pendergrast J., Tait G., et al. (2013). The relationship between fibrinogen levels after cardiopulmonary bypass and large volume red cell transfusion in cardiac surgery: An observational study. *Anesthesia and Analgesia, 117*(1), 14-22.

Kihara M., Davies R., KearsleyFleet L., Watson K.D., Lunt M., Symmons D.P.M., et al. (2017). Use and effectiveness of tocilizumab among patients with rheumatoid arthritis: An observational study from the british society for rheumatology biologics register for rheumatoid arthritis. *Clinical Rheumatology, 36*(2), 241-250.

Kilic S., Hermanides R.S., Ottervanger J.P., Kolkman E., Dambrink J.H.E., Roolvink V., et al. (2017). Effects of radial versus femoral artery access in patients with acute myocardial infarction: A large centre prospective registry. *Netherlands Heart Journal, 25*(1), 33-39.

Kim K.H., Kim Y.M., Kim M.C., & Jung, G. J. (2014). Analysis of prognostic factors and outcomes of gastric cancer in younger patients: A case control study using propensity score methods. *World Journal of Gastroenterology, 20*(12), 3369-3375.

Kim M.H., Johnston S.S., Chu B.C., Dalal M.R., & Schulman, K. L. (2011). Estimation of total incremental health care costs in patients with atrial fibrillation in the united states. *Circulation: Cardiovascular Quality and Outcomes, 4*(3), 313-320.

Kimmelstiel C., Pinto D., Aronow H.D., Weintraub A.R., Dangas G., Fan W., et al. (2016). Bivalirudin is associated with improved in-hospital outcomes compared with heparin in percutaneous vascular interventions: Observational, propensity-matched analysis from the premier hospital database. *Circulation: Cardiovascular Interventions, 9*(1) (pagination), ate of Pubaton: 01 Jan 2016.

Ko D.T., Wijeysundera H.C., Yun L., Austin P.C., Cantor W.J., & Tu, J. V. (2011). Effectiveness of preprocedural statin therapy on clinical outcomes for patients with stable coronary artery disease after percutaneous coronary interventions. *Circulation: Cardiovascular Quality and Outcomes, 4*(4), 459-466.

Kor D.J., Erlich J., Gong M.N., Malinchoc M., Carter R.E., Gajic O., et al. (2011). Association of prehospitalization aspirin therapy and acute lung injury: Results of a multicenter international observational study of at-risk patients. *Critical Care Medicine, 39*(11), 2393-2400.

Krebs E.E., Becker W.C., Zerzan J., Bair M.J., McCoy K., & Hui, S. (2011). Comparative mortality among department of veterans affairs patients prescribed methadone or long-acting morphine for chronic pain. *Pain, 152*(8), 1789-1795.

Kubota K., Egi M., & Mizobuchi, S. (2017). Haptoglobin administration in cardiovascular surgery patients: Its association with the risk of postoperative acute kidney injury. *Anesthesia and Analgesia, 124*(6), 1771-1776.

Kumar A., Matheny M.E., Ho K.K.L., Yeh R.W., Piemonte T.C., Waldman H., et al. (2015). The data extraction and longitudinal trend analysis network study of distributed automated postmarket cardiovascular device safety surveillance. *Circulation: Cardiovascular Quality and Outcomes, 8*(1), 38-46.

Kurlansky P., Herbert M., Prince S., & Mack, M. (2016). Coronary artery bypass graft versus percutaneous coronary intervention. *Circulation, 134*(17), 1238-1246.

Lange A., Kasperk C., Alvares L., Sauermann S., & Braun, S. (2014). Survival and Cost Comparison of Kyphoplasty and Percutaneous Vertebroplasty Using German Claims Data.

Lasa J.J., Rogers R.S., Localio R., Shults J., Raymond T., Gaies M., et al. (2016). Extracorporeal cardiopulmonary resuscitation (E-CPR) during pediatric in-hospital cardiopulmonary arrest is associated with improved survival to discharge. *Circulation, 133*(2), 165-176.

Le Manach Y., Collins G.S., Ibanez C., Goarin J.P., Coriat P., Gaudric J., et al. (2012). Impact of perioperative bleeding on the protective effect of beta-blockers during infrarenal aortic reconstruction. *Anesthesiology, 117*(6), 1203-1211.

Le Manach Y., Ibanez Esteves C., Bertrand M., Goarin J.P., Fleron M.H., Coriat P., et al. (2011). Impact of preoperative statin therapy on adverse postoperative outcomes in patients undergoing vascular surgery. *Anesthesiology, 114*(1), 98-104.

Lee Y.G., Kim I., Yoon S.S., Park S., Cheong J.W., Min Y.H., et al. (2013). Comparative analysis between azacitidine and decitabine for the treatment of myelodysplastic syndromes. *British Journal of Haematology, 161*(3), 339-347.

Leong C., AlessiSeverini S., Enns M.W., Nie Y., Sareen J., Bolton J., et al. (2017). Cerebrovascular, cardiovascular, and mortality events in new users of selective serotonin reuptake inhibitors and

serotonin norepinephrine reuptake inhibitors: A propensity score-matched population-based study. *Journal of Clinical Psychopharmacology, 37*(3), 332-340.

Lewinter C., Bland J.M., Crouch S., Cleland J.G.F., Doherty P., Lewinter M.M., et al. (2014). Impact of aspirin and statins on long-term survival in patients hospitalized with acute myocardial infarction complicated by heart failure: An analysis of 1706 patients. *European Journal of Heart Failure, 16*(1), 95-102.

Lin P.J., Zhong Y., Fillit H.M., Chen E., & Neumann, P. J. (2016). Medicare expenditures of individuals with alzheimer's disease and related dementias or mild cognitive impairment before and after diagnosis. *Journal of the American Geriatrics Society, 64*(8), 1549-1557.

Lindahl B., Baron T., Erlinge D., Hadziosmanovic N., Nordenskjold A., Gard A., et al. (2017). Medical therapy for secondary prevention and long-term outcome in patients with myocardial infarction with nonobstructive coronary artery disease. *Circulation, 135*(16), 1481-1489.

Liu P.H., Lee Y.H., Hsu C.Y., Huang Y.H., Chiou Y.Y., Lin H.C., et al. (2014). Survival advantage of radiofrequency ablation over transarterial chemoembolization for patients with hepatocellular carcinoma and good performance status within the milan criteria. *Annals of Surgical Oncology, 21*(12), 3835-3843.

LopezCortes L.E., Cisneros J.M., FernandezCuenca F., Bou G., Tomas M., GarnachoMontero J., et al. (2014). Monotherapy versus combination therapy for sepsis due to multidrug-resistant acinetobacter baumannii: Analysis of a multicentre prospective cohort. *Journal of Antimicrobial Chemotherapy, 69*(11), 3119-3126.

Maekawa K., Tanno K., Hase M., Mori K., & Asai, Y. (2013). Extracorporeal cardiopulmonary resuscitation for patients with out-of-hospital cardiac arrest of cardiac origin: A propensity-matched study and predictor analysis. *Critical Care Medicine, 41*(5), 1186-1196.

Mannino S., Villa M., Apolone G., Weiss N.S., Groth N., Aquino I., et al. (2012). Effectiveness of adjuvanted influenza vaccination in elderly subjects in northern italy. *American Journal of Epidemiology, 176*(6), 527-533.

Margolis D.J., Gupta J., Hoffstad O., Papdopoulos M., Glick H.A., Thom S.R., et al. (2013). Lack of effectiveness of hyperbaric oxygen therapy for the treatment of diabetic foot ulcer and the prevention of amputation a cohort study. *Diabetes Care, 36*(7), 1961-1966.

MartinezSanchez P., Fuentes B., MartinezMartinez M., RuizAres G., FernandezTravieso J., SanzCuesta B.E., et al. (2013). Treatment with statins and ischemic stroke severity: Does the dose matter?. *Neurology, 80*(19), 1800-1805.

MartinezSelles M., Gomez Doblas J.J., Carro Hevia A., Garcia de la Villa B., FerreiraGonzalez I., Alonso Tello A., et al. (2014). Prospective registry of symptomatic severe aortic stenosis in octogenarians: A need for intervention. *Journal of Internal Medicine, 275*(6), 608-620.

Masuda H., Takahashi Y., Nishida Y., & Asai, S. (2012). Comparison of the effect of mesalazine and sulfasalazine on laboratory parameters: A retrospective observational study. *European Journal of Clinical Pharmacology, 68*(11), 1549-1555.

Matsue Y., Shiraishi A., Kagiyama N., Kume T., Okura H., Suzuki M., et al. (2016). Renal function on admission modifies prognostic impact of diuretics in acute heart failure: A propensity score matched and interaction analysis. *Heart and Vessels, 31*(12), 1980-1987.

Maura G., Blotiere P.O., Bouillon K., Billionnet C., Ricordeau P., Alla F., et al. (2015). Comparison of the short-term risk of bleeding and arterial thromboembolic events in nonvalvular atrial fibrillation patients newly treated with dabigatran or rivaroxaban versus vitamin K antagonists a french nationwide propensity-matched cohort study. *Circulation, 132*(13), 1252-1260.

McClellan S.R., Panattoni L., Chan A.S., & TaiSeale, M. (2016). Patient-initiated electronic messages and quality of care for patients with diabetes and hypertension in a large fee-for-service medical group results from a natural experiment. *Medical Care, 54*(3), 287-295.

McIsaac D.I., Bryson G.L., & Van Walraven, C. (2014). Elective, major noncardiac surgery on the weekend: A population-based cohort study of 30-day mortality. *Medical Care, 52*(6), 557-564.

Meltzer C., Klau M., Gurushanthaiah D., Tsai J., Meng D., Radler L., et al. (2016). Surgeon volume in thyroid surgery: Surgical efficiency, outcomes, and utilization. *Laryngoscope, 126*(11), 2630-2639.

Minnerup J., Wersching H., Teuber A., Wellmann J., Eyding J., Weber R., et al. (2016). Outcome after thrombectomy and intravenous thrombolysis in patients with acute ischemic stroke: A prospective observational study. *Stroke, 47*(6), 1584-1592.

Mogensen K.M., Robinson M.K., Casey J.D., Gunasekera N.S., Moromizato T., Rawn J.D., et al. (2015). Nutritional status and mortality in the critically ill. *Critical Care Medicine, 43*(12), 2605-2615.

Molnar M.Z., Streja E., Kovesdy C.P., Hoshino J., Hatamizadeh P., Glassock R.J., et al. (2012). Estimated glomerular filtration rate at reinitiation of dialysis and mortality in failed kidney transplant recipients. *Nephrology Dialysis Transplantation, 27*(7), 2913-2921.

NadeauFredette A.C., Chan C.T., Cho Y., Hawley C.M., Pascoe E.M., Clayton P.A., et al. (2015). Outcomes of integrated home dialysis care: A multi-centre, multi-national registry study. *Nephrology Dialysis Transplantation, 30*(11), 1897-1904.

O'Brien D.D., Shanks A.M., Talsma A., Brenner P.S., & Ramachandran, S. K. (2014). Intraoperative risk factors associated with postoperative pressure ulcers in critically ill patients: A retrospective observational study. *Critical Care Medicine, 42*(1), 40-47.

Ouldamer L., Caille A., Giraudeau B., & Body, G. (2015). Quilting suture of mastectomy dead space compared with conventional closure with drain. *Annals of Surgical Oncology, 22*(13), 4233-4240.

Park D.W., Chun B.C., Kwon S.S., Yoon Y.K., Choi W.S., Sohn J.W., et al. (2012). Red blood cell transfusions are associated with lower mortality in patients with severe sepsis and septic shock: A propensity-matched analysis. *Critical Care Medicine, 40*(12), 3140-3145.

Park D.W., Egi M., Nishimura M., Chang Y., Suh G.Y., Lim C.M., et al. (2016). The association of fever with total mechanical ventilation time in critically ill patients. *Journal of Korean Medical Science, 31*(12), 2033-2041.

Park S.Y., Lee S.O., Choi S.H., Kim Y.S., Woo J.H., Baek S., et al. (2012). Efficacy and safety of low-dose ganciclovir preemptive therapy in allogeneic haematopoietic stem cell transplant recipients compared with conventional-dose ganciclovir: A prospective observational study. *Journal of Antimicrobial Chemotherapy, 67*(6) (pp 1486-1492), Arte Number: ks043. ate of Pubaton: June 2012.

Pasquali S.K., Hall M., Li J.S., Peterson E.D., Jaggers J., Lodge A.J., et al. (2010). Corticosteroids and outcome in children undergoing congenital heart surgery: Analysis of the pediatric health information systems database. *Circulation, 122*(21), 2123-2130.

Pasquier P., Gayat E., Rackelboom T., La Rosa J., Tashkandi A., Tesniere A., et al. (2013). An observational study of the fresh frozen plasma: Red blood cell ratio in postpartum hemorrhage. *Anesthesia and Analgesia, 116*(1), 155-161.

Patterson J.A., Irving D.O., Isbister J.P., Morris J.M., Mayson E., Roberts C.L., et al. (2015). Age of blood and adverse outcomes in a maternity population. *Transfusion, 55*(11), 2730-2737.

Pawaskar M., Bonafede M., Johnson B., Fowler R., Lenhart G., & Hoogwerf, B. (2013). Medication utilization patterns among type 2 diabetes patients initiating exenatide BID or insulin glargine: A retrospective database study. *BMC Endocrine Disorders, 13*(pagination), Arte Number: 20. ate of Pubaton: 22 Jun 2013.

Paxton E.W., Inacio M.C.S., Kurtz S., Love R., Cafri G., & Namba, R. S. (2015). Is there a difference in total knee arthroplasty risk of revision in highly crosslinked versus conventional polyethylene?. *Clinical Orthopaedics and Related Research, 473*(3), 999-1008.

Peng K.P., Chen Y.T., Fuh J.L., Tang C.H., & Wang, S. J. (2014). Increased risk of bell palsy in patients with migraine : A nationwide cohort study. *Neurology, 84*(2), 116-124.

Perlas A., Chan V.W.S., & Beattie, S. (2016). Anesthesia technique and mortality after total hip or knee arthroplasty. *Anesthesiology, 125*(4), 724-731.

Pollack C.E., Rastegar A., Keating N.L., Adams J.L., Pisu M., & Kahn, K. L. (2015). Is self-referral associated with higher quality care?. *Health Services Research, 50*(5), 1472-1490.

Pollett S., Baxi S.M., Rutherford G.W., Doernberg S.B., Bacchetti P., & Chambers, H. F. (2016). Cefazolin versus nafcillin for methicillin-sensitive staphylococcus aureus bloodstream infection in a california tertiary medical center. *Antimicrobial Agents and Chemotherapy, 60*(8), 4684-4689.

Ramcharran D., Qiu H., Schuemie M.J., & Ryan, P. B. (2017). Atypical antipsychotics and the risk of falls and fractures among older adults: An emulation analysis and an evaluation of additional confounding control strategies. *Journal of Clinical Psychopharmacology, 37*(2), 162-168.

Ranzani O.T., Ferrer M., Esperatti M., Giunta V., Bassi G.L., Carvalho C.R.R., et al. (2012). Association between systemic corticosteroids and outcomes of intensive care unit-acquired pneumonia. *Critical Care Medicine, 40*(9), 2552-2561.

Rathbun A.M., Yau M.S., Shardell M., Stuart E.A., & Hochberg, M. C. (2017). Depressive symptoms and structural disease progression in knee osteoarthritis: Data from the osteoarthritis initiative. *Clinical Rheumatology, 36*(1), 155-163.

Reames B.N., Scally C.P., Thumma J.R., & Dimick, J. B. (2015). Evaluation of the effectiveness of a surgical checklist in medicare patients. *Medical Care, 53*(1), 87-94.

Reents T., Buiatti A., Ammar S., Dillier R., Semmler V., Telishevska M., et al. (2015). Catheter ablation of ventricular arrhythmias using a fluoroscopy image integration module. *PACE - Pacing and Clinical Electrophysiology, 38*(6), 700-705.

Rodriguez A., Diaz E., MartinLoeches I., Sandiumenge A., Canadell L., Diaz J.J., et al. (2011). Impact of early oseltamivir treatment on outcome in critically ill patients with 2009 pandemic influenza A. *Journal of Antimicrobial Chemotherapy, 66*(5) (pp 1140-1149), Arte Number: kq511. ate of Pubaton: May 2011.

Rokx C., Gras L., van de Vijver D.A.M.C., Prins J.M., Kuijpers T.W., Scherpbier H.J., et al. (2016). Virological responses to lamivudine or emtricitabine when combined with tenofovir and a protease inhibitor in treatment-naive HIV-1-infected patients in the dutch AIDS therapy evaluation in the netherlands (ATHENA) cohort. *HIV Medicine, 17*(8), 571-580.

Rosato S., Santini F., Barbanti M., Biancari F., D'Errigo P., Onorati F., et al. (2016). Transcatheter aortic valve implantation compared with surgical aortic valve replacement in low-risk patients. *Circulation: Cardiovascular Interventions, 9*(5) (pagination), Arte Number: e003326. ate of Pubaton: 01 May 2016.

Rosenblatt L., Farr A.M., Johnston S.S., & Nkhoma, E. T. (2016). Risk of cardiovascular events among patients initiating efavirenz-containing versus efavirenz-free antiretroviral regimens. *Open Forum Infectious Diseases, 3*(2) (pagination), Arte Number: ofw061. ate of Pubaton: 01 Ar 2016.

Sansonnens J., Taffe P., & Burnand, B. (2016). Higher occurrence of nausea and vomiting after total hip arthroplasty using general versus spinal anesthesia: An observational study. *BMC Anesthesiology, 16*(1) (pagination), Arte Number: 44. ate of Pubaton: 26 Ju 2016.

Schneider M., Zuckerman I.H., Onukwugha E., Pandya N., Seal B., Gardner J., et al. (2011). Chemotherapy treatment and survival in older women with estrogen receptor-negative metastatic breast cancer: A population-based analysis. *Journal of the American Geriatrics Society, 59*(4), 637-646.

Schwann N.M., Hillel Z., Hoeft A., Barash P., Mohnle P., Miao Y., et al. (2011). Lack of effectiveness of the pulmonary artery catheter in cardiac surgery. *Anesthesia and Analgesia, 113*(5), 994-1002.

Seicean A., Alan N., Seicean S., Worwag M., Neuhauser D., Benzel E.C., et al. (2014). Impact of increased body mass index on outcomes of elective spinal surgery. *Spine, 39*(18), 1520-1530.

Seicean S., Seicean A., Alan N., Plana J.C., Budd G.T., & Marwick, T. H. (2013). Cardioprotective effect of beta-adrenoceptor blockade in patients with breast cancer undergoing chemotherapy follow-up study of heart failure. *Circulation: Heart Failure, 6*(3), 420-426.

Seiffge D.J., Hooff R.J., Nolte C.H., Bejot Y., Turc G., Ikenberg B., et al. (2015). Recanalization therapies in acute ischemic stroke patients impact of prior treatment with novel oral anticoagulants on bleeding complications and outcome a pilot study. *Circulation, 132*(13), 1261-1269.

Seitz D.P., Gill S.S., Gruneir A., Austin P.C., Anderson G., Reimer C.L., et al. (2011). Effects of cholinesterase inhibitors on postoperative outcomes of older adults with dementia undergoing hip fracture surgery. *American Journal of Geriatric Psychiatry, 19*(9), 803-813.

Shah M., Tsadok M.A., Jackevicius C.A., Essebag V., Eisenberg M.J., Rahme E., et al. (2014). Warfarin use and the risk for stroke and bleeding in patients with atrial fibrillation undergoing dialysis. *Circulation, 129*(11), 1196-1203.

Shih C.J., Chen H.T., Chao P.W., Kuo S.C., Li S.Y., Yang C.Y., et al. (2016). Angiotensin-converting enzyme inhibitors, angiotensin II receptor blockers and the risk of major adverse cardiac events in patients with diabetes and prior stroke: A nationwide study. *Journal of Hypertension, 34*(3), 567-575.

Shimizu F., Muto S., Taguri M., Ieda T., Tsujimura A., Sakamoto Y., et al. (2017). Effectiveness of platinum-based adjuvant chemotherapy for muscle-invasive bladder cancer: A weighted propensity score analysis. *International Journal of Urology, 24*(5), 367-372.

Shin T.G., Choi J.H., Jo I.J., Sim M.S., Song H.G., Jeong Y.K., et al. (2011). Extracorporeal cardiopulmonary resuscitation in patients with inhospital cardiac arrest: A comparison with conventional cardiopulmonary resuscitation. *Critical Care Medicine, 39*(1), 1-7.

Slade E.P., McCarthy J.F., Valenstein M., Visnic S., & Dixon, L. B. (2013). Cost savings from assertive community treatment services in an era of declining psychiatric inpatient use. *Health Services Research, 48*(1), 195-217.

Sobolev B.G., Fradet G., Kuramoto L., & Rogula, B. (2012). An observational study to evaluate 2 target times for elective coronary bypass surgery. *Medical Care, 50*(7), 611-619.

Spelman T., Kalincik T., Jokubaitis V., Zhang A., Pellegrini F., Wiendl H., et al. (2016). Comparative efficacy of first-line natalizumab vs IFN-beta or glatiramer acetate in relapsing MS. *Neurology: Clinical Practice, 6*(2), 102-115.

Sutter R., De Marchis G.M., Semmlack S., Fuhr P., Ruegg S., Marsch S., et al. (2017). Anesthetics and Outcome in Status Epilepticus: A Matched Two-Center Cohort Study.

TaiSeale M., Wilson C.J., Stone A., Durbin M., & Luft, H. S. (2014). Patients' body mass index and blood pressure over time diagnoses, treatments, and the effects of comorbidities.

Takagi T., Kondo T., Tachibana H., Iizuka J., Omae K., Kobayashi H., et al. (2016). A propensity score-matched comparison of surgical precision obtained by using volumetric analysis between robot-assisted laparoscopic and open partial nephrectomy for T1 renal cell carcinoma: A retrospective non-randomized observational study of initial outcomes. *International Urology and Nephrology, 48*(10), 1585-1591.

Tanabe S., Yasunaga H., Ogawa T., Koike S., Akahane M., Horiguchi H., et al. (2012). Comparison of outcomes after use of biphasic or monophasic defibrillators among out-of-hospital cardiac arrest patients: A nationwide population-based observational study. *Circulation: Cardiovascular Quality and Outcomes, 5*(5), 689-696.

Tang C.H., Chen T.H., Wang C.C., Hong C.Y., Huang K.C., & Sue, Y. M. (2013). Renin-angiotensin system blockade in heart failure patients on long-term haemodialysis in taiwan. *European Journal of Heart Failure, 15*(10), 1194-1202.

Taniguchi T., Morimoto T., Shiomi H., Ando K., Kanamori N., Murata K., et al. (2017). High- versus low-gradient severe aortic stenosis: Demographics, clinical outcomes, and effects of the initial aortic valve replacement strategy on long-term prognosis. *Circulation: Cardiovascular Interventions, 10*(5) (pagination), Arte Number: e004796. ate of Pubaton: 01 May 2017.

Tavakoli H., FitzGerald J.M., Chen W., Lynd L., Kendzerska T., Aaron S., et al. (2017). Ten-year trends in direct costs of asthma: A population-based study. *Allergy: European Journal of Allergy and Clinical Immunology, 72*(2), 291-299.

Teshale E.H., Xing J., Moorman A., Holmberg S.D., Spradling P.R., Gordon S.C., et al. (2016). Higher all-cause hospitalization among patients with chronic hepatitis C: The chronic hepatitis cohort study (CHeCS), 2006-2013. *Journal of Viral Hepatitis, 23*(10), 748-754.

Tribouilloy C., Rusinaru D., Marechaux S., Jeu A., Ederhy S., Donal E., et al. (2012). Increased risk of left heart valve regurgitation associated with benfluorex use in patients with diabetes mellitus: A multicenter study. *Circulation, 126*(24), 2852-2858.

Tsadok M.A., Jackevicius C.A., Essebag V., Eisenberg M.J., Rahme E., Humphries K.H., et al. (2012). Rhythm versus rate control therapy and subsequent stroke or transient ischemic attack in patients with atrial fibrillation. *Circulation, 126*(23), 2680-2687.

Vaara S.T., Pettila V., Kaukonen K.M., Bendel S., Korhonen A.M., Bellomo R., et al. (2014). The attributable mortality of acute kidney injury: A sequentially matched analysis. *Critical Care Medicine, 42*(4), 878-885.

Van Der Wal G., Brinkman S., Bisschops L.L.A., Hoedemaekers C.W., Van Der Hoeven J.G., De Lange D.W., et al. (2011). Influence of mild therapeutic hypothermia after cardiac arrest on hospital mortality. *Critical Care Medicine, 39*(1), 84-88.

Van Hout F.M.A., Hogervorst E.K., Rosseel P.M.J., Van Der Bom J.G., Bentala M., Van Dorp E.L.A., et al. (2017). Does a platelet transfusion independently affect bleeding and adverse outcomes in cardiac surgery?. *Anesthesiology, 126*(3), 441-449.

Van Lier F., Van Der Geest P.J., Hoeks S.E., Van Gestel Y.R.B.M., Hol J.W., Sin D.D., et al. (2011). Epidural analgesia is associated with improved health outcomes of surgical patients with chronic obstructive pulmonary disease. *Anesthesiology, 115*(2), 315-321.

Wada T., Yasunaga H., Horiguchi H., Matsubara T., Fushimi K., Nakajima S., et al. (2016). Outcomes of argatroban treatment in patients with atherothrombotic stroke : Observational nationwide study in japan. *Stroke, 47*(2), 471-476.

Wahl A., Juni P., Mono M.L., Kalesan B., Praz F., Geister L., et al. (2012). Long-term propensity score-matched comparison of percutaneous closure of patent foramen ovale with medical treatment after paradoxical embolism. *Circulation, 125*(6), 803-812.

Walkey A.J., & Wiener, R. S. (2011). Utilization patterns and patient outcomes associated with use of rescue therapies in acute lung injury. *Critical Care Medicine, 39*(6), 1322-1328.

Wang V., Liu C.F., Bryson C.L., Sharp N.D., & MacIejewski, M. L. (2011). Does medication adherence following a copayment increase differ by disease burden?. *Health Services Research, 46*(6 PART 1) (pp 1963-1985), ate of Pubaton: eember 2011.

Wasko M.C.M., Dasgupta A., Hubert H., Fries J.F., & Ward, M. M. (2013). Propensity-adjusted association of methotrexate with overall survival in rheumatoid arthritis. *Arthritis and Rheumatism, 65*(2), 334-342.

Weisberg D.F., Gordon K.S., Barry D.T., Becker W.C., Crystal S., Edelman E.J., et al. (2015). Long-term prescription of opioids and/or benzodiazepines and mortality among HIV-infected and uninfected patients. *Journal of Acquired Immune Deficiency Syndromes, 69*(2), 223-233.

Weiss S.L., Fitzgerald J.C., Balamuth F., Alpern E.R., Lavelle J., Chilutti M., et al. (2014). Delayed antimicrobial therapy increases mortality and organ dysfunction duration in pediatric sepsis. *Critical Care Medicine, 42*(11), 2409-2417.

Wiewel M.A., Van Vught L.A., Scicluna B.P., Hoogendijk A.J., Frencken J.F., Zwinderman A.H., et al. (2017). Prior use of calcium channel blockers is associated with decreased mortality in critically ill patients with sepsis: A prospective observational study. *Critical Care Medicine, 45*(3), 454-463.

Williams S.B., Lei Y., Nguyen P.L., Gu X., Lipsitz S.R., Yu H.Y., et al. (2012). Comparative effectiveness of cryotherapy vs brachytherapy for localised prostate cancer. *BJU International, 110*, E92-E98.

Yang C.Y., Chang Z.F., Chau Y.P., Chen A., Yang W.C., Yang A.H., et al. (2015). Circulating Wnt/beta-catenin signalling inhibitors and uraemic vascular calcifications. *Nephrology Dialysis Transplantation, 30*(8), 1356-1363.

Yang J.H., Hahn J.Y., Song P.S., Song Y.B., Choi S.H., Choi J.H., et al. (2014). Percutaneous coronary intervention for nonculprit vessels in cardiogenic shock complicating ST-segment elevation acute myocardial infarction. *Critical Care Medicine, 42*(1), 17-25.

Yuan Q., Cheng H., Yang C., Wang Y., Wang E., Qiu B., et al. (2016). Is intracranial pressure monitoring of patients with diffuse traumatic brain injury valuable? an observational multicenter study. *Neurosurgery, 78*(3), 361-368.

Zhou C., Selles R.W., Slijper H.P., Feitz R., van Kooij Y., Moojen T.M., et al. (2016). Comparative effectiveness of percutaneous needle aponeurotomy and limited fasciectomy for dupuytren's contracture: A multicenter observational study. *Plastic and Reconstructive Surgery, 138*(4), 837-846.

Zusterzeel R., Spatz E.S., Curtis J.P., Sanders W.E., Selzman K.A., Pina I.L., et al. (2015). Cardiac resynchronization therapy in women versus men: Observational comparative effectiveness study from the national cardiovascular data registry. *Circulation: Cardiovascular Quality and Outcomes, 8*, S4-S11.

Zweig T., Enke J., Mannion A.F., Sobottke R., Melloh M., Freeman B.J.C., et al. (2017). Is the duration of pre-operative conservative treatment associated with the clinical outcome following surgical

decompression for lumbar spinal stenosis? A study based on the spine tango registry. *European Spine Journal, 26*(2), 488-500.

Appendix C – Chapter 3

Appendix C.1: Percentage of completeness of analysis variables

Variable	Experiment 1 (%)	Experiment 2 (%)	Experiment 3 (%)
Age (2 – 59 months)	99.7	99.5	99.3
Indrawing (present/absent)	100.0	96.3	95.8
Level of consciousness – AVPU	_	95.5	92.3
Central cyanosis	_	95.9	95.1
Grunting	_	94.2	93.8
Ability to drink	_	91.2	91.5
Gender (male/female)	99.6	99.0	98.7
Cough duration (days)	84.9	83.4	83.2
Crackles (present/absent)	97.4	94.7	92.6
Weight (Kg)	96.3	96.0	94.8
Pallor (0, +, +++)	96.7	94.5	93.9
Capillary refill	83.3	78.0	78.2
Fever (present/absent)	98.2	97.6	96.5
Temperature	94.1	92.6	91.4
Convulsions (present/absent)	96.3	94.3	93.7
Vomiting (yes/no)	97.1	95.2	94.3
Referral (yes/no)	83.3	73.6	73.4
Length of illness (days)	98.4	98.0	97.3
Thrush (present/absent)	90.4	83.9	78.2
Quinine/artesunate (prescribed/not prescribed)	100.0	100.0	100.0
Wheeze (present/absent)	97.1	94.5	90.6
Respiratory rate	87.4	85.4	83.5
IV fluid prescription	100.0	100.0	100.0
Outcome (died/alive)	99.5	99.2	98.8

Correctness of penicillin and gentamicin dosing

It was also examined if the patients received correct dosages of penicillin and gentamicin: For penicillin, a dose of 40,000 – 60,000 I.U/Kg was considered normal and for gentamicin, a dose of 6 – 9 mg/Kg. These were +/- 20% of recommended dosages in the Kenyan paediatric protocols. Majority of the patients were prescribed normal dosages of penicillin and gentamicin (see **Appendix C.2**).

Appendix C.2: Correctness of penicillin and gentamicin prescription

	Experiment 1		Experiment 2		Experiment 3	
	Penicillin	Gentamicin	Penicillin	Gentamicin	Penicillin	Gentamicin
Under dose	3%	10%	3%	12%	4%	12%
Normal	93%	87%	92%	85%	91%	84%
Over dose	4%	3%	5%	3%	5%	4%

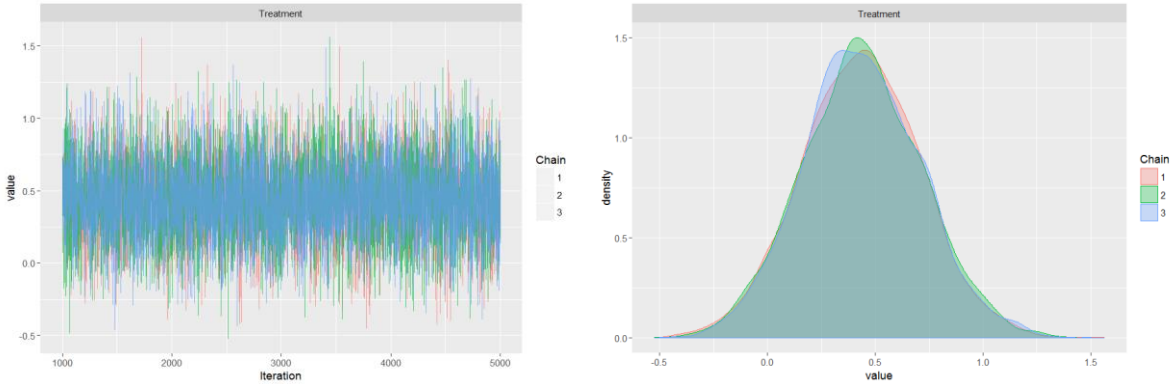
Appendix C.3: Bayesian model convergence diagnostics

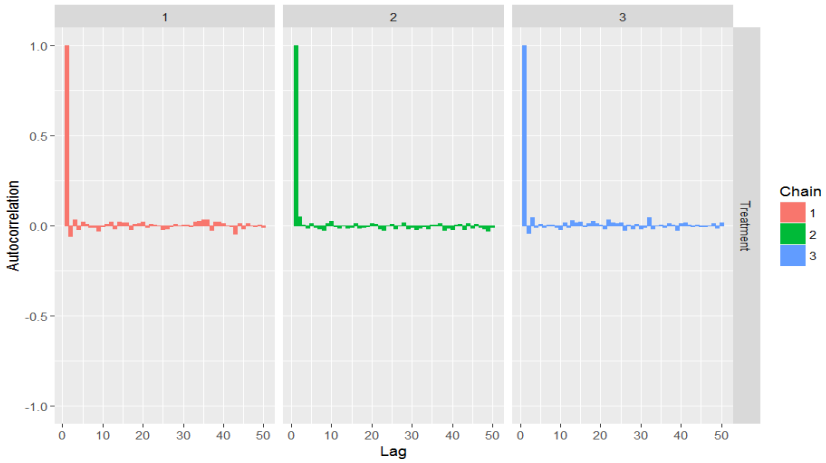
In the Bayesian framework, how well the posterior distributions were explored were examined using trace (history) and density plots of sampled values – and convergence to the target distribution assessed using Geweke plots and potential scale reduction factors (PSRF). Geweke plots subdivide the chain of sampled values into bins – and by default 10 bins are selected, and standardises the difference between sampled values in the first bin and those in the tenth bin. If the standardised difference falls outside -2 and +2 this would be an indication of non-convergence of the chain (184). On the other hand, since these analyses used 3 chains to sample values for a given parameter, PSRF was used to investigate the ratio of variance of within chain sampled values to the variance of between chain sampled values. If this ratio is greater than 1 then this would be an indication of non – convergence (185).

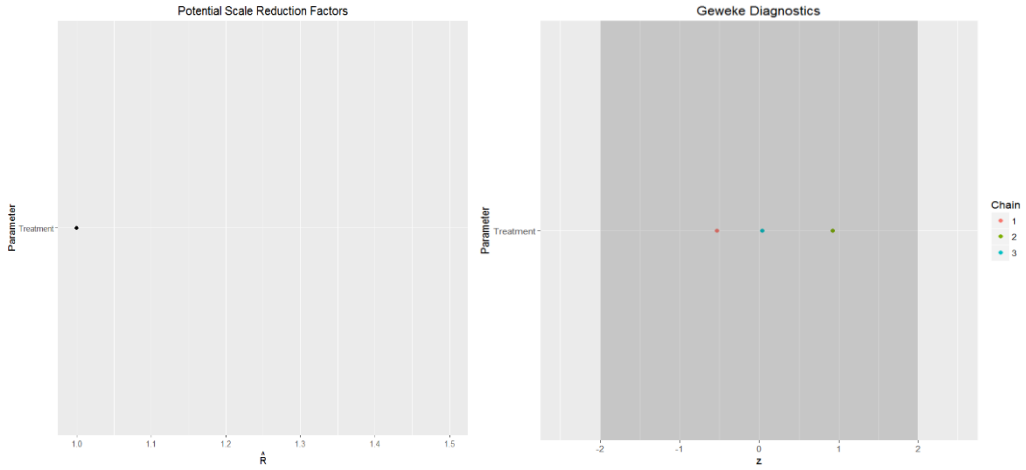
The Bayesian models in experiments 1 – 3 used 5000 iterations with burn-in of 1000. Even though multivariable regression models were used, here I present the diagnostics related to the estimated treatment effect parameters for illustrative purposes of convergence. From the plots presented below, convergence for all the models was satisfactory. The monitored coefficients were in terms of log RR.

i) Experiment 1

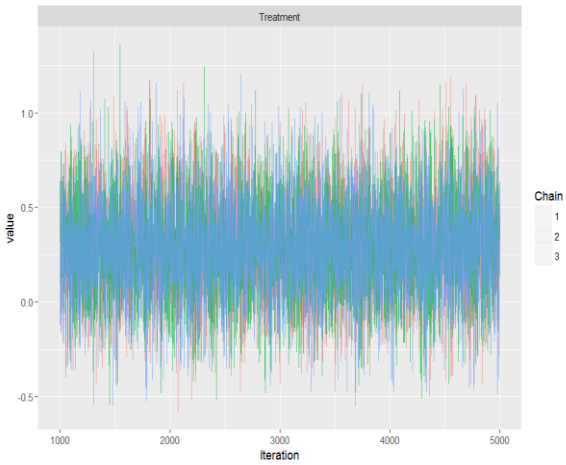
Regression without PS adjustments

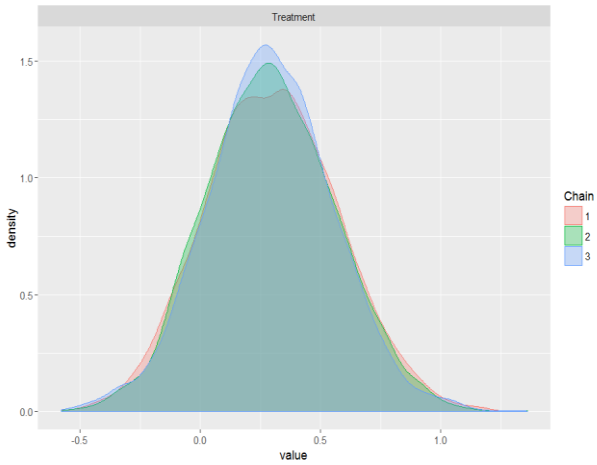


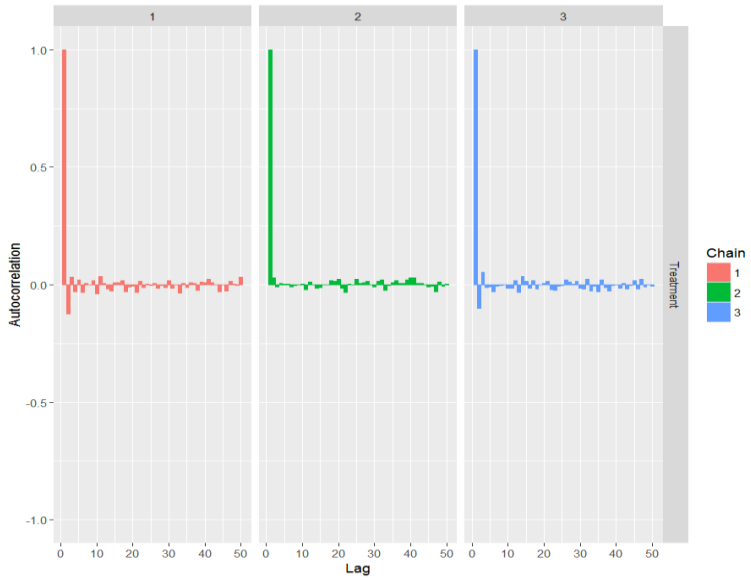


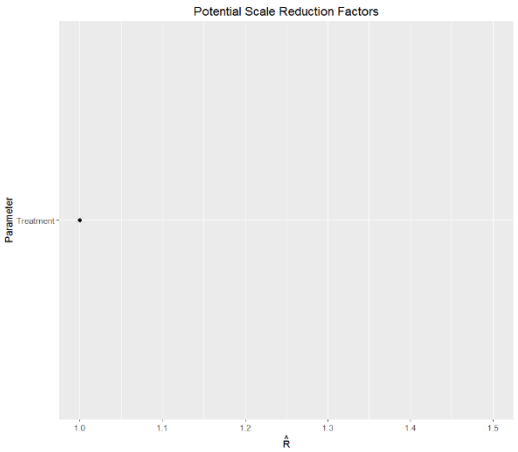


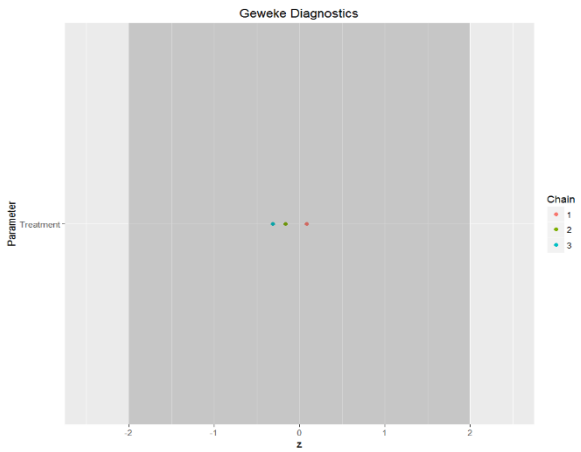
Regression with PS weighted data





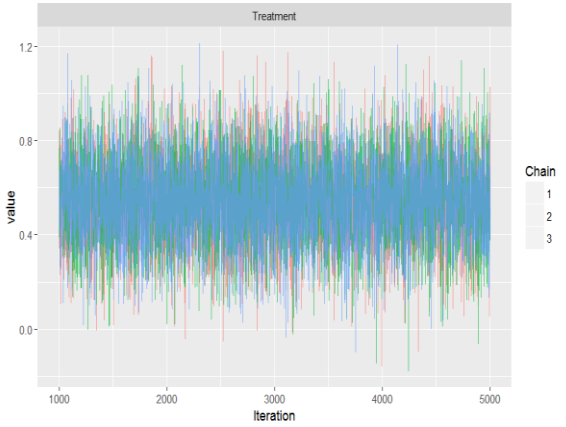


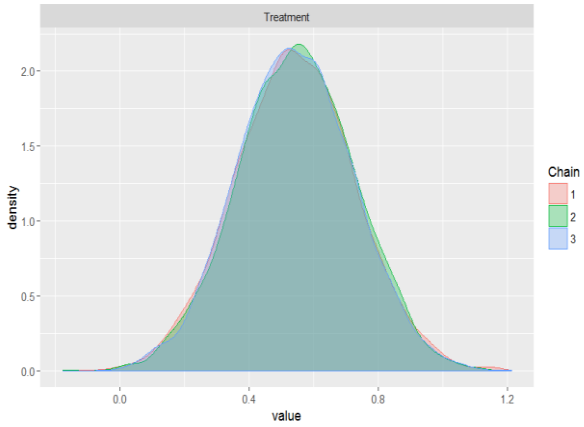


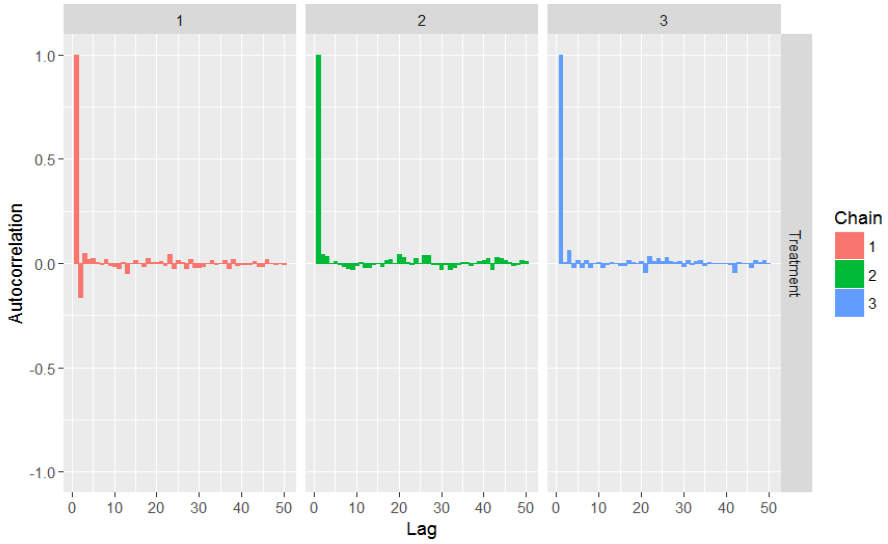


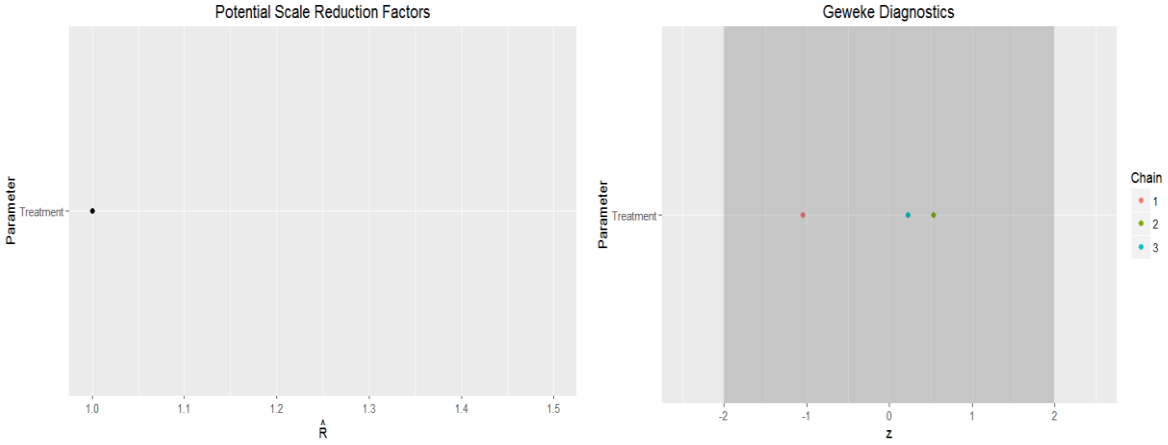
ii) Experiment 2

Regression without PS adjustments

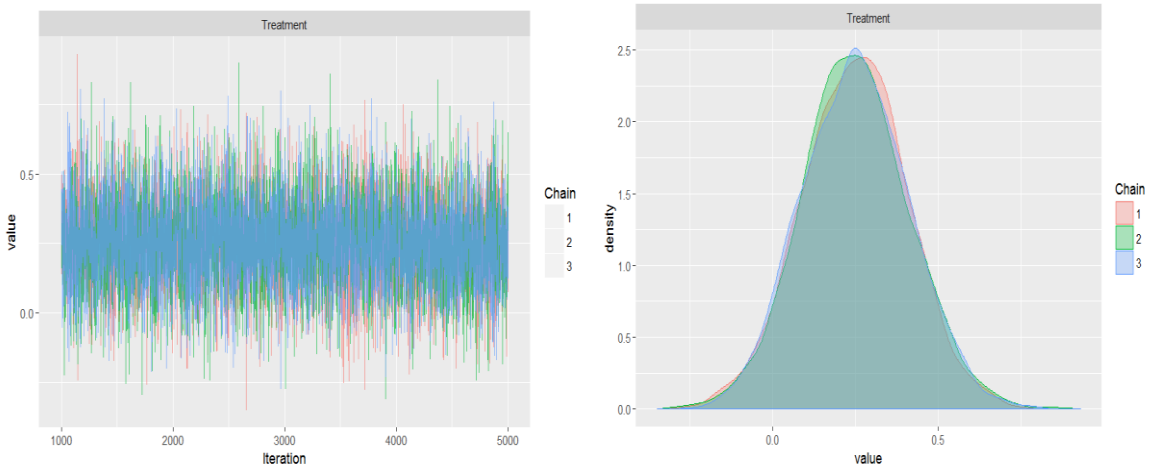


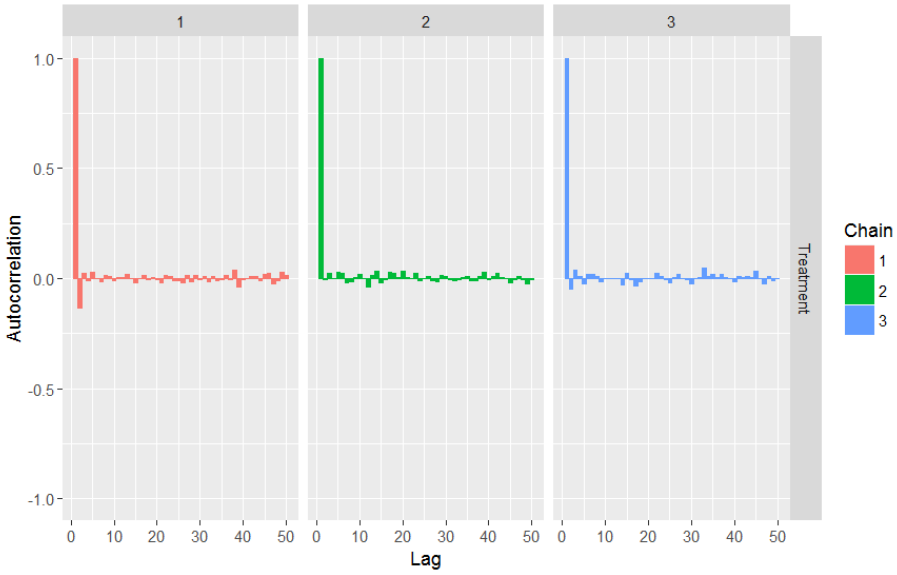


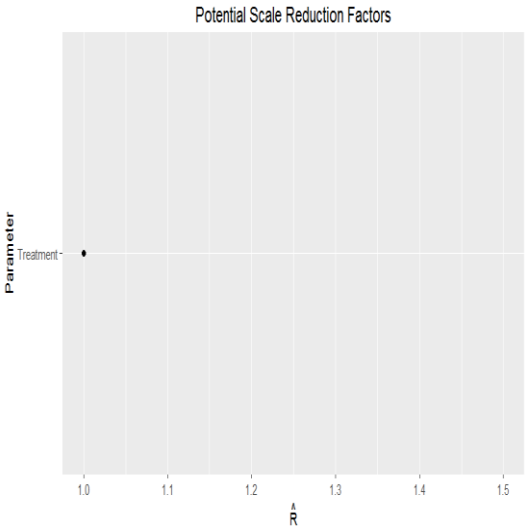


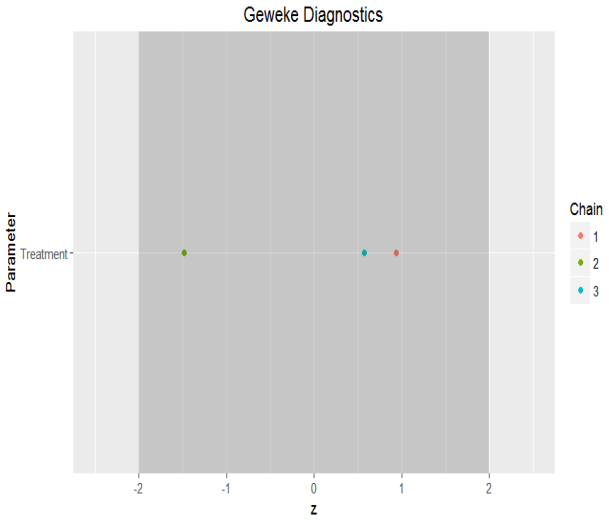


Regression with PS weighted data



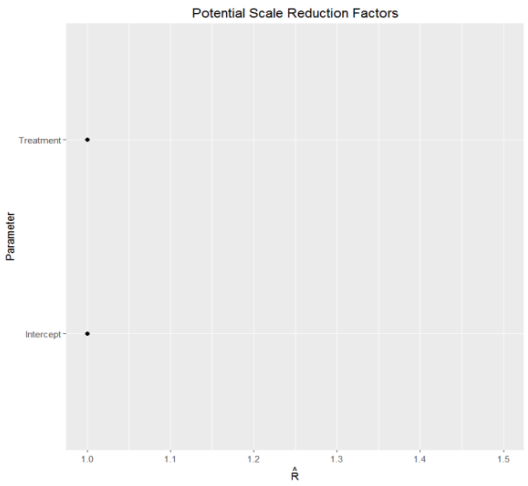


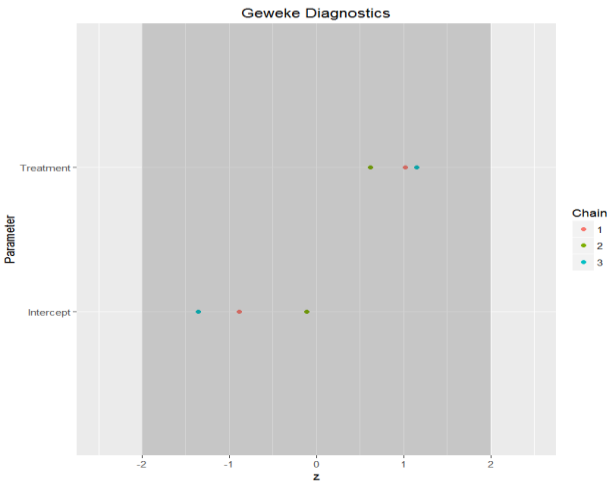




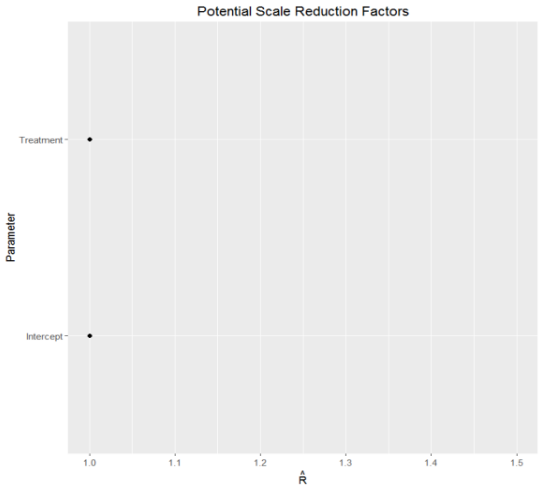
iii) Experiment 3

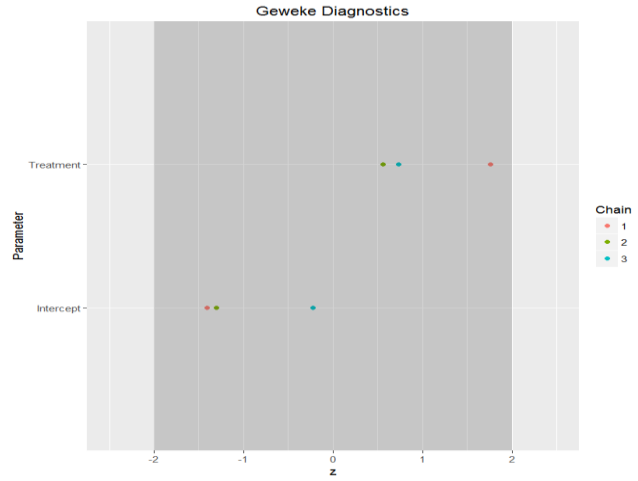
Subclass 1



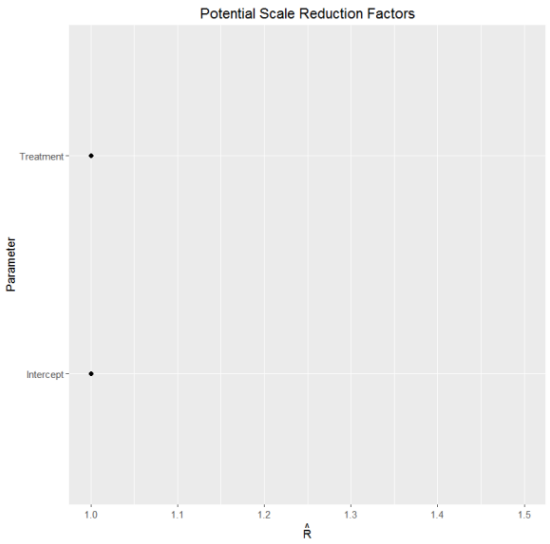


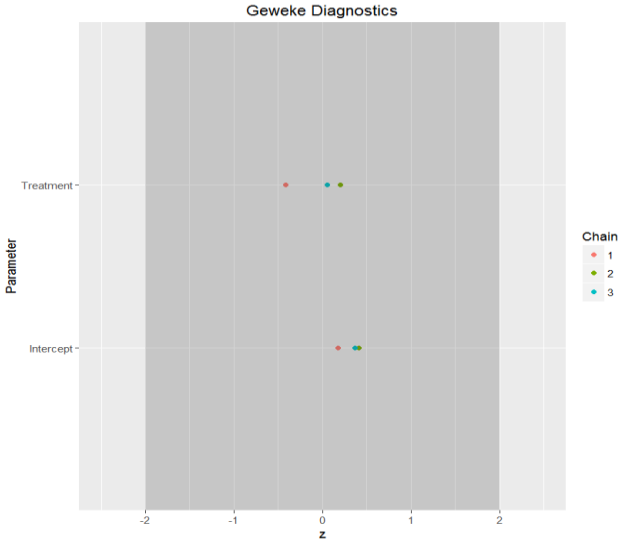
Subclass 2



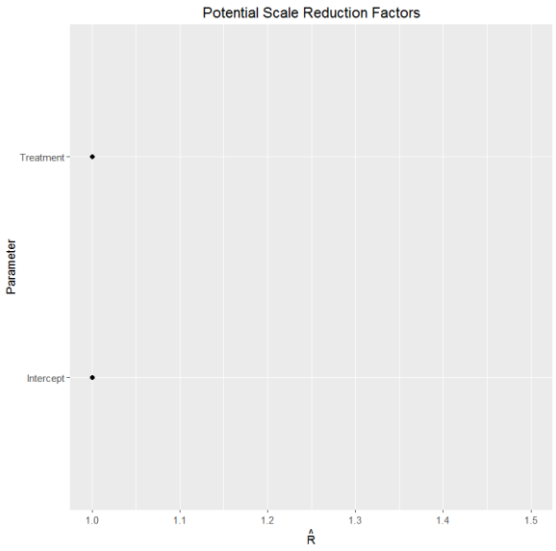


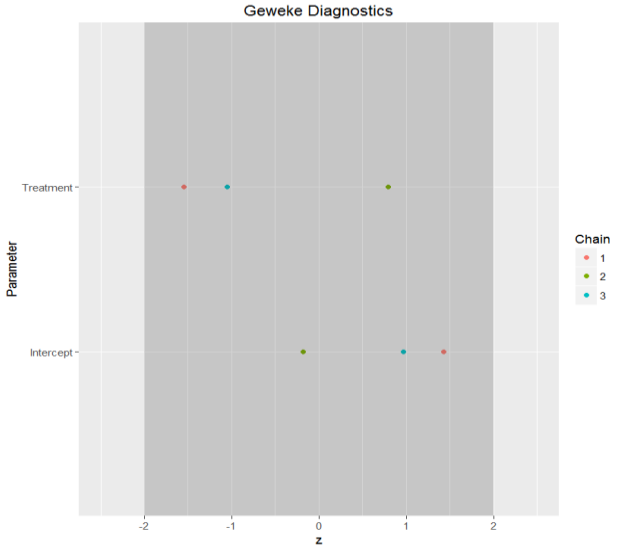
Subclass 3



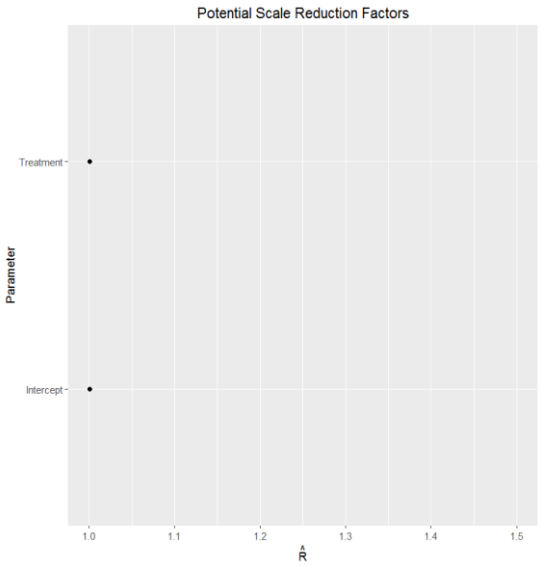


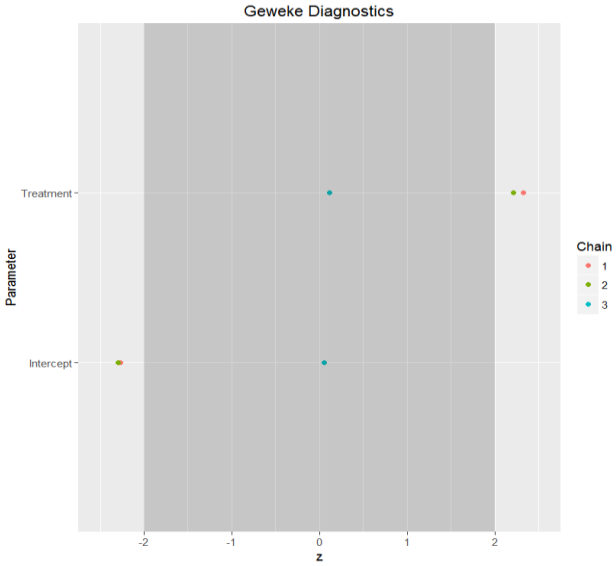
Subclass 4





Subclass 5


Potential Scale Reduction Factors
Treatment
Intercept
Parameter
1.0
1.1
1.2
1.3
1.4
1.5
R


Geweke Diagnostics
Treatment
Intercept
Parameter
-2
-1
0
1
2
z
Chain
1
2
3

Appendix D – Chapter 4

Appendix D.1: Percent in bias reduction per variable – experiment 1

Variables	PS weighting				PS optimal full matching				PS sub – classification			
	MIA	MIAY	MIW	MIWY	MIA	MIAY	MIW	MIWY	MIA	MIAY	MIW	MIWY
Child sex	74.2	86.4	72.2	84.5	77.5	75.5	-385.4	-152.5	39.1	13.2	60.6	-15.7
Age (months)	-12.1	-16.4	-15.0	-5.8	69.6	69.1	-53.2	-69.1	64.1	3.4	-46.4	-51.5
Weight (Kg)	17.0	22.6	15.3	31.0	47.6	56.0	-15.8	-28.3	61.4	73.5	-23.6	-17.8
Pallor	31.8	67.2	44.4	63.8	39.5	42.8	-6.4	13.5	3.2	34.2	-68.6	-26.0
Respiratory rate	90.3	81.7	94.2	88.1	77.9	84.1	68.2	82.2	75.7	71.6	53.0	62.1
Difficulty in breathing	90.6	84.2	92.8	87.7	75.3	72.8	-113.1	-106.6	84.7	83.7	85.2	84.1
Capillary refill	74.9	76.2	85.2	88.0	97.7	97.1	37.4	45.7	88.9	92.2	73.3	70.7
Fever	-477.9	-228.7	-512.8	-116.3	-134.2	48.0	-1569.4	-801.1	-369.8	-41.7	-925.9	-440.5
Convulsions	78.1	84.6	54.2	-2.0	-467.5	-346.2	-15.0	24.3	-420.1	-476.9	-134.6	-204.5
Vomiting	-74.4	-231.6	-352.0	-657.2	-419.7	-903.0	-1289.1	-2455.1	-354.4	-587.6	-269.3	-539.0
Cough duration (days)	76.6	80.2	65.4	76.4	-12.5	22.9	-80.8	-130.2	-66.1	-59.5	7.0	-3.6
WAZ score	49.4	49.0	50.9	51.3	98.9	98.6	87.2	85.4	97.8	98.0	87.3	87.7
Co - morbidities	65.3	61.1	65.9	66.2	30.3	27.2	82.9	92.9	79.8	76.8	18.0	-19.8
Crackles	68.1	75.2	71.6	72.3	47.3	61.1	-53.5	-63.8	67.3	67.6	-38.6	-5.0
Hospital referral	74.4	62.6	87.9	85.5	74.8	40.8	-12.5	-33.9	54.0	48.5	26.6	51.1
Length of illness	80.4	78.3	-47.5	53.3	-357.0	-21.0	-393.1	-186.2	-525.7	-74.1	-139.5	-39.2
Thrush	4.4	22.8	-14.2	19.3	61.7	74.6	-44.0	-8.5	-1.4	26.3	-90.8	-106.6
Wheeze	86.1	59.2	69.5	69.1	39.6	-6.8	-28.2	-87.5	-17.4	-52.0	52.9	10.7

*Positive values indicate reduced bias while negative shows increased bias.

Appendix D.2: Percent in bias reduction per variable – experiment 2

Variables	PS weighting				PS optimal full matching				PS sub – classification			
	MIA	MIAY	MIW	MIWY	MIA	MIAY	MIW	MIWY	MIA	MIAY	MIW	MIWY
Child sex	47.9	48.2	55.2	47.9	98.1	86.0	-189.1	-209.9	67.2	52.6	-44.2	-40.9
Age (months)	-12.6	-12.8	-7.6	5.3	91.7	85.8	-16.2	-39.1	76.7	70.0	-16.7	-22.3
Weight (Kg)	6.3	8.6	8.6	18.3	81.2	64.7	25.1	-8.0	78.1	65.9	-9.7	-10.4
Pallor	73.3	79.1	72.3	83.9	92.1	72.6	66.0	44.7	78.0	76.0	54.7	68.7
Respiratory rate	79.4	74.6	76.7	63.3	106.8	91.8	34.4	28.9	43.1	45.5	42.4	32.7
Difficulty in breathing	95.6	93.7	92.3	90.7	79.9	72.2	31.1	8.4	81.8	86.0	81.0	84.9
Capillary refill	90.1	88.7	83.4	84.9	93.9	98.0	47.6	54.8	91.3	90.6	65.6	64.9
Fever	26.5	25.0	36.9	31.4	76.6	98.5	-58.5	-55.7	36.8	36.1	-39.5	-56.0
Convulsions	9.7	41.3	28.9	87.6	-519.7	-1194.1	-514.5	-307.5	-288.3	-135.1	-356.0	-692.2
Vomiting	-2210.8	-1802.7	-2094.9	-1791.9	-3228.4	-1304.0	-4299.7	-4712.7	-1729.1	-822.2	-5304.4	-3742.1
Cough duration (days)	-39.9	-41.9	-33.4	-24.0	31.6	26.1	-2698.9	-2533.1	82.3	72.5	-130.0	-101.5
WAZ score	47.1	46.0	49.7	49.5	98.1	99.7	94.0	90.9	107.8	98.0	94.1	93.3
Co - morbidities	81.6	81.5	82.4	82.5	29.7	45.6	89.7	81.2	86.2	86.8	38.3	23.1
Crackles	89.6	88.7	81.6	83.0	63.7	70.1	19.6	48.6	74.5	76.9	59.9	68.9
Hospital referral	77.5	79.9	91.9	90.0	89.9	99.8	57.2	31.1	71.1	76.3	76.3	70.0
Length of illness (days)	15.5	7.7	11.0	22.4	15.2	1.9	-1864.2	-1788.8	49.8	62.3	-58.5	-30.8
Thrush	-637.4	-112.1	-1472.3	74.1	-1003.9	44.4	-1870.2	-42.6	-1168.6	21.8	-2889.3	-634.3
Wheeze	53.9	67.9	57.6	83.2	-15.3	-61.3	-1302.8	-1117.8	-53.2	-15.3	-7.7	-41.8
Indrawing	23.0	38.5	6.7	11.5	9.3	54.5	-376.8	-229.0	-12.4	-0.2	-32.2	-35.0
Oxygen order	95.3	94.6	91.2	93.9	90.7	92.0	-12.1	-2.0	81.4	80.4	61.9	57.6
Ability to drink	66.6	72.9	81.5	69.7	7.9	-4.1	15.7	-65.4	16.6	12.6	39.1	-18.1
Central cyanosis	34.1	43.0	32.6	43.9	124.2	84.2	-2109.6	-1773.2	79.3	69.9	-35.4	-58.8
Grunting	96.7	97.1	91.2	91.6	100.0	99.3	44.2	55.2	98.3	95.6	58.6	64.5

Appendix D.3: Percent in bias reduction per variable – experiment 3

Variables	PS sub-classification			
	MIA	MIAY	MIW	MIWY
Child sex	-141.6	-97.1	-142.9	-158.1
Age (months)	77.9	78.0	73.9	67.2
Weight (Kg)	71.7	70.0	64.4	68.4
Pallor	56.7	67.2	25.8	56.3
Respiratory rate	80.0	79.8	85.7	86.8
Difficulty in breathing	94.8	94.4	92.5	87.5
Capillary refill	82.9	83.7	75.1	80.0
Fever	49.9	58.2	72.1	35.3
Convulsions	39.5	25.6	32.3	11.1
Vomiting	56.0	61.4	32.6	26.1
Cough duration (days)	-95.2	-114.1	-275.3	-147.5
WAZ score	97.3	97.0	95.4	95.2
Co – morbidities	-58.8	-55.5	6.1	34.7
Crackles	90.5	90.9	88.7	89.0
Hospital referral	80.1	76.1	78.2	76.8
Length of illness (days)	-59.8	-129.3	-206.9	-164.8
Thrush	-295.9	-771.0	-438.4	-735.0
Wheeze	39.6	39.1	55.4	37.8
Indrawing	73.9	74.6	76.0	72.6
Oxygen order	91.3	90.5	83.8	81.1
Ability to drink	82.5	86.3	85.7	86.9
Central cyanosis	85.3	84.9	46.6	57.7
Grunting	97.5	97.7	89.2	84.6

Appendix D.4: Experiment 1: Treatment effect estimates under different approaches (within PS estimation)

	Optimal full matching		**PS weighting**		**PS sub-classification**	
	MIW	MIWY	MIW	MIWY	MIW	MIWY
Unrestricted	1.28 [0.79, 2.07]	1.05 [0.65, 1.71]	1.46 [0.85, 2.44]	1.23 [0.69, 2.20]	1.46 [0.82, 2.61]	1.31 [0.72, 2.38]
0 – 100	1.28 [0.79, 2.07]	1.05 [0.65, 1.71]	1.46 [0.85, 2.44]	1.23 [0.69, 2.20]	1.46 [0.82, 2.61]	1.31 [0.72, 2.38]
1 – 99	1.26 [0.92, 1.72]	1.17 [0.79, 1.76]	1.43 [0.90, 2.27]	1.17 [0.66, 2.06]	1.40 [0.79, 2.47]	1.28 [0.72, 2.29]
2.5 – 97.5	1.25 [1.00, 1.54]	1.13 [0.73, 1.72]	1.38 [0.88, 2.16]	1.12 [0.64, 1.97]	1.35 [0.77, 2.37]	1.23 [0.70, 2.15]
5 – 95	1.22 [0.98, 1.50]	1.10 [0.71, 1.67]	1.39 [0.90, 2.15]	1.12 [0.65, 1.93]	1.34 [0.77, 2.32]	1.30 [0.75, 2.25]

Appendix D.5: Experiment 1: Treatment effect estimates under different approaches (across PS estimation)

	Optimal full matching		**PS weighting**		**PS sub-classification**	
	MIA	MIAY	MIA	MIAY	MIA	MIAY
Unrestricted	0.83 [0.50, 1.32]	0.97 [0.58, 1.62]	1.18 [0.61, 2.26]	1.07 [0.58, 1.99]	1.31 [0.70, 2.44]	1.38 [0.73, 2.61]
0 – 100	0.83 [0.50, 1.32]	0.97 [0.58, 1.62]	1.18 [0.61, 2.26]	1.07 [0.58, 1.99]	1.31 [0.70, 2.44]	1.38 [0.73, 2.61]
1 – 99	0.79 [0.48, 1.29]	0.88 [0.53, 1.45]	1.15 [0.60, 2.20]	0.98 [0.54, 1.78]	1.20 [0.65, 2.23]	1.28 [0.68, 2.41]
2.5 – 97.5	0.75 [0.46, 1.23]	0.79 [0.48, 1.30]	1.08 [0.57, 2.05]	0.96 [0.53, 1.73]	1.22 [0.67, 2.21]	1.30 [0.69, 2.44]
5 – 95	0.70 [0.43, 1.14]	0.72 [0.45, 1.16]	1.06 [0.56, 1.99]	0.96 [0.54, 1.71]	1.24 [0.69, 2.23]	1.33 [0.71, 2.49]

Appendix D.6: Experiment 2: Treatment effect estimates under different approaches (within PS estimation)

	Optimal full matching		**PS weighting**		**PS sub-classification**	
	MIW	MIWY	MIW	MIWY	MIW	MIWY
Unrestricted	0.92 [0.69, 1.20]	0.88 [0.66, 1.13]	1.04 [0.76, 1.40]	1.01 [0.70, 1.26]	1.62 [0.99, 2.53]	1.63 [0.98, 2.51]
0 – 100	0.92 [0.69, 1.20]	0.88 [0.66, 1.13]	1.04 [0.76, 1.40]	1.01 [0.70, 1.26]	1.62 [0.99, 2.53]	1.63 [0.98, 2.51]
1 – 99	0.91 [0.71, 1.17]	0.86 [0.67, 1.11]	1.04 [0.78, 1.38]	1.00 [0.72, 1.39]	1.58 [1.00, 2.50]	1.60 [0.99, 2.57]
2.5 – 97.5	0.89 [0.72, 1.10]	0.85 [0.68, 1.06]	1.03 [0.76, 1.39]	1.00 [0.74, 1.35]	1.50 [0.95, 2.35]	1.52 [0.93, 2.39]
5 – 95	0.82 [0.67, 1.00]	0.85 [0.70, 1.03]	1.05 [0.84, 1.31]	1.00 [0.76, 1.31]	1.46 [0.96, 2.22]	1.49 [0.95, 2.39]

Appendix D.7: Experiment 2: Treatment effect estimates under different approaches (across PS estimation)

	Optimal full matching		**PS weighting**		**PS sub-classification**	
	MIA	MIAY	MIA	MIAY	MIA	MIAY
Unrestricted	1.12 [0.89, 1.43]	1.11 [0.87, 1.36]	1.02 [0.78, 1.38]	0.99 [0.72, 1.27]	1.49 [0.89, 2.05]	1.52 [0.91, 2.25]
0 – 100	1.12 [0.89, 1.43]	1.11 [0.87, 1.36]	1.02 [0.78, 1.38]	0.99 [0.72, 1.27]	1.49 [0.89, 2.05]	1.52 [0.91, 2.25]
1 – 99	1.11 [0.91, 1.36]	1.11 [0.89, 1.38]	1.01 [0.8, 1.27]	0.99 [0.74, 1.32]	1.38 [0.83, 2.27]	1.41 [0.86, 2.31]
2.5 – 97.5	1.07 [0.91, 1.26]	1.03 [0.85, 1.25]	1.01 [0.83, 1.23]	1.00 [0.77, 1.30]	1.43 [0.90, 2.26]	1.45 [0.92, 2.29]
5 – 95	1.06 [0.92, 1.22]	1.05 [0.90, 1.23]	1.01 [0.86, 1.19]	1.00 [0.79, 1.26]	1.36 [0.89, 2.08]	1.42 [0.98, 2.16]

Appendix D.8: Experiment 3: Pooled treatment effect estimates under different approaches

	MIW	**MIWY**	**MIA**	**MIAY**
Unrestricted	1.51 [0.77, 1.92]	1.63 [0.98, 2.51]	1.60 [1.25, 2.05]	1.57 [1.20, 1.92]
0 – 100	1.51 [0.77, 1.92]	1.62 [0.98, 2.51]	1.60 [1.25, 2.03]	1.57 [1.21, 1.92]
1 – 99	1.48 [0.77, 2.83]	1.56 [0.95, 2.59]	1.58 [1.28, 1.95]	1.57 [1.24, 1.98]
2.5 – 97.5	1.45 [0.77, 2.72]	1.55 [0.97, 2.48]	1.50 [1.24, 1.81]	1.49 [1.21, 1.84]
5 – 95	1.40 [0.75, 2.63]	1.45 [0.94, 2.23]	1.49 [1.27, 1.75]	1.46 [1.22, 1.74]

Appendix E – Chapter 5

Appendix E.1: Analysis protocol

Introduction

Diarrhoea is one of the leading causes of under-five mortality in Kenya (186). Oral rehydration solution (ORS) have been shown to be effective in treating diarrhoea (187). A number of primary and meta-analyses of Zinc trials including (25, 136-139) demonstrated that addition of Zinc to ORS treatments shortens the duration of diarrhoea – and as a result informed the WHO diarrhoea treatment guidelines for children in developing countries (188). These recommendations have been adopted in the Kenyan national treatment guidelines for children with diarrhoea and all levels of severity of dehydration (28).

The use of Zinc is based in part on the fact that children in developing countries may often be Zinc deficient to some extent due to poor nutrition (136, 189).The prevalence rates of Zinc deficiency vary by country (140) and this has potentially contributed to between country variations of Zinc effectiveness as demonstrated in a systematic review by Patel (2012) (26). Even though the use of Zinc supplementation has been recommended by WHO, majority of healthcare workers in developing countries are slow at adopting it (190). Additional evidence from lower and middle income countries (LMIC) on the value of Zinc, particularly in non-trial settings that may carefully select patients and in hospitalised patients rarely included in these studies, may help encourage healthcare workers to use Zinc routinely in treatment for diarrhoea, even those in hospitals. In addition, a recent systematic review by Lazerrini and Wanzira (25) on effectiveness of Zinc concluded that there is still insufficient evidence to

support the benefits of Zinc in children younger than six months, those with acute diarrhoea and in well-nourished children.

This heterogeneity in Zinc use outcomes between places and insufficient evidence on effectiveness raises the need to examine Zinc effectiveness through community pragmatic trials in LMIC. As discussed in the thesis background, trials including pragmatic trials are usually costly and require time to plan and execute. An alternative would be to use routine datasets to examine the effects of Zinc treatment in analyses that aim to approximate RCT (whose limitations have also been discussed in the background which include non – random treatment allocation and missing data).

Objective

The primary focus of this analysis was to examine the effectiveness of Zinc supplementation in reducing time (in days) to discharge for children (aged 1 – 5 and 6 – 59 months) admitted with diarrhoea in Kenyan hospitals. Systematic reviews have included data on mostly trials conducted in hospital settings and did not find an effect on mortality but do suggest a valuable effect on duration of diarrhoea (25). As CIN patient population was hospitalised, this analysis therefore focused on length of stay (to model time to getting discharged alive) likely to be influenced by duration of symptoms. In secondary analyses within each age group, this analysis aimed to examine the effectiveness of Zinc amongst those classified as either severely-moderately malnourished or well nourished. Further, as part of exploratory analyses, time to experiencing inpatient mortality

(in those who received Zinc versus those who did not) was examined in the two age categories.

Study Design

This was an observational study based on analyses of data routinely collected from hospital paediatric wards in Kenya's CIN. The CIN provided, as in this case, the basis for cheaply and quickly studying effectiveness for a wider range of treatments – and as a result informed thinking on the external validity of available efficacy trials and provided comparative effectiveness data where suitable trials are lacking.

The design process broadly consisted of the following aspects:

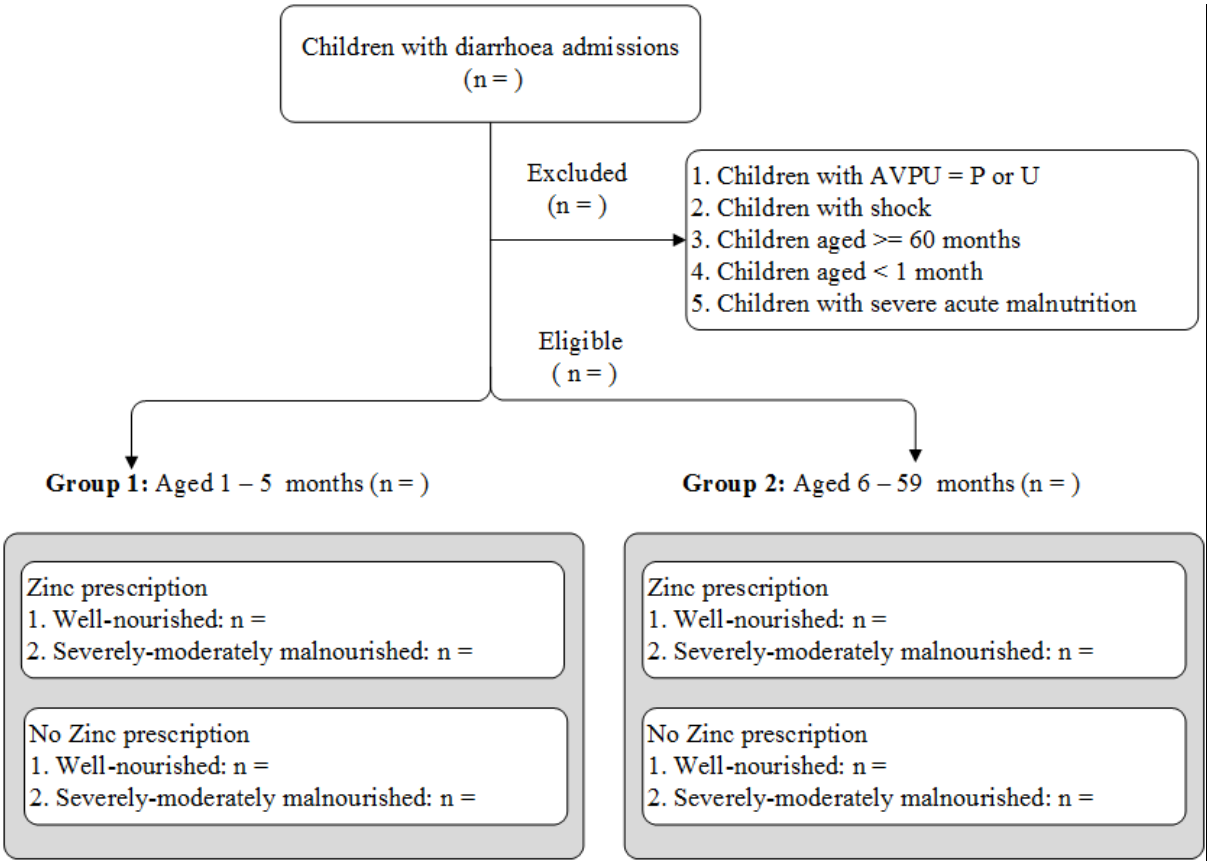
f) Definition of inclusion and exclusion criteria.
g) Understanding the diarrhoea diagnosis and treatment assignment processes for Zinc[31].
h) Verification of sample size if sufficient for any meaningful analyses.
i) Creation of comparable treatment arms – which analytically addressed the challenges of non – random treatment assignment and missing data.

(i) Inclusion and exclusion criteria

The study included children with a diagnosis of diarrhoea at admission, while children with AVPU[32] = P or U, and those with shock were excluded (see figure 1).

[31] Understanding the treatment assignment process is useful in determining key variables for analysis.
[32] AVPU – measures the level of consciousness: A – Alert, V – Verbal , P – Pain and U – Unresponsive.

Figure 1: *Inclusion and exclusion of study population*



(ii) Understanding the diarrhoea diagnosis and treatment assignment processes.

Diarrhoea treatment happens in three steps: First, clinical signs are assessed; these then are integrated to assess the extent of dehydration – which later informs appropriate treatment. Dehydration in the Kenyan treatment guidelines has been classified into shock, severe, some and no dehydration. Children with shock present with weak/absent pulse, AVPU<A, cold hands plus temperature gradient, capillary refill > 3 seconds, sunken eyes and slow skin pinch. In case of these signs then a child should be treated for hypoglycaemia and treated with intravenous fluid then later given oral rehydration solution once able to drink.

While children with severe dehydration present with inability to drink, AVPU<A plus sunken eyes and return of skin pinch greater than or equal to two seconds. These children are treated in two steps: Step one involves administration of intravenous fluid followed by administration of oral rehydration solution once a patient is able to drink (which can occur with minutes or hours). Those with some dehydration are able to drink and present with at least two of the following signs: sunken eyes, return of skin pinch 1 - 2 seconds and restlessness/irritability. Lastly, those with diarrhoea without dehydration present with only one sign of some dehydration, and these patients should be treated with oral rehydration solution alone. See (28) for specific details on recommended doses of oral rehydration solution and intravenous fluids. According to the guidelines, all patients with diarrhoea regardless of dehydration classification should be given oral Zinc.

Sample Size

The sample size calculations used the formula cited by Weaver (2017) (191):

$$events = \frac{(z_{\alpha/2} + z_{\beta})^2}{p1 \times p2 \times \log SHR^2}$$

Where: $z_{\alpha/2}\ and\ z_{\beta}$ are standard normal percentiles; p 1 and p 2 are proportions allocated to (treatment) groups.

In order to determine the number of patients that would be discharged by day 14 when treated with Zinc, data were simulated for various values of Sub-Distribution Hazard Ratios (ranging from 1.1 to 1.25) that would need to be observed. From the simulations, it seemed a minimum of 1000 would be

sufficient to detect a SHR of 1.2 in either of the age groups (assuming SHR of 1.2 is clinically plausible).

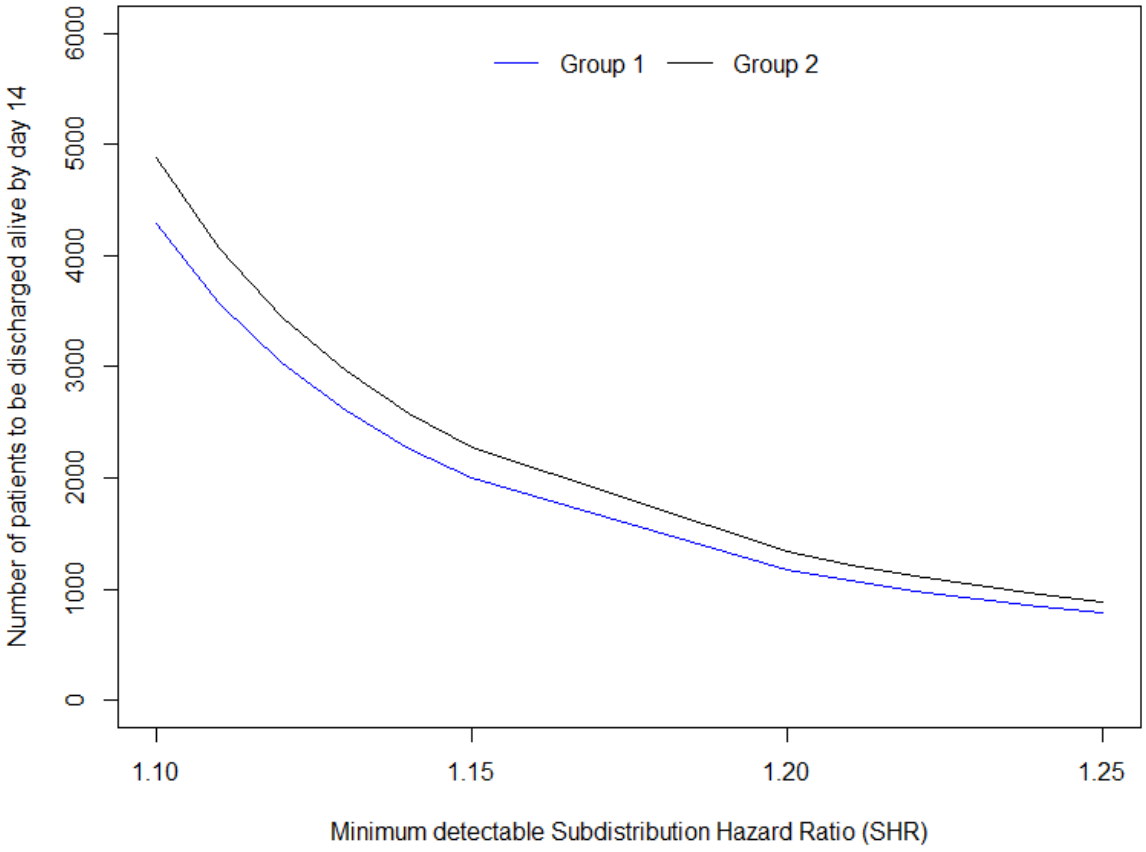


Figure 2: Sample size verification – using significance level of 5% and power of 80%. For group 1 (aged 1 – 5 months): p1 = 72%, p2 = 28% and group 2 (aged 6 – 59 months): p1 = 77%, p2=23%.

Variables

Outcome variables

The analysis used time to getting discharged alive as the primary outcome. However, time to experiencing inpatient mortality was also analysed as part of exploratory analyses.

Independent variables

These variables were grouped into key and auxiliary. Key variables have been outlined in the treatment protocol (28) (also explained in the treatment assignment process) and they should influence diarrhea/dehydration classification and hence treatment. While auxiliary variables may be examined for an explicit influence on treatment assignment, although according to the formal rules (the guidelines) they are not considered reasons to alter treatment assignment. See table 1 for a summary of key and auxiliary variables that were used in the analyses.

Table 1: Summary of key and auxiliary variables

Key Variables	Auxiliary variables
Pulse strength (weak, normal)	Age (months)
AVPU (Alert, Verbal, Pain, Unresponsive)	Gender (male/female)
Extremities warm to (hand, shoulder, elbow)	Weight
Capillary refill (seconds)	Pallor (0, +, +++)
Sunken eyes (Yes, No)	Temperature (Degrees centigrade)
Skin pinch (Seconds)	Fever (present/absent)
Blood transfusion prescribed (Yes, No) – proxy for Hb < 5g/dl	Convulsions (present/absent)
Ability to Drink (Yes, No)	Vomiting (yes/no)
	Length of illness (days)
	Visible severe wasting (yes/no)
	Thrush (present/absent)
	Oedema (none, foot, knee, face)
	ORS prescription
	IV fluid prescription
	Weight age z – score
	Wheeze (present/absent)
	Comorbidities (HIV+, Meningitis, Malaria and or pneumonia)
	Hospital (H1 – H14)

Statistical Analysis

Statistical analysis proceeded in the following four steps:

Step 1 – missing clinical signs data were multiply imputed. Outcome data were included in the imputation models. Then subsets of data by each of the age categories (1 – 5 and 6 – 59 months) obtained for each of the imputed datasets.

Step 2[33] **–** for each imputed dataset (for each of the age categories), patients in the alternative treatment arms were matched to overcome non – random

[33] After multiple imputations, PS will be estimated and matching done on each of the imputed datasets. This method allows to account for uncertainties in PS estimation rather than using averaged PS across the datasets.

treatment allocation. In order to create matched groups, this study compared PS methods: optimal full matching and weighting[34]. Standardised mean differences (and where necessary density plots) were used as diagnostic checks for covariate balance and overlap (107, 108). The method that results in the minimum average absolute standardised mean differences for the majority of the variables was considered appropriate (37).

Step 3: conducting outcome analysis.

For each imputed dataset, analysis estimated time to being discharged. For modelling time to being discharged – a competing risk regression model suggested by Scheike (2011) (145) was used to estimate overall treatment effects. In this case, the focus was on time to being discharged alive (or in stable clinical condition) in which mortality was treated as a competing event. Patients who were discharged against medical advice, referred or absconded from the hospitals were censored as actual length of stay was unknown.

An additive interaction was used in modelling the use of Zinc by nourishment status (143). As such, I derived a variable with four levels representing those who: (i) received Zinc and were well-nourished; (ii) received Zinc and were malnourished; (iii) did not receive Zinc and were well-nourished and; (iv) did not receive Zinc and were malnourished. The reference group used in the analyses, was well-nourished children who received Zinc as they were expected to be discharged sooner than the other three subgroups. To ensure the four subgroups had comparable patient characteristics, I followed the step by step approach

[34] Sub-classification will not be considered in this analyses as it performed poorer than PS weighting and optimal full matching.

suggested by Spreeuwenberg (2010) (144). This involved: (i) Fitting PS models using multinomial logistic regression with a probit link function to obtain four propensity scores per patient; (ii) Examining PS distribution overlap. This was important as a patient in a treatment subgroup should have some probability of being assigned/belonging to other subgroups. Biased estimates may be obtained in case of PS non-overlap; (iii) Examining covariate balance across the four subgroups. Covariables used in creating propensity scores in all the analyses included key signs and symptoms suggested for diagnosing and assessing severity of diarrhoea and dehydration in the Kenyan paediatric guidelines together with variables considered *a priori* to influence the clinical outcomes of interest such as fluid regimen prescribed and comorbidities. These models were fitted on each imputed dataset (adjusting for other variables used in PS models) and results pooled using Rubin rules.

Step 4: sensitivity analysis was conducted to investigate effects of unmeasured confounders and validity of estimates obtained through multiple imputation.

Exploring effects of unmeasured confounders

Sensitivity analysis for unmeasured confounders involved the use of instrumental variables (IV) (192). IV tries to find a natural experiment in the data. This analysis used the timing to admission (weekend/weekday) variable (also discussed in Chapter 4).

Examining validity of multiple imputation

The sensitivity analysis for MI validity was be based on pattern mixture models (193). Pattern mixture models involve formulation of different missing data

patterns. This analysis involved missing data in more than one covariate and therefore patterns were formulated by grouping of patients according to the number of variables for which they have observed data (same as discussed in Chapter 4). After formulation of missing data patterns then imputation followed by outcome analysis were conducted on each using separate models (each had less power but the aim was to check consistency of results). Each pattern estimated treatment effect and these were pooled by weighting the estimates using the proportion of patients in each pattern (193).

Appendix E.2: Additional analyses

The appendix is organised into the following sub – sections:

- Amount of missing data reported and multiple imputation diagnostics
- Examining Constance of effects of covariates in age groups 1 (1 – 5 months) and 2 (6 – 59 months) – time to being discharged alive being the outcome
- Examining distribution of patient characteristics in a four level variable formulated in an additive interaction between Zinc prescription and nourishment status
- Cumulative incidences by nourishment status
- Distribution of variables by levels of weekend/weekday admission variable
- Examining constant effects of covariates in groups 1 (1 – 5 months) and 2 (6 – 59 months) – time to experiencing mortality being the outcome
- Mortality cumulative incidences
- Propensity score weighting by patterns 1 – 3

a) Amount of missing data reported and multiple imputation diagnostics

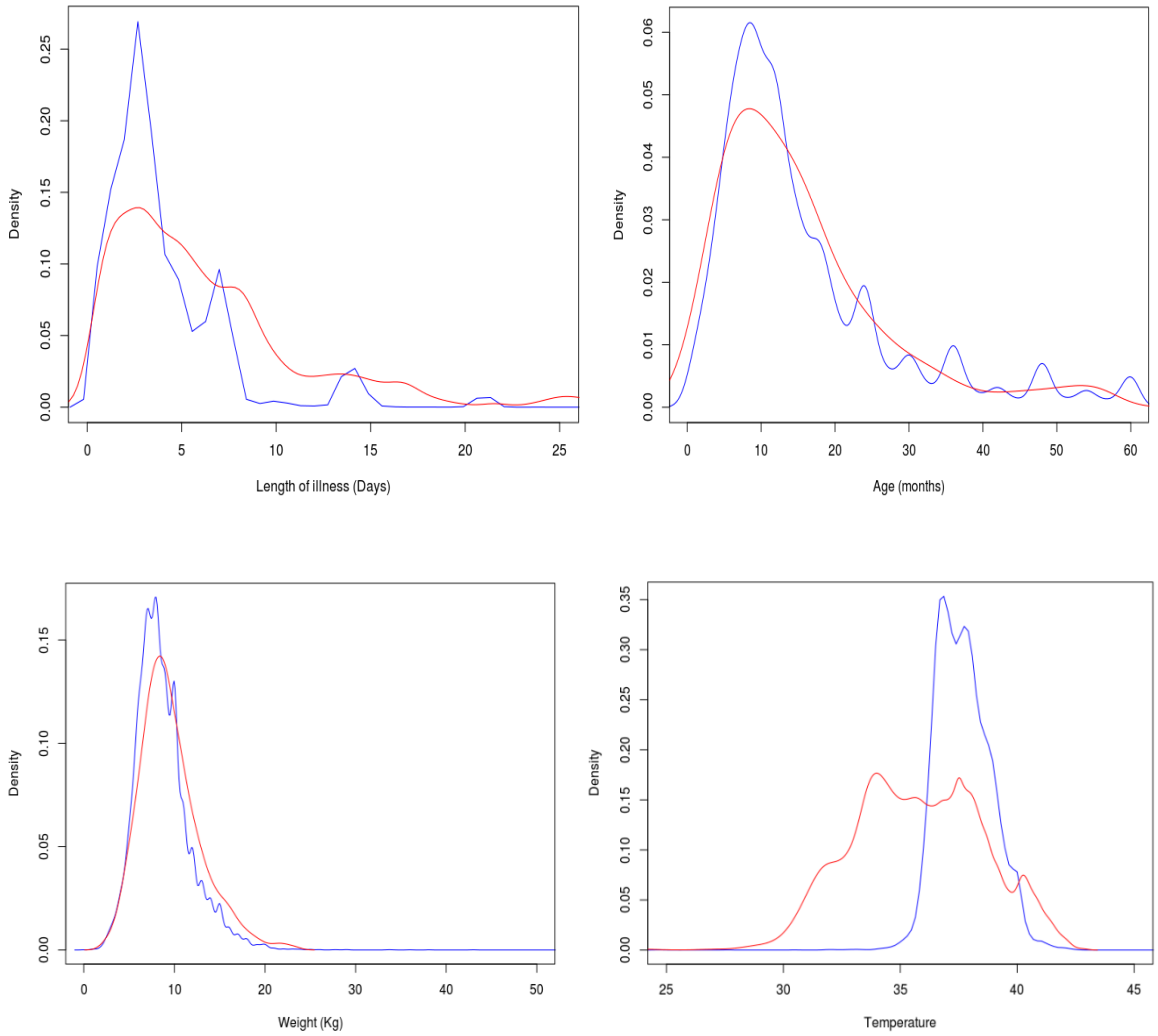
Appendix E.2.1 presents all the variables that were used in the PS models. These data (for those aged < 6 months and those aged>= 6 months together) had missing data of at most 23%.

Appendix E.2.1: Percentage of missing data

Variable	%
Pulse	10.2
AVPU	4.5
Capillary refill	17.0
Sunken eyes	10.7
Skin pinch	8.8
Blood transfusion order	0.3
Ability to drink	8.8
Skin temperature	21.9
Age (months)	0.4
Child sex	0.8
Weight	3.9
Pallor	4.3
Temperature	10.3
Fever	3.1
Convulsions	5.6
Vomiting	2.3
Hospital referral	23.2
Length of illness	2.3
Severe wasting	21.6
Thrush	13.1
Oedema	6.7
Oral fluid	0.0
IV fluid	0.0
Wheeze	4.9
Hospital	0.0
Diarrhoea > 14 days	14.4
HIV	0.0
Pneumonia	0.0
Malaria	0.0
Meningitis	0.0
Outcome (alive/died)	0.9
Length of stay	0.2

After assessing completeness of the variables, 10 datasets were multiply imputed using chained equations. Thereafter, average densities of observed versus imputed values were examined for the following continuous variables: length of illness, age, weight and temperature (**Appendix E.2.2**). The densities for observed

and imputed values were overlapping showing the plausibility of the imputed values and that the imputation models had worked correctly.



Appendix E.2.2: Distribution of observed versus average imputed values (continuous variables)

b) Examining constant effects of covariates in age groups 1 (1 – 5 months) and 2 (6 – 59 months) – time to being discharged alive being the outcome.

Scheike (2011) explained the use of Kolmogorov-Sminorv and Cramer von Mises test statistics to examine if covariates had varying effects across the discharge time points. These tests were conducted on group 1 and 2 PS weighted datasets. Pulse and blood transfusion order were modelled with varying effects across discharge time points in group 1 (1 – 5 months). On the other hand, sunken eyes, blood transfusion order, oral and IV prescription were modelled with varying effects in group 2 (6 – 59 months) (see **Appendix E.2.3**).

Appendix E.2.3: p – value for PS adjusted models

	Group 1 (1 – 59 months)		**Group 2 (6 – 59 months)**	
	Kolmogorov-Smirnov	**Cramer Von Mises**	**Kolmogorov-Smirnov**	**Cramer Von Mises**
Zinc prescription (Yes)	0.14	0.15	0.06	0.08
Pulse (Weak)	**0.01***	**0.00***	0.68	0.86
AVPU (Verbal response)	0.51	0.33	0.91	0.96
Skin temperature (Hand)	0.11	0.17	0.70	0.51
Skin temperature (Shoulder)	0.28	0.31	0.09	0.12
Capillary refill (> 3 sec)	0.71	0.58	0.81	0.90
Capillary refill (Indeterminate)	0.64	0.56	0.07	0.18
Sunken eyes (Yes)	0.33	0.42	**0.00***	**0.00***
Skin pinch (immediate)	0.66	0.83	0.79	0.83
Skin pinch (>=2 sec)	0.15	0.19	0.46	0.46
Blood transfusion order (Yes)	**0.01***	**0.00***	**0.00***	**0.00***
Ability to drink (Yes)	0.48	0.64	0.36	0.16
Oral fluid prescription (Yes)	0.15	0.07	**0.00***	**0.00***
IV fluid prescription (Yes)	0.15	0.20	**0.00**	**0.00***

***** p – value < 5%

c) **Examining distribution of patient characteristics in a four level variable formulated in an additive interaction between Zinc prescription and nourishment status.**

Appendix E.2.4: Patient distributions across the four subgroups and p – values for testing differences before and after PS adjustments (1 – 5 months)

	Zinc-wellnourished	Zinc-malnourished	No Zinc-wellnourished	No Zinc-malnourished	Before multiple PS correction (p - value)	After multiple PS correction (p - value)
Pulse						
Normal	95	91	96	92	0.050	0.998
Weak	5	9	4	8		
AVPU						
Alert	98	96	97	97	0.307	1.000
Verbal response	2	4	3	3		
Capillary refill						
<= 3 Sec	95	94	97	94	0.249	1.000
> 3 Sec	1	1	1	3		
Indeterminate	4	5	2	3		
Sunken eyes						
No	73	63	82	69	0.000	0.989
Yes	27	37	18	31		
Skin pinch						
1 -2 secs	21	26	20	29	0.000	1.000
Immediate	70	58	74	53		
more than or equal to 2secs	9	16	6	18		
Blood transfusion order						
No	98	95	98	96	0.090	0.999
Yes	2	5	2	4		
Ability to drink						
No	15	19	18	27	0.008	1.000
Yes	85	81	82	73		
Skin temperature						
Elbow	3	4	3	5	0.230	1.000
Hand	96	94	94	92		
Shoulder	1	2	3	3		
Child sex						
Female	46	42	49	39	0.142	0.999
Male	54	58	51	61		
Pallor						
mild/moderate	7	15	9	17	0.001	1.000
None	89	82	89	81		

Severe	3	3	2	2		
Fever						
No	19	20	20	27	0.173	1.000
Yes	81	80	80	73		
Convulsions						
No	92	93	89	95	0.053	0.999
Yes	8	7	11	5		
Vomiting						
No	39	42	39	44	0.648	1.000
Yes	61	58	61	56		
Hospital referral						
No	83	81	75	69	0.000	1.000
Yes	17	19	25	31		
Severe wasting						
No	97	83	95	80	0.000	0.909
Yes	3	17	5	20		
Thrush						
No	97	95	96	94	0.275	1.000
Yes	3	5	4	6		
Oedema						
Face	0	0	0	0	0.057	0.116
Foot	1	0	2	0		
Knee	0	0	0	0		
None	99	100	98	100		
Oral fluid prescribed						
No	12	21	20	31	0.000	0.997
Yes	88	79	80	69		
IV fluid prescribed						
No	71	57	68	59	0.000	0.996
Yes	29	43	32	41		
Wheeze						
No	97	95	90	97	0.001	0.999
Yes	3	5	10	3		
History of pesistent diarrhoea						
No	97	96	99	97	0.135	0.999
Yes	3	4	1	3		
HIV						
hiv+	1	3	1	5	0.001	0.995
hiv-	99	97	99	95		
Pneumonia						
pneum+	49	50	57	61	0.010	1.000
pneum-	51	50	43	39		
Malaria						
malaria+	25	16	17	7	0.000	0.998

malaria-	75	84	83	93		
Meningitis						
meningitis+	4	5	9	8	0.007	1.000
meningitis-	96	95	91	92		
Continuous variables: mean (sd)						
Weight	6 (2)	4 (1)	6 (2)	4 (1)	0.000	0.644
Temperature	38 (1)	38 (1)	38 (1)	38 (1)	0.002	0.999
Length of illness	5 (5)	7 (12)	5 (5)	7 (12)	0.000	0.998

Appendix E.2.5: Patient distributions across the four subgroups and p – values for testing differences before and after PS adjustments (6 – 59 months)

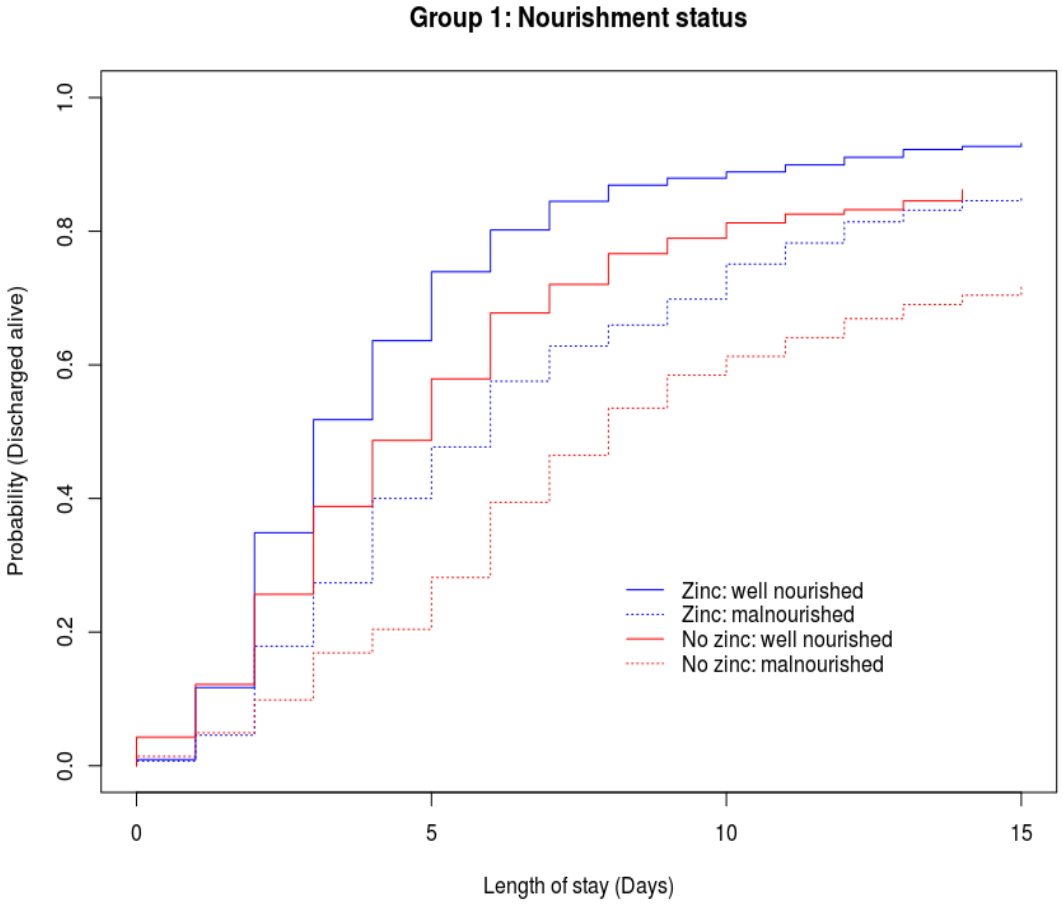
	Zinc-wellnourished	Zinc-malnourished	No Zinc-wellnourished	No Zinc-malnourished	Before multiple PS correction (p - value)	After multiple PS correction (p - value)
Pulse						
Normal	95	93	94	91	0.000	0.962
Weak	5	7	6	9		
AVPU						
Alert	98	97	97	96	0.008	0.999
Verbal response	2	3	3	4		
Capillary refill						
<= 3 Sec	95	95	95	95	0.503	1.000
> 3 Sec	1	1	1	1		
Indeterminate	4	4	4	4		
Sunken eyes						
No	67	61	75	67	0.000	0.985
Yes	33	39	25	33		
Skin pinch						
1 -2 secs	23	23	22	24	0.000	0.982
Immediate	71	64	72	60		
more than or equal to 2secs	7	13	7	16		
Blood transfusion order						
No	98	97	96	95	0.000	0.999
Yes	2	3	4	5		
Ability to drink						
No	14	17	18	18	0.000	0.986
Yes	86	83	82	82		
Skin temperature						
Elbow	3	5	4	6	0.000	1.000
Hand	95	93	94	92		
Shoulder	2	2	2	3		
Child sex						

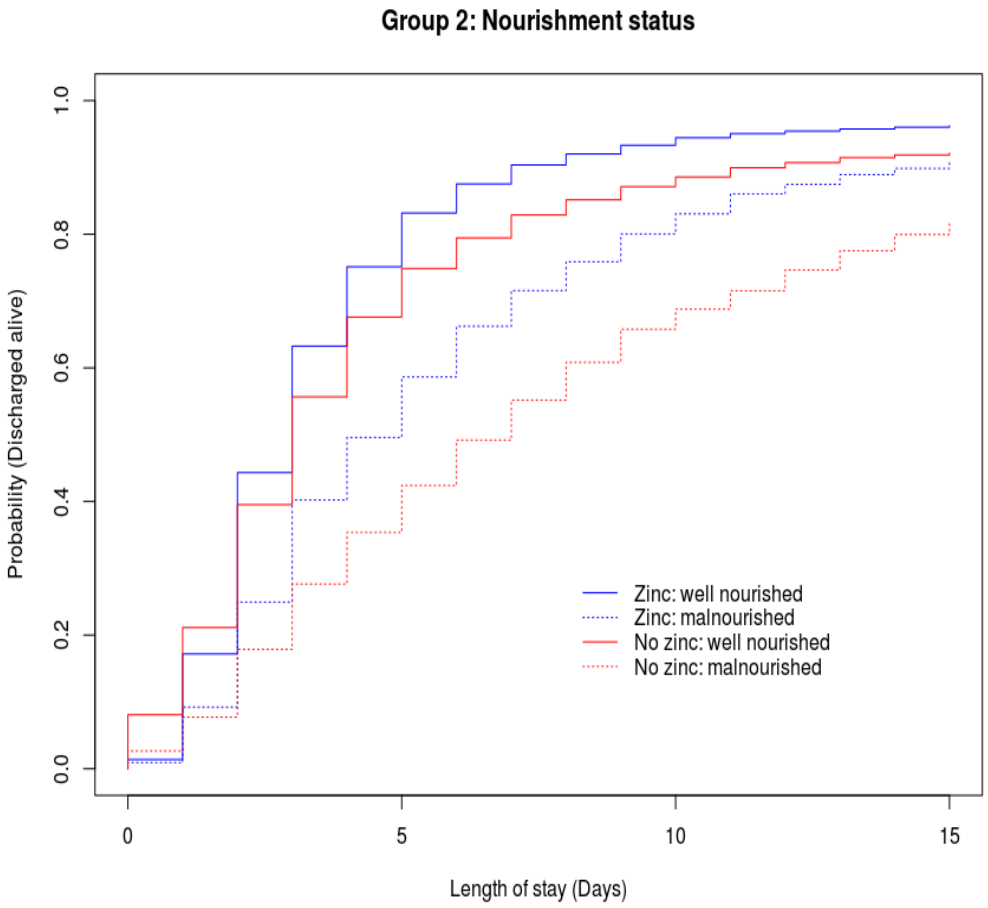
Female	46	40	48	43	0.000	0.987
Male	54	60	52	57		
Pallor						
mild/moderate	8	12	9	16	0.000	0.999
None	90	85	87	80		
Severe	2	3	4	4		
Fever						
No	24	22	20	24	0.007	0.990
Yes	76	78	80	76		
Convulsions						
No	88	91	85	89	0.000	0.999
Yes	12	9	15	11		
Vomiting						
No	19	20	24	29	0.000	1.000
Yes	81	80	76	71		
Hospital referral						
No	84	83	80	75	0.000	0.999
Yes	16	17	20	25		
Severe wasting						
No	98	85	97	77	0.000	0.060
Yes	2	15	3	23		
Thrush						
No	97	95	96	94	0.000	0.981
Yes	3	5	4	6		
Oedema						
Face	0	0	1	0	0.000	1.000
Foot	2	0	3	0		
Knee	0	0	0	0		
None	98	100	96	100		
Oral fluid prescribed						
No	11	15	25	26	0.000	0.877
Yes	89	85	75	74		
IV fluid prescribed						
No	69	62	65	62	0.000	0.939
Yes	31	38	35	38		
Wheeze						
No	98	97	96	96	0.000	0.999
Yes	2	3	4	4		
History of pesistent diarrhoea						
No	97	96	97	96	0.002	0.993
Yes	3	4	3	4		
HIV						
hiv+	1	3	1	3	0.000	0.763
hiv-	99	97	99	97		

Pneumonia						
pneum+	28	36	37	43	0.000	0.909
pneum-	72	64	63	57		
Malaria						
malaria+	31	18	27	17	0.000	0.869
malaria-	69	82	73	83		
Meningitis						
meningitis+	4	5	7	6	0.000	0.992
meningitis-	96	95	93	94		
Continuous variables: mean (sd)						
Weight	10 (3)	7 (2)	10 (3)	7 (2)	0.000	0.000
Temperature	38 (2)	38 (1)	38 (1)	38 (1)	0.000	0.992
Length of illness	4 (7)	6 (11)	5 (11)	7 (13)	0.000	0.943

d) Cumulative incidences by nourishment status

The discharge probabilities by nourishment status for groups 1 (1 – 5 months) and 2 (6 – 59 months) showed that those who were well nourished were likely to be discharged sooner than those who were malnourished. Also those who received Zinc were more likely to be discharged sooner than those who did not receive Zinc (**Appendix E.2.6**).





Appendix E.2.6 (age groups 1 and 2) :Discharge probabilities by nourishment status

e) Distribution of variables by levels of weekend/weekday admission variable

To assess whether weekend/weekday admission variable would form a natural and random experiment, the distributions of covariates were examined across the levels of the instrumental variable (weekend/weekday) in groups 1 and 2. The distribution of each of the patient characteristics between weekend and weekday admission was approximately similar for the variables with majority having absolute standardised mean differences of <=10% (**Appendix E.2.7**). Also likelihood ratio tests showed that weekend/weekday admission was significantly associated with Zinc prescription in both age groups. Further, mediation analysis showed significant association between time to discharge and weekend/weekday

admission through treatment (Zinc/ No Zinc) also in both age groups. These suggested that timing of admission variable satisfied the assumptions as a valid IV.

Appendix E.2.7: Imbalance of covariates between children who were admitted during the weekends and weekdays

Variable	Level	Group 1: 1 – 5 months (n = 1058)			Group 2: 6 – 59 months (n = 8056)		
		Weekend	Weekday	ASMD	Weekend	Weekday	ASMD
Pulse	Normal	85	84	0.03	92	92	0.00
	Weak	15	16		8	8	
AVPU	Alert	97	96	0.05	97	97	0.02
	Verbal response	3	4		3	3	
Capillary refill	<= 3 Sec	94	94	0.00	92	92	0.00
	> 3 Sec	1	1		1	1	
	Indeterminate	5	5		7	7	
Sunken eyes	No	69	70	0.02	60	60	0.00
	Yes	31	30		40	40	
Skin pinch	1 -2 secs	25	24	0.08	25	22	**0.12**
	Immediate	55	57		63	68	
	>= 2 secs	20	19		12	10	
Blood transfer order	No	99	99	0.00	96	97	0.09
	Yes	1	1		4	3	
Ability to drink	No	21	19	0.08	18	18	0.00
	Yes	79	81		82	82	
Skin tempterature	Elbow	4	5	0.03	5	5	0.00
	Hand	95	94		93	93	
	Shoulder	1	1		2	2	
Chid sex	Female	45	46	0.06	42	42	0.00
	Male	55	54		58	58	
Pallor	Mild/moderate	13	13	0.08	10	11	0.03
	None	82	84		90	86	
	Severe	5	3		1	3	
Fever	No	21	22	0.05	26	25	0.02
	Yes	79	78		74	75	
Convulsions	No	94	93	0.01	89	90	0.03
	Yes	6	7		11	10	
Vomiting	No	37	35	**0.13**	17	19	0.06
	Yes	63	65		83	81	
Hospital referral	No	80	74	**0.19**	85	81	0.09
	Yes	20	26		15	19	
Severe wasting	No	92	90	0.05	97	96	0.03

	Yes	8	10		3	4	
Thrush	No	97	96	0.04	98	96	0.04
	Yes	3	4		2	4	
Oedema	Face	0	0	0.01	0	0	0.02
	Foot	1	1		1	1	
	Knee	0	0		0	0	
	None	99	99		99	98	
Oral fluid	No	12	13	0.02	29	26	0.06
	Yes	88	87		71	74	
IV fluid	No	63	62	0.02	65	59	**0.15**
	Yes	37	38		25	41	
Wheeze	No	98	94	0.07	97	97	0.00
	Yes	2	6		3	3	
Diarrhoea>14 days	No	99	98	0.05	98	98	0.00
	Yes	1	2		2	2	
HIV	hiv+	1	3	0.10	1	2	0.04
	hiv-	99	97		99	98	
Pneumonia	pneum+	41	52	**0.23**	32	34	0.08
	pneum-	59	49		68	66	
Malaria	malaria+	15	12	**0.19**	26	25	0.01
	malaria-	85	88		74	75	
Meningitis	meningitis+	8	6	**0.20**	2	4	0.06
	meningitis-	92	94		98	96	
Weight		5.28	5.31	0.07	8.95	8.90	0.03
Temperature		37.31	37.52	0.04	37.51	37.55	0.02
Length of illness		4.9	6.32	**0.25**	4.63	4.78	0.05

Note:

The IV did not balance on the variables with ASMD > 10%. These were however further adjusted for in the Scheike regression model.

f) Examining constant effects of covariates in groups 1 (1 – 5 months) and 2 (6 – 59 months) – time to experiencing mortality being the outcome.

Using both Kolmogorov-Smirnov and Cramer von Mises tests, it was indicated that the effect of Zinc on mortality was not constant across the discharge time points (for both groups). Therefore, Zinc effect estimates with corresponding 95% C.I were obtained for each time point using the Scheike's flexible model. In group 1, Zinc seemed to have no effect on mortality until after a week while Zinc

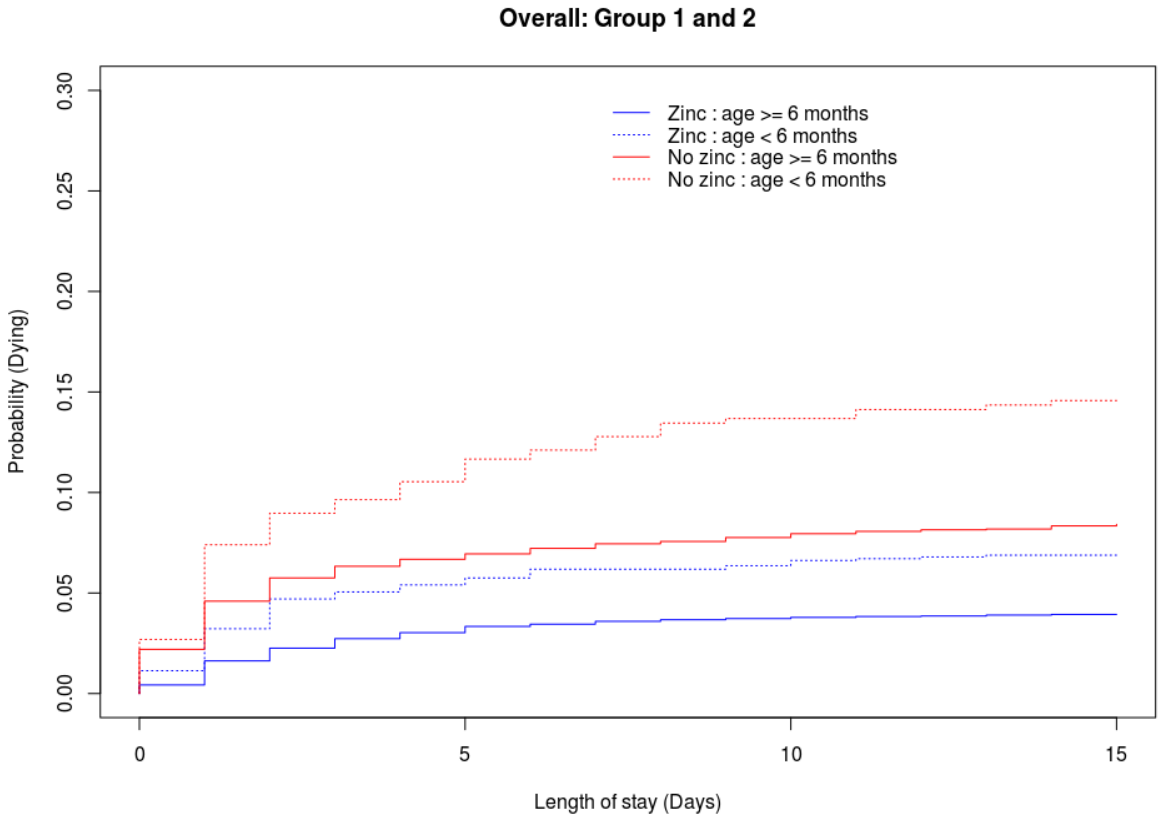
seemed to have a significant effect on mortality in group 2 – with those who received Zinc having reduced risk of dying (results presented in the main text – Figure 5.05).

Appendix E.2.8: p – values for PS adjusted models

	Group 1 (1 – 5 months)		Group 2 (6 – 59 months)	
	Kolmogorov-Smirnov	Cramer Von Mises	Kolmogorov-Smirnov	Cramer Von Mises
Zinc prescription (Yes)	**0.01**	**0.01**	**0.00**	**0.00**
Pulse (Weak)	0.30	0.33	0.49	0.67
AVPU (Verbal response)	**0.00**	**0.00**	0.25	0.04
Skin temperature (Hand)	0.29	0.31	0.24	0.16
Skin temperature (Shoulder)	0.90	0.86	0.15	0.15
Capillary refill (> 3 sec)	0.41	0.62	0.41	0.04
Capillary refill (Indeterminate)	**0.00**	**0.01**	**0.01**	**0.02**
Sunken eyes (Yes)	0.65	0.70	0.32	0.30
Skin pinch (immediate)	0.97	0.93	0.41	0.44
Skin pinch (>=2 sec)	**0.04**	**0.02**	0.45	0.63
Blood transfusion order (Yes)	**0.01**	**0.01**	**0.02**	**0.02**
Ability to drink (Yes)	**0.00**	**0.00**	0.29	0.23
Oral fluid prescription (Yes)	**0.01**	**0.00**	0.50	0.55
IV fluid prescription (Yes)	**0.04**	**0.01**	0.43	0.12

g) Mortality cumulative incidences

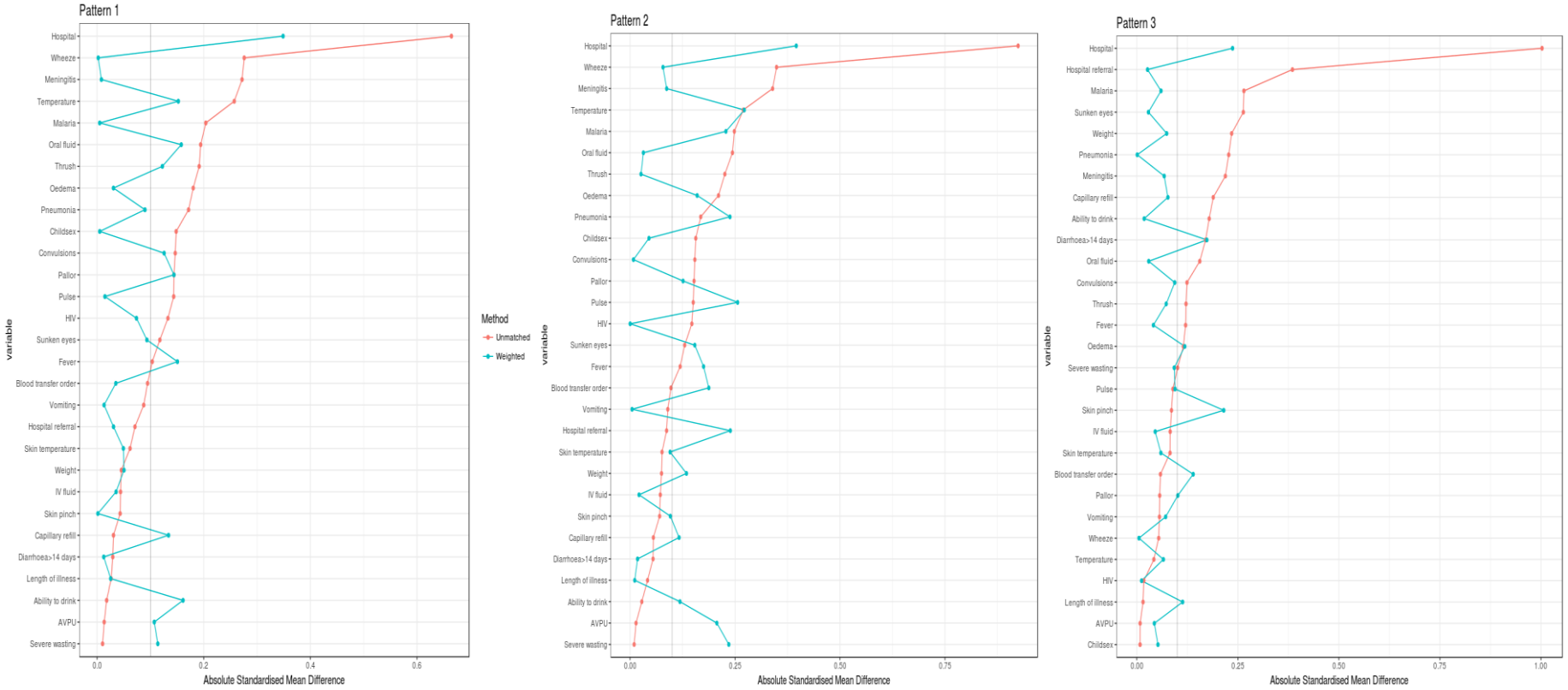
In **Appendix E.2.9**, those who received Zinc in both age groups seemed to have reduced risk of dying compared to those who did not receive Zinc.



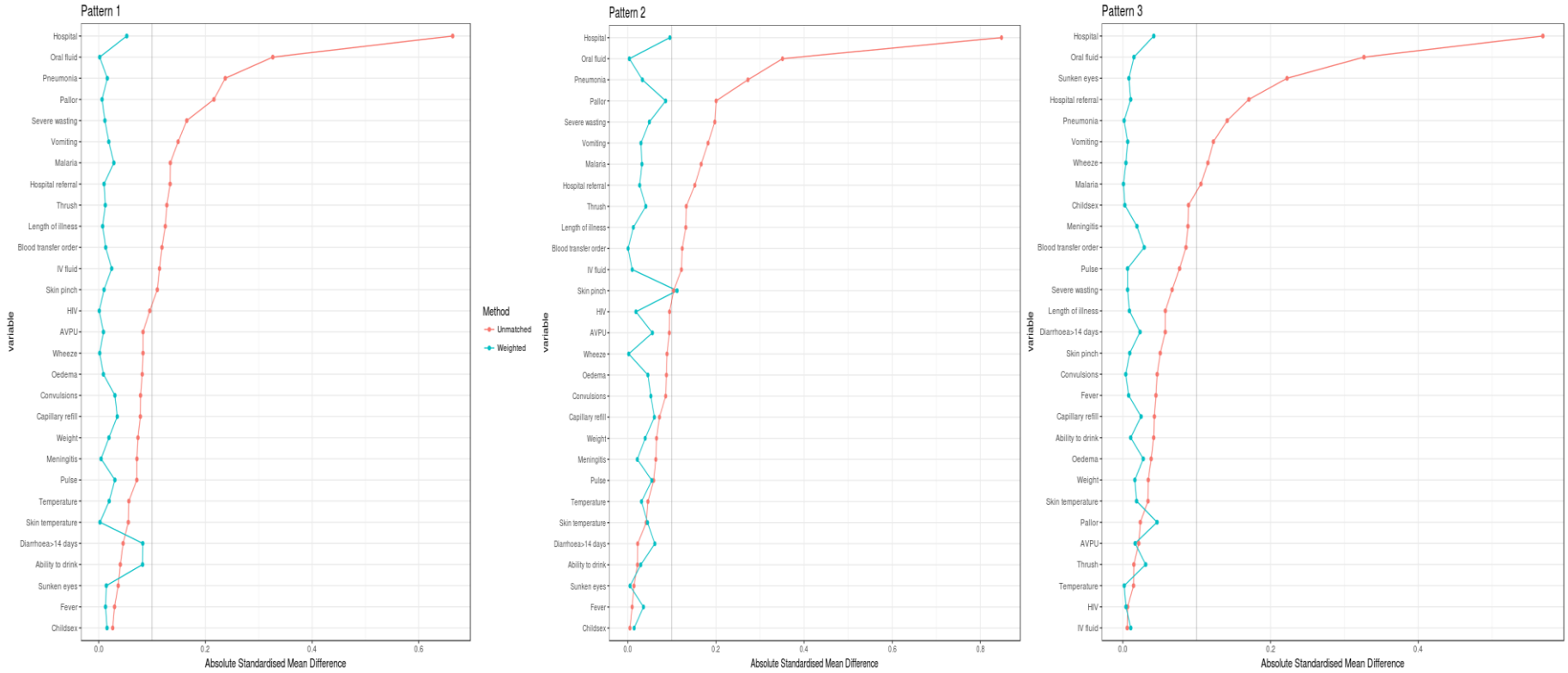
Appendix E.2.9: Probability of dying across discharge time points in groups 1 and 2

h) Propensity score weighting by patterns 1 – 3

Appendix E.2.10: Group 1



Appendix E.2.11: Group 2



Appendix F – Chapter 6

Appendix F.1

Box 6.4.1: Proof by Leyrat (2017) of inconsistence of the across method

Suppose T represents a treatment variable, Y an outcome, X a confounder (having values missing at random), and R a missing data indicator where:

$$X \sim Bernouilli\ (0.5)$$

$$T \sim \begin{cases} Bernoulli\ (0.1)\ if\ X = 0 \\ Bernoulli\ (0.9)\ if\ X = 1 \end{cases}$$

$$R \sim \begin{cases} 1 & if\ T = 0 \\ Bernoulli\ (0.1)\ if\ T = 1 \end{cases}$$

$$Y \sim \begin{cases} Bernoulli\ (0.1)\ if\ X = 0\ and\ or\ T = 0 \\ Bernoulli\ (0.9) \quad if\ X = T = 1 \end{cases}$$

When X is observed (X_{obs}) then R = 1, and when X is missing (X_{miss}) then R = 0. Here the model linking treatment (T) to the confounder X is $\log\left(\frac{\Pr(T=1)}{(1-\Pr(T=1))}\right) = -2.19 + 4.39\ X$. As the distribution of the outcome is the same as the treatment, then the expected outcome (E [$Y^{T=1}$]) is 0.5. When both T and Y are equal to 1 and $e(.)$ a function of the propensity score, then the expected PS (based on the treatment model above) is E [e (X_{obs}, X_{miss})] = 0.89. When T and Y are equal to 1 then the confounder X can take the following three values: missing, 0 and 1 resulting in expected PS of 0.89, 0.1 and 0.9. Therefore:

$$E\left[\frac{TY}{E[e(X_{obs}, X_{miss})]}\right] = \frac{\Pr(T=1, X=0, R=1, Y=1)}{0.1} + \frac{\Pr(T=1, X=1, R=1, Y=1)}{0.9} + \frac{\Pr(T=1, X=0, R=0, Y=1)}{0.89} = 0.47$$

Clearly $E\left[\frac{TY}{E[e(X_{obs}, X_{miss})]}\right]$ $is\ not\ equivalent\ to\ E\ [Y^{z=1}]$. This demonstrates that the across method may produce an inconsistent estimator.

Appendix F.2: Distribution of patients by clinical characteristics between gentamicin plus penicillin (P + G) and Penicillin (P) per experiment (Coloured cells show cases where G+P have more positives on the characteristics).

	Experiment 1		Experiment 2		Experiment 3	
	G + P	P	G + P	P	G + P	P
Central cyanosis (present)	-	-	1%	0%	2%	0%
Indrawing (present)	100%	100%	78%	76%	72%	63%
Grunting (present)	-	-	31%	13%	40%	12%
Pallor						
mild/moderate	5%	4%	6%	4%	7%	6%
None	93%	94%	92%	94%	91%	92%
Severe	2%	1%	2%	1%	2%	2%
Able to drink	-	-	85%	87%	82%	87%
APU						
Alert	-	-	96%	97%	94%	97%
Pain response	-	-	1%	1%	3%	1%
Unresponsive	-	-	1%	0%	1%	0%
Verbal response	-	-	1%	2%	2%	2%
Difficulty in breathing (present)	72%	66%	72%	66%	72%	60%
Capillary refill						
1 Sec	60%	72%	60%	70%	60%	68%
2 Sec	36%	26%	36%	27%	35%	29%
>2 Sec	3%	2%	4%	2%	4%	3%
Fever (present)	80%	80%	82%	81%	81%	82%
Comorbidities						
None	82%	85%	79%	83%	78%	78%
Malaria	11%	9%	12%	10%	13%	13%
Diarrhoea/Dehydration	2%	3%	2%	3%	3%	3%
Both	5%	4%	6%	4%	6%	6%
Convulsions (present)	4%	5%	6%	6%	7%	8%
Vomitting	37%	37%	36%	37%	38%	41%
Cough (present)	90%	90%	89%	90%	87%	88%
Difficulty in feeding	32%	27%	36%	31%	37%	31%
Crackles (present)	57%	52%	57%	50%	54%	45%
Hospital referral	18%	17%	20%	17%	22%	17%
Thrush (present)	2%	2%	2%	2%	2%	2%
Wheeze (present)	16%	15%	16%	15%	16%	14%
Oxygen ordered	23%	12%	21%	13%	28%	11%
IV fluid prescribed	5%	3%	6%	4%	7%	5%
Quinine prescribed	3%	4%	5%	5%	5%	8%
Artesunate prescribed	10%	6%	13%	6%	13%	10%