The development of process measures for the quality of inpatient care given to children hospitalised with common illnesses in Kenya

CHARLES LWANGA OMONDI OPONDO

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

of the

University of London

2014

Department of Medical Statistics

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by KEMRI–Wellcome Trust Research Programme, Nairobi, Kenya

Research group affiliation(s): Health Services Unit, Public Health Research Department

Declaration

I, Charles Opondo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: [redacted] **Date:** Wednesday 17th December 2014

Acknowledgements

I would like to express my deep gratitude to colleagues, friends and family who have supported me in many ways over the course of this work.

I acknowledge the valuable contribution of Elizabeth Allen and Tansy Edwards whose initial work on measurement of quality of care inspired me to undertake these studies.

I am extremely grateful to Elizabeth Allen, Mike English and Jim Todd for their constant advice and valuable comments that helped conceive and refine the concepts at the core of this thesis.

I am grateful to Andrew Hutchings for his guidance during the literature review.

I thank Bianca DeStavola for providing access to much-needed software and manuals, and for assistance in conducting sections of my analysis.

I acknowledge Ambrose Agweyu and Graham Wheeler for sharing their thoughts which helped improve some of the methodological approaches used to address key issues explored in this work.

I appreciate the funding provided by the KEMRI-Wellcome Trust Research Programme in Nairobi, Kenya, which supported my work. I am also grateful to the Wellcome Trust UK for a fellowship award (#089349) which supported the first six months of my studies, and an award (#076827) to Mike English which supported the study from which most of the data analysed in this work arose.

Lastly, my family have provided lots of encouragement over the years, and I am eternally grateful to them. I am particularly thankful to my partner Nelly for her constant moral and material support.

Even with all this support my thesis will inevitably contain some errors and shortcomings; for these I am fully responsible.

Abstract

Assessing the quality of health services remains a major challenge for the less industrialized nations of the world. Hospitals consume high proportions of national health budgets but approaches to examining their capacities and performance are still rudimentary. Better assessment strategies are essential for understanding what investments in hospital care and health systems are achieving and whether resources are being equitably distributed. Assessing hospital care for children is a particular challenge given the multiplicity of steps to be correctly undertaken in examining patients, diagnosing disease and giving appropriate treatment.

This thesis describes the development of a measure of quality of care for children admitted to hospital, which can be routinely deployed to form part of future nation-wide health system assessments. The measure is based on standards of paediatric care described by national clinical practice guidelines for Kenya adapted from WHO guidelines. It is constructed using data from a trial of a new intervention to improve quality of care in first-level referral hospitals in Kenya. The proposed measure is subsequently validated using data from observations of routine hospital care for children admitted to a different set of hospitals over a period not covered by the primary data.

A number of statistical techniques are used in this regard: these include factor analysis to explore the dimensions of process of care, logistic regression to study the association between the new measure and mortality, and multilevel modelling to explore the amount of variability in the data lost through modification of the structure of the measure. These analyses show that the items making up the measure are consistent with three conceptual domains of clinical process described in guidelines. They also provide evidence that the measure is associated with other exposures and outcomes in ways that strengthen its validity and suitability of purpose. Specifically, there is strong evidence that adherence to each of six generic recommendations of how care should be provided is associated with a reduction in odds of death by a fifth. This thesis thus demonstrates the usefulness of a generic approach to measuring quality of care, and highlights key issues to be addressed when extending this approach to other settings.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction and a Review of Literature

1.1. The concept of quality in health care

Measuring quality of health care is an important aspect of any health system since it provides the information necessary to monitor and improve service delivery. However quality measurement is not straight-forward because healthcare is complex and quality of care is a difficult-to-define concept [Marcinowicz *et al.* 2009, Chanthong *et al.* 2009, De Maeseneer & De Sutter 2004]. It is known that different aspects of care contribute to its overall quality, and for this reason it is important to deconstruct 'quality' to allow for a clearer understanding. The most commonly discussed framework for measuring quality was proposed by Avedis Donabedian when he described three attributes of quality of care namely structures, processes and outcomes [Donabedian 1988].

Structures are the resources that support the delivery of health services, including health workers, medical devices and equipment, infrastructure and drugs. Processes refer to what is actually done by health workers in providing care, such as taking clinical history, performing physical examination, making a diagnosis supported by clinical evidence and history, and charting an appropriate course of treatment to restore health. Outcomes are the consequences of care, such as death, recovery, satisfaction with care received and duration of hospital stay. Good outcomes are ultimately what is always sought in healthcare so good quality of care can be conceptualised as the route to achieving them. The relevance of outcomes to overall quality of care is generally clear and acceptable to providers and users of health care services and is thus rarely questioned [Donabedian 2005].

Assessments of structure make for sensible measures of quality because it is assumed that without the appropriate necessary resources favourable outcomes cannot be achieved. This is highlighted in low-income countries where insufficiently funded and overstrained health systems often suffer serious resource constraints which tend to adversely affect the provision of good care [Linden *et al.* 2012, Leatherman *et al.* 2010]. Not only does poor structure limit the ability and desire of health workers to provide care, but it may also reduce demand for services [Collier *et al.* 2003]. While a link between structure and outcomes may appear obvious there is still little clear evidence of a cause-effect relationship between them, other than ecological-level associations between resource inputs and outcomes. It can therefore be argued that given a basic set

of resources what is actually done by health workers in providing care – the process of care – is possibly what best captures quality [Rutten *et al.* 2010, Williams *et. al*. 2006]. When process measures are informed by well-established evidence of a link to good outcomes, expert opinion or guidelines, they become relevant to what quality is perceived to be [Wobrock *et al.* 2009]. For example, in the management of myocardial infarction, timely thrombolysis has been linked to lower morbidity and mortality in RCTs, hence the use of 'door-to-needle' times as a process-of-care indicator of quality [Corfield *et al.* 2004, FTT Collaborative Group 1994].

1.1.1. Current strategies for measuring quality of care in hospitals in developed countries

Quality of care measurement strategies are highly varied because of the varied needs of different health systems [Peabody *et al.* 2006]. In developed countries availability of resources is less of a limiting factor to the provision of good hospital care, so quality assessments tend to focus on outcome and process measures. For example a multi-country evaluation of disparities in quality of care for different socioeconomic classes in Canada, England, New Zealand and the United States focused on measures of cancer survival and screening rates, asthma mortality rates, suicide rates, smoking rates and acute myocardial infarction 30-day case fatality rates as indicators of quality of care [Hussey *et al.* 2007]. Similarly the declining trend in childhood cancer mortality in economically developed countries as has been considered an indicator of improved quality of medical care [La Vecchia *et al.* 1998].

Mortality is generally a rare event. Differences in death rates attributable to quality of care received are therefore harder to detect, often requiring large datasets collected over long periods of time to observe sufficient events to detect these differences. Furthermore outcomes tend to be affected by factors such as socio-economic disparities and case mix which are unrelated to care given. For example the decline in childhood mortality from infectious causes observed in the early 1900's (even before mainstream use of antibiotics) was attributable to factors unrelated to quality of healthcare, such as improved nutrition [Cutler & Meara 2001]. These factors fall beyond the realm of hospital care and create 'risk-sets' of patients which are more powerful pre-determinants of outcomes than the quality of care given or received [Lilford & Pronovost 2010]. If information on these factors is available their effect can be estimated or adjusted for [Zaslavsky 2001]. Unfortunately in low-income settings the

detailed data required to detect and/or explain trends in outcomes is often unavailable, incomplete or too poorly coded to be useful [Kihuba *et al.* 2014, Williams & Boren 2008, WHO 2006].

Patient satisfaction measures have been developed as an alternative outcome specifically to identify, from service consumers' perspective, shortfalls in service delivery for input into quality improvement drives [Doyle *et al.* 2010]. In developed countries this undertaking is aided by the availability of very large databases of data routinely collected in inpatient surveys which are virtually non-existent in low income countries. But as some studies have reported, patients may not focus on the technical aspects of care – whose suitability they often lack the competence to judge [Donabedian 2005, Harutyunyan *et al.* 2010] – which directly contribute to restoration of health, but rather on social and interpersonal aspects such as perception of kindness, communication and respect which are very subjective by nature [Schoenfelder *et al.* 2011]. Thus the use of outcome measures especially in the absence of sufficiently detailed data may fail to correctly identify 'problem areas' in the continuum of care.

Process-of-care measures make up the core of the clinical audit culture in health systems of the developed world where legal and professional pressures and availability of resources to support audits has led to their integration into routine clinical practice [Maher 1996]. Organizations such as the National Institute of Clinical Excellence (NICE) in the UK and the Agency for Healthcare Research and Quality (AHRQ) in the United States define national standards of care, promote the use of process-of-care data to measure performance against these standards, and direct implementation of changes necessary to improve care [NICE 2002]. Additionally process measures have been used in evaluating effectiveness of interventions such as the HAPPY AUDIT which aimed to reduce inappropriate antibiotic prescriptions for respiratory tract infections in Europe and South America [Bjerrum *et al.* 2010], and in pay-for-performance initiatives to identify levels of compliance with process targets to be rewarded [Petersen *et al.* 2006, Rosenthal *et al.* 2005]. There is little use of structure measures of quality of care possibly because of little variation in core staff and equipment across places. However there is widespread use of structure to measure equity in healthcare service provision [Macinko & Starfield 2002] (for example, in terms of the comparative number of specialist clinicians [Cooper *et al.* 2002]) and its association with outcomes at a population level [Vogel & Ackerman 1998, Shi 1994, Shi 1992].

1.1.2. Quality of care measurements in a low income country

A dearth of resources to support care and translation of existing evidence into practice in low income countries has steered quality measurement towards structure and process measures because substantial shortfalls in these areas may prevent any meaningful progress in improving outcomes [Peabody *et al.* 2006]. In Kenya assessments of structure have taken the form of Service Provision Assessments (SPA) which are rapid cross-sectional assessments of resource availability in health facilities using detailed check-lists [NCAPD *et al.* 2011]. SPAs focus on child health, maternity and newborn care, family planning, sexually transmitted and other infectious diseases, and HIV/AIDS, reporting on hospitals' infrastructure, resources, systems, drug and vaccines supply and availability and counselling services. This wide scope gives a very comprehensive picture of the 'supply side' of healthcare but yields voluminous reports which often do not easily identify where deficiencies are localised within institutions.

Process-of-care based quality assessments have also been undertaken in Kenya albeit in non-routine settings. A study investigating the effect of training of health workers on the use of guidelines for case management of severely ill children, supervision and feedback on quality of care for these children used 14 indicators of correct care for key illnesses as outcome measures to compare the intervention and control hospitals [Ayieko *et al.* 2011]. The multi-country evaluation (MCE) also measured correctness of case management practices for different illnesses. In one approach to measurement this MCE, instead of reporting performance on each of the several binary indicators separately, summarised them into a single mean index of integrated child assessment which was then used to compare quality of care provided by health workers with different durations of training [Huicho *et al.* 2008].

Even when process of care has been measured for the same illness definitions of correct care have varied widely. A review of malaria care quality found that while some studies defined correct treatment as use of any anti-malarial drug (sometimes ignoring the correctness of dosage), others were stricter, accepting the use of guideline-recommended drugs only [Zurovac & Rowe 2006]. This lack of uniformity of indicators across studies could be attributed to the differences in guideline recommendations from illness to illness, coupled with non-systematic selection of process of care indicators. Another problem is that indicators tend to be task-specific and different tasks are not as easy as each other – checking for fever, for instance, is a

relatively straight forward task compared to performing an auscultation for abnormal breathing sounds. An intervention effect based on the simpler task might appear larger than that measured using the more complex one if task complexity is not taken into account. It therefore becomes very difficult to compare effects of different interventions whose endpoints are not measured on the same set of indicators [Rowe 2013]. A third problem is that multiple indicators may be used to measure quality of care for the same illness – a Delphi study to rate the acceptability of quality indicators found that experts recommended the use of more than 10 indicators each for 4 different illnesses [Ntoburi *et al.* 2010]. As a result there is a lot of heterogeneity in process-of-care measures across studies, a fact which greatly limits their comparability.

1.1.3. Benchmarking quality of care on established standards and guidelines

Infectious diseases are responsible for more than 65% of deaths in children aged 1–59 months globally, majority of these being in low income countries [Liu *et al.* 2012, Black *et al.* 2010]. Together, pneumonia, malaria and diarrhoea contribute to over 60% of these deaths. Effective treatments for these illnesses are available, and strategies for delivering them to children who need them have been developed. A notable and widely promoted strategy is the Integrated Management of Childhood Illnesses (IMCI) strategy. This was published by the WHO as a draft algorithm to guide health workers in the management of a range of childhood illnesses responsible for majority of morbidity and mortality [WHO 1997]. IMCI algorithms use syndromes – collections of signs and symptoms of illness and laboratory test results – to identify and classify childhood illnesses by severity and select the most appropriate treatment. Several countries have adapted these generic guidelines to fit their own local disease patterns. They have also designed and implemented training programmes for health workers on the use of guidelines [Lambrechts *et al.* 1999].

Kenya adopted the IMCI approach in 2000 and has had traditional IMCI training for over a decade (even though it fell out of favour within the decade due to its high cost). This syndromic approach has also been applied to hospital level care [Berkley *et al.* 2005, English *et al.* 2003]. To help implement the hospital care component of IMCI new guidelines for care of children hospitalised with acute illnesses – the Basic Paediatric Protocols [MoH 2006] – were designed. To promote their uptake the Emergency Triage Assessment and Treatment Plus Life Support (ETAT+) course for health workers who provide initial care for hospitalised children [Irimu *et al.* 2008] was

designed and a cluster randomised trial conducted to evaluate whether the intervention was effective. A summary of the trial is presented in section 1.4.

There is some evidence that health workers do not always provide optimal care [Reyburn *et al.* 2008, English *et al.* 2004, Duke & Tamburlini 2003, Nolan *et al.* 2001]. The resulting poor care has been identified as a key limiting factor in the quest for better health outcomes [Peabody *et al.* 2006]. Reasons for poor care may include constraints in the resources necessary to support delivery of services [Opondo *et al.* 2009], negative pre-conceptions about some recommended treatments, health workers' (over)confidence in their own capacity to treat severe illness without referral [Walter *et al.* 2009] and medico-legal constraints limiting the range of services that available staff can provide [Simoes *et al.* 2003], among others. This know-do gap results in a degradation of the quality of care provided to children admitted to hospital with life-threatening illnesses, and although it is widely assumed that poor health in low income countries is mostly brought about by limited access to health services (hence the widespread use of interventions that focus on providing additional resources for health), inadequate or incorrect processes of care may in fact be the main limiting factor to the attainment of good health [Das 2011]. Despite this, policy makers in many low-income health systems have not yet developed nor implemented quality measurement as part of routine quality assurance for hospital care [Ntoburi *et al.* 2008].

Progress is necessary on all fronts: ability to measure structure for the assessment of coverage and equity, ability to measure process to identify whether care provided complies with the best available evidence, and ability to measure and understand outcomes. This thesis is about process, specifically describing the design of a measure capable of effectively identifying and reporting problem areas in the process of care. The use of a score to measure how closely care provided corresponds to established standards of practice is an attractive way to simplify, summarise and report quality and compare it between individual patients, clinicians, diseases and hospitals, and to identify areas of poor care for improvement. If health systems are to meet health and equity goals then policy-makers, planners and researchers seeking to understand health systems need quality measures presented in a form that can answer questions such as, 'Are all hospitals meeting a minimum standard?' and 'Which aspects of hospital care are poorest?'.

1.1.4. Summary

Whereas measuring structure is a relatively simple undertaking, structure measurements fall short in assessing overall quality of care. Outcomes, while arguably the most important, are also the most demanding measures in terms of quantity and quality of data required and are prone to interference from factors unrelated to care received. Existing process measures are made up of a potentially very large set of disease-specific indicators which are hard to use as general measures of quality. Nevertheless process measures remain attractive for monitoring the performance of systems of care in a low-income setting because process of care is frequently the target of improvement efforts in the 'service delivery arena', an important health systems building block. Thus the interest of this work is to develop a measure of quality with a focus on process of care for children admitted to hospital, which spans common illnesses that are targets of system or service improvement interventions in many low-income countries and which can be measured at the individual patient level. The proposed measure is based on current clinical management guidelines and would need to be updated alongside future updates to these guidelines.

1.2. Literature review

Before embarking on the development of a new measure, a review of the literature was undertaken to first determine if previous efforts had been made in quality of care measurement in low income settings, and to identify the approaches taken by other studies that attempted to summarise or report aggregate measures at the patient level.

1.2.1. Specific objectives of the review

i. To identify literature describing the development of summary measures of quality of care for children in low income settings.

ii. To identify literature providing descriptions of methodological approaches to developing summary scores for measuring quality of care.

1.2.2. Methods

The literature search was done on MEDLINE/PubMed and Embase. Details of the search terms are presented in Appendix A.1. A broad literature search of articles published in English was undertaken and 2 different criteria applied to the results to meet objectives (i) and (ii). The reason for this approach to the search was that any

literature potentially addressing the first objective was likely to be a subset of a larger body of literature that describes the development of measures of quality of care, and these could be sought using slightly less restrictive search terms and inclusion/exclusion criteria. For the first objective the review was limited to studies measuring quality of care for children aged 0-59 months, admitted to hospital with illnesses requiring non-surgical treatment (specifically acute illnesses included in the inpatient IMCI guidelines), in a low-income setting, using process of care as a measure of quality. Studies in an older population, developed country setting, or focusing on other attribute of quality of care were excluded. For the second objective any study reporting the development of a summary measure, score or composite indicator for measuring quality of process or outcomes, and those from non-low income and non-paediatric care settings were included. Data extracted from the studies include study objectives, setting, population, and a summary of statistical methods and findings. Key references were flagged and reviewed for more information pertaining to methods used. Quality assessment tools for methodological studies are still poorly developed and during this review two reports were found which broadly discussed methods of constructing measures of complex constructs [Harris *et al.* 2009, OECD 2008]. A list of suggested steps for constructing composite indicators, shown in Appendix A.2, was adapted from these reports and used to characterise the studies.

1.2.3. Results

The searches yielded 277 articles of which 211 were discarded after initial screening by title and a further 41 after screening by abstract. Of the remaining 25 articles, the full text of 4 pre-1995 reports could not be located. 21 articles were retrieved and screened by content and 6 of them discarded for irrelevance. The remaining 15 articles were found to report on the development of composite quality measures – objective (ii) – but none met the criteria set out in objective (i) of this review. Figure 1.2-1 illustrates results of each step of the search. The 15 articles found to address objective (ii) are summarised in Table 1.2-1, with a more comprehensive description in Appendix A.3.

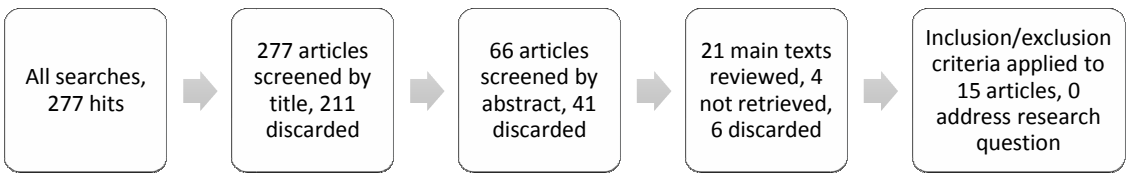


Figure 1.2-1: Summary of the literature search

Table 1.2-1: Summary of the reviewed studies

Check-list adapted from the OECD handbook for constructing composite indicators and Harris et al. 2009. See Appendix A.2 for explanations of suggested steps

Suggested steps	Chen 2011	Wierenga 2011	Bamm 2010	Mael 2010	Suhonen 2010	Chevat 2009	Klassen 2009	Saloojee 2009	Siebes 2008	Najjar-Pellet 2008	Llewellyn 2007	Siebes 2006	Sixma 2000	Ashton 1999	Symmons 1995
Perspective of quality measurement	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Data selection, sample size	+	-	-	-	-	-	-	-	-	+	+	+	+	+	0
Handling of missing data*	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0
Multivariate analysis	+	0	+	0	+	+	+	+	+	0	0	0	+	+	0
Weighting, aggregation, generation of a summary measure	0	0	0	0	+	+	-	0	-	+	+	0	0	+	+
Assessment of robustness and sensitivity analysis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reliability and validity	+	+	+	+	+	+	+	+	+	0	+	+	+	+	+
Links to other indicators and measures	0	+	+	0	0	+	+	+	+	0	0	+	0	0	+

Key: (+) reported and appropriate; (-) reported but not appropriate; (0) not reported; *some multivariate methods adjust for missingness-at-random but do not justify this assumption

1.2.4. Overview of the studies reviewed to explore methodologies

A majority of the studies were conducted in developed countries, mostly North America and Europe, except two: one in South Africa [Saloojee *et al.* 2009] and another in China [Chen *et al.* 2011]. Most study populations were patients receiving care, their caregivers or health workers providing care and 3 studies reported work in a paediatric care setting.

All studies provided good descriptions of the clinical area in which the measure was to be applied and perspectives from which quality was considered. For example 5 studies described score development using the Measure of Process of Care (MPOC) as a starting point [Bamm *et al.* 2010, Saloojee *et al.* 2009, Klassen *et al.* 2009, Siebes *et al.* 2008, Siebes *et al.* 2006]. Developed by a multidisciplinary team of clinical neuro-development and rehabilitation researchers in Canada the MPOC was designed to measure the process of family-focused intervention services for children's rehabilitation services by using the behaviour and interaction of providers and caretakers. This focus was taken based on the assumption that parents of children receiving care would appreciate the identification and modification of processes they perceived as poor. A similar study measuring quality of pharmaceutical care for Dutch elderly patients discussed the links between prevention, diagnosis, treatment and follow-up of diseases in this group and the risks of adverse drug events, poly-pharmacy, under-treatment, functional decline and other important outcomes [Wierenga *et al.* 2011]. To generate items and domains the MPOC studies focused on care-giving as a construct to generate a pool of potential representative items while Wierenga *et al.* used an established quality indicator set from the Assessing Care of Vulnerable Elders (ACOVE) study [Wenger *et al.* 2001]. The items had been designed to capture quantifiable perceptual elements of the construct of interest using semi-quantitative (Likert-type) and binary responses.

All 15 studies used either a multivariate analysis or expert opinion or a mix of both to refine, clarify and group the list of selected items into 'scales' or 'domains' which represent a specific dimension of quality. Multivariate methods used were based on factor analysis (FA) which assumes that there exist latent variables which are not directly observed but are measured through other observed *variables* (alternatively referred to here as 'items'). FA aims to measure these latent variables or factors and estimate their effects on some outcome. This technique may also be used for data reduction by decomposing the set of items measuring an underlying construct into a

series of uncorrelated linear combinations or groupings of related items. If there is no *a priori* belief on the nature of groupings of items, then this multivariate method is useful for identifying these groupings. In situations where there is reasonable knowledge of the phenomenon or construct that the set of observed variables represent, a confirmatory factor analysis (CFA) can be set up to refine item groupings. However the results of FA tend to be inconsistent, yielding different item groupings whenever small changes occur in the underlying datasets and this could explain why validation studies replicating quality measures in new settings sometimes fail to confirm the usefulness of measures [Saloogee *et al.* 2009]. Moreover groupings solely determined by statistical processes may not be relevant to real-life settings. For these reasons other studies favoured expert opinion for item grouping refined through a consensus approach such as the Delphi method [Wierenga *et al.* 2011, Najjar-Pellet *et al.* 2008], or a combination of statistical methods reinforced by expert opinion or consensus [Chen *et al.* 2011, Suhonen *et al.* 2010]. The Delphi technique refines expert opinion in several rounds during which information is collected from panelists, collated and presented back to them to form the basis of the next round, allowing them to revise their opinion based on this feedback until consensus is reached. Through repeated feedback cycles this technique helps refine the selection and aggregation of items to be measured.

The review highlighted some inadequacies in statistical approaches. For example considerations for sample size estimation and handling of missing data were poorly reported in all studies. Less than half of the reviewed studies even mentioned their sample sizes, and those which did used Nunally's suggested criterion for scale validation of 5-10 observations per item [Streiner & Norman 1989]. The rest were nested within broader studies for which sample sizes had been determined based on the main objectives but were not reported, and none reported the statistical considerations used to estimate them. Similarly no study discussed whether missing data was encountered, the mechanisms of missingness or its implications on the final measure. One study examined the effect of differential weighting of items as suggested by a panel of experts, concluding that it made no difference to the eventual score's predictive validity [Ashton *et al.* 1999]. However none explored any potential changes due to sampling variation or different methods of score generation. Nevertheless reliability and validity testing was extensively done in all studies. Internal consistencies were tested using methods that compare performance over time or across different raters. Validity

checks included comparison with other tools for measuring the same construct and checks on between-scale correlations.

Outputs of the studies were broadly categorized into two: those with a composite indicator set proposed for measuring the complex construct being studied, and those which reported an overall score which was the result of summarising scores assigned to each item in the composite indicator list using some suitable method (such as summation of individual scores where there was no justification for weighting).

1.2.5. Discussion

Some useful techniques for measuring quality of health care and other complex constructs have been described in the literature. In general the process of constructing a measure involves first describing the perspective to be applied in the measurement, followed by identifying or generating a list of items potentially relevant to characterising quality in that perspective, refining the list and grouping items into scales which represent clinically-relevant domains of quality; finally using the items as criteria for measurement and where appropriate, summarising them into aggregate measures and reporting findings. However none of the studies discussed the implications of missing data. Additionally techniques for addressing uncertainty due to sampling variation and methodological differences are also poorly developed, and the need for easy-to-use tools to collect, process, and report assessment findings quickly and efficiently is still unmet.

As discussed earlier shortfalls in data management in health systems of many low-income countries are likely to pose special difficulties in obtaining complete data; indeed missingness of data in case record forms is commonplace [Gathara *et al.* 2011, Mwakyusa *et al.* 2006, English *et al.* 2004]. As such, a well-designed quality measure must include checks of data completeness and considerations for imputing or ignoring missing data, and analyses to check the robustness of the measure in the face of missing data. The reporting needs of different levels of the healthcare system must also be considered. Most of the studies above do not do this; where a summary measure is provided the unit of aggregation is at a higher level, usually the hospital. This limits the ability of the measure to distinguish between quality of care for individuals with different characteristics – e.g. diagnosis, co-morbidities, carer, all of which are potential modifiers of quality of care and outcomes – which would have been achievable using a

measure that encompasses quality of care across multiple diseases as an individual (patient)-level measure.

1.2.6. Summary

This literature review demonstrates the potential usefulness of process measures of quality of care. It highlights methods of developing aggregate measures that could be useful in a low-income country like Kenya where standards of care backed by good clinical evidence and supported by health policy makers are already in place but quality measurement is still rudimentary. The review also shows that methods to deal with challenges such as poor documentation of process of care and the multiplicity of guideline recommendations across illnesses and severity classifications are still poorly developed. This work was therefore set up to address some of these shortfalls while demonstrating the use of routine process-of-care data in identifying weaknesses in management of common childhood illnesses during a critical time of care when most deaths occur.

1.3. Objectives of the research work

1.3.1. Overall objective

The overall objective of the work presented in this thesis is to design and test a patient-level score to measure quality of the initial admission care for children under 60 months admitted with common childhood illnesses to hospitals in Kenya by focusing on multiple tasks that comprise the process of care.

1.3.2. Specific objectives

1. To design a quality measure that summarises hospital care for children admitted to hospital in a low-income setting:

    i. to identify process-of-care tasks and domains of care for children admitted to hospital with malaria, pneumonia or diarrhoea/dehydration;

    ii. to explore simple and intuitive scoring approaches for summarising process-of-care tasks into a measure of quality of care at different levels (domain, disease, patient, clinician, hospital) and to combine them to an overall score;

iii. to identify a coherent set of methods and tools for reporting the measure to its various audiences.

2. To demonstrate the validity, reliability and potential advantages of the quality of care measure:

   i. to show that the items within domains of the quality of care measure are internally consistent;

   ii. to explore links between this measure and outcomes of care;

   iii. to test the generalisability of the measure to different situations.

1.4. Background of the research work

This thesis relies on and extends previous studies undertaken by the Health Services Unit (previously known as the Child and Newborn Health Group and later Health Services Research Group) which is in the Public Health Research Department of the KEMRI-Wellcome Trust Research Programme in Nairobi, Kenya, to which the author belongs. Key elements of the Unit's work which this thesis builds upon are the development of the Basic Paediatric Protocols ('guidelines') for management of childhood illnesses, a training programme for health workers on the use of the guidelines, and a trial to investigate the effect of training and the use of the guidelines, supervision and feedback on the quality of hospital care for children admitted in Kenyan hospitals with the illnesses responsible for majority of deaths in this setting.

1.4.1. Development of the basic paediatric protocols

Previous studies had noted that hospital care for children in low income countries was poor and that improvement was possible by promoting the application of existing evidence into practice [English *et al.* 2004, Nolan *et al.* 2001]. Guidelines on how to provide care were rare and even where available they had neither been updated nor put to actual use. Therefore evidence-based clinical practice guidelines (CPGs) for paediatric care in Kenya were developed to improve diagnosis and treatment of childhood illnesses [Irimu *et al.* 2008]. The guidelines were developed through adaptation of the World Health Organization (WHO) guidelines to focus on illnesses responsible for a majority of childhood deaths in this setting, namely anaemia, birth asphyxia, diarrhoea/dehydration, malaria, meningitis, neonatal sepsis, pneumonia and

prematurity/low birth weight, and also on basic life support including resuscitation management of shock and convulsion.

The CPGs were designed to aid clinicians in providing the best care possible within the available resources because most children were initially attended to by nurses, clinical officers (health workers with a diploma in clinical medicine), medical interns and junior medical officers who did not have extensive training or experience in paediatric care. Health policy makers within the Ministries of Health, medical schools and research organisations were co-opted to advise, adapt and further develop the CPGs culminating in the production and publication of the 'Basic Paediatric Protocols' booklet [MoH 2006]. Involving a large range of stakeholders, it was hoped, would help facilitate widespread acceptance of the guidelines.

A training programme on the use of the guidelines, named Emergency Triage Assessment and Treatment plus admission care (ETAT+), was developed, tested at a national referral hospital, a smaller rural hospital and a medical training centre [Irimu *et al.* 2008] and refined using feedback from these pilot runs. The ETAT+ course includes lectures and practical training to impart knowledge and build skills on life support using the airway-breathing-circulation-disability (ABCD) model. Eventually, increasing need and demand for ETAT+ training spurred the continued scaling up of training coverage in local and regional hospitals and medical schools [English *et al.* 2011].

1.4.2. The Kenyan district hospitals trial

A cluster-randomised controlled trial of an intervention to improve referral care for children admitted to Kenyan district hospitals with acute illnesses began in September 2006 and was delivered over 18 months until April 2008 with a 12-month continuation phase in the intervention sites up to April 2009 to assess whether any changes observed were sustained [Ayieko *et al.* 2011]. The intervention was to promote and support hospital implementation of the CPGs described above. Elements of the intervention included a five-and-a-half day training programme for health care workers on the ETAT+ approach to case management in line with guidelines, external supervision by paediatricians from the study team, an on-site facilitator at each hospital to provide a link to the external supervisor and support for general delivery of the intervention.

Practice guideline booklets, job aids such as drug and fluid dose charts and structured paediatric admission record forms were provided to both groups at the start of the trial. In place of the training programme the control group received a one-and-a-half day lecture introducing the guidelines. Surveys were conducted every six months. At each survey data on case management practices were collected from a sample of approximately 400 admission records at each hospital selected using their admission dates from a randomisation list generated using Stata™. Both groups of hospitals received written feedback of survey findings three to six weeks after each survey. The intervention hospitals additionally received face-to-face feedback.

1.4.3. Use of an assessment score to measure health workers' adherence to guidelines

In investigating the effect of the intervention in the Kenyan district hospitals study, health worker performance was assessed using process-of-care indicators of adherence to key guideline recommendations for treatment of malaria, pneumonia and diarrhoea/dehydration. Binary variables were defined for process-of-care tasks, errors of treatment and supportive care, such as correct drug, fluid or oxygen prescription, and use of recommended severity classification of illness. Proportions of admission records in which the admitting clinicians correctly performed 13 indicator tasks were reported for each hospital, and the differences in proportions in the intervention vs. control group for each indicator used as a hospital-level measure of the intervention effects.

Separately to summarise health workers' assessment of clinical signs and symptoms – a patient-level measure – an assessment score for each child was reported by dividing the number of documented signs and symptoms by the total number expected depending on the child's diagnosis. For example if a child was diagnosed with severe pneumonia and severe malaria it was expected that 8 signs and symptoms for pneumonia and 6 for malaria be documented. If only 10 signs were documented then the child's assessment score was $^{10}/_{14}$ i.e. 0.625. Assessment scores were averaged at hospital level then used to compare the intervention and control groups.

Both methods of assessing the effectiveness of the intervention provided evidence of a greater improvement in the intervention than control hospitals by the main study end-point. However this approach to reporting the trial endpoints had potentially limited interpretability for several reasons. First, the assessment score represented only one

aspect ('domain') of process of care as laid out in the guidelines. The 13 indicators spanning diagnosis and treatment were reported separately as independent tasks and this was unable to shed light on each individual's experience of the continuum of care from assessment to diagnosis to treatment. Additionally measures could only be reported for each specific illness independently. For these reasons it was not possible to summarise overall hospital or clinician-level performance across tasks and across diseases.

Secondly, some guideline-recommended aspects of correctness of diagnosis and treatment such as correct route, frequency and duration of drug therapy were not included as trial outcome indicators, as shown in Table 1.4-1.

Table 1.4-1: A comparison of items in the assessment score for malaria in the Kenyan district hospitals study with the full guideline recommendations for malaria case management

Signs and symptoms	**Part of malaria assessment score/ process indicators**
Fever indicated as present or absent or temperature indicated	Yes
Convulsions indicated as present or absent	Yes
Acidotic breathing indicated as present or absent	Yes
Degree of pallor indicated as '0', '+' or '+++'	Yes
Ability to drink or breastfeed reported in the affirmative or negative	Yes
Level of consciousness classified as 'alert', 'responsive to voice', 'responsive to pain' or 'unconscious'	Yes
Diagnosis	
Malaria is classified explicitly as 'severe' or 'non-severe'	Yes
Classification is consistent with observed signs and symptoms	No
Treatment	
Coartem or quinine only is prescribed for severe or non-severe malaria respectively	No
Route of administration of Coartem is oral, quinine is intramuscular or intravenous	No
Number of Coartem tablets prescribed is 1, 2, 3 or 4 for child weighing 5.0–14.9kg, 15.0–24.9kg, 25.0–34.9kg or 35.0kg+, respectively; loading dose of quinine is 20mg/kg *and* maintenance dose is 10mg/kg	Yes
Frequency of daily dosing of each drug is twice daily	No
Duration of treatment with antimalarial is specified	No

Thus a clinician could have been rated highly on the assessment with scores of 6/6 but still be failing to provide optimal care, with diagnosis and treatment scores of 0 or 1, as is the case in patient IDs 3, 6, 10, 12 and 16 in Appendix A.4. Thirdly, many children were diagnosed as having more than one illness yet the method employed in combining assessments for different illnesses into an assessment score left no scope for distinguishing between care for the different illnesses; this would have been useful for

identifying illnesses for which care was poor. Missing documentation of clinical enquiry was common in these data, as shown in Table 1.4-2; not only did this characterise care as poor but also made it difficult to determine the correctness of classification and treatment according to guidelines. For example less than half of the 8,090 children with malaria had all assessments documented.

Table 1.4-2: Completeness and appropriateness of assessment, classification and treatment of malaria in the Kenyan district hospitals study according to guideline recommendations

Assessment (fraction performed)	**Number of cases (percentage of total malaria cases)**		
Complete ($^6/_6$)	3,887 (48.1%)		
High ($^4/_6$ to $^5/_6$)	1,172 (14.5%)		
Low ($^1/_6$ to $^3/_6$)	2,935 (36.3%)		
Incomplete ($^0/_6$)	96 (1.2%)		
Classification			
Severe	2,954 (36.5%) classified, 1,728 (58.5%) of these correctly		
Non-severe	1,851 (22.9%) classified, 886 (47.9%) of these correctly		
Not classified	3,285 (40.6%)		
Treatment	**Severe, n=2,954**	**Non-severe, n=1,851**	**Not classified, n=3,285**
Drug correctly selected	2,533 (85.7%)	401 (21.7%)	(cannot be determined)
Route correctly specified	1,935 (65.5%)	382 (20.6%)	261 (7.9%)
Dose correct for body weight	20 (0.7%)	346 (18.7%)	76 (2.3%)
Frequency correctly specified	1,738 (58.8%)	386 (20.9%)	202 (6.1%)
Duration (any) specified	3 (0.1%)	401 (21.7%)	966 (29.4%)

Classification of severity was a key guideline recommendation yet only 4,805 (59.6%) cases had this done; the rest were either incorrectly classified or unclassifiable due to insufficient assessment. Consequently only 2,934 children received the correct drug for their severity classification, the remaining 5,156 being a mix of incorrectly treated children and those for whom correctness of treatment could not be determined. Classification and treatment errors are not equal in relation to their perceived consequences on quality of care [Rowe *et al.* 2003]: a child incorrectly classified and treated for a more severe illness than they actually have would still likely have a favourable outcome but resources would have been wasted; on the other hand a child being inappropriately treated for an illness of lower severity than they had could be at risk of adverse outcomes.

1.4.4. Rationale for scope of proposed work

The work undertaken in this thesis will attempt to demonstrate how to measure the process of care for under-5 year olds admitted to hospital with the three commonest childhood illnesses in a low-income setting – namely pneumonia, diarrhoea and malaria. This is because these illnesses are responsible for more than half of deaths from infectious diseases in childhood. Recent estimates show that pneumonia was the single biggest driver of under-5 mortality, the cause of between 1.1 and 1.6 million deaths annually from 2008 to 2010 while diarrhoea was responsible for 15% of under-5 mortality worldwide in 2008 and 9.9% in 2010 [Liu *et al.* 2012, Black *et al.* 2010]. Malaria was the cause of 564,000 deaths (7.4% of total mortality). Additionally, as illustrated in sections 2.2 and 2.3, they are also responsible for a majority of hospital admissions in Kenya.

Although the global burden of malaria morbidity and mortality are on the decline in many places [WHO 2011] – mostly due to the development of effective treatment and control strategies – there is still some evidence that use of inappropriate treatments or incorrect use of recommended treatments is widespread [Ajayi *et al.* 2008, Oshikoya & Ojo 2007, Nshakira *et al.* 2002]. These missed opportunities to provide optimum care likely lead to preventable deaths that unnecessarily increase mortality. This thesis forms a potential starting point to the improvement of care by describing the systematic development of a tool for identifying weaknesses in care delivery through measurement of process of care.

1.4.5. The author's contribution to previous work

The author was involved with the development stages of the ETAT+ training programme as a trainee. Feedback from the early trainees was used to refine the training programme. During the main trial he supported and supervised data collection, collated survey findings, wrote and disseminated post-survey reports to participating hospitals and conducted data management. He also helped refine data collection and entry procedures for the trial and developed operating procedures for electronic data collection in the last three surveys. Many of the techniques he developed at this stage have been adapted for use in data collection and management for other studies at the Programme since then. At the end of the study he analysed data and co-authored key reports of main and secondary findings [Opondo *et al.* 2011, Ayieko *et al.* 2011,

Opondo *et al.* 2009]. While developing the proposal for this work he undertook the review of trial endpoints which highlighted the weaknesses of the assessment score and the other outcome measures, and the need to expand their scope.

1.5. Summary

The primary and secondary analyses of the Kenyan district hospitals trial data demonstrated the need to expand the scope of quality measures through: identification and inclusion of all guideline-recommended aspects of process of care to improve the ability of the measure to characterise a patient's admission episode; development of summary measures separately for each illness, to enable identification of illnesses for which care is poor and to better characterise care for children with co-morbidities; domain-level summaries of quality within and across illnesses to better identify problem areas of care; and identification of different forms of missing data and strategies to handle them and their consequences on quality of care evaluation.

Chapter 2 – Data for Designing and Testing the Proposed Measure

2.1. Introduction

2.1.1. Rationale for choice of data

The poor state of clinical data capture, storage and retrieval in Kenya and many low income countries makes the quality measurements targeted at improving care difficult to undertake. It has long been recognised that data quality could be greatly improved by the use of dedicated electronic medical records and information systems such as those which support the process of care in many developed countries. However these are rare in the Kenyan health system. The use of these technologies has to date been limited to externally-funded programmes focused on single diseases or one group of diseases [Tierney *et al.* 2010, Forster *et al.* 2008] and collaborations between select health facilities and academic institutions [Rotich *et al.* 2003, Hannan *et al.* 2001, Hannan *et al.* 2000] with varied success at actually improving patient records and patient care.

Despite the poor state of clinical data management systems, procedures for manual data retrieval from existing paper-based patient records have effectively been used for data collection in trials and observational studies examining processes of care. For example detailed case record data have previously been successfully obtained by retrospective random sampling and abstraction of patients' admission records [MoH 2013, Ayieko *et al.* 2011]. This is a rapid, low-cost and non-invasive approach which provides clinical data that reasonably replicate the information stored in most Kenyan hospital records. It is often data summarised from admission records which are eventually used to generate morbidity and mortality statistics at hospital level; these are then collated all the way up to the national level to produce a country's health indicators. Patient records are thus an important source of information for inference on the state of the health system. The data obtained from them are, in the absence of better approaches to quality assessment, very important component of a quality of care measure intended for routine use.

2.1.2. Intended uses for the data

Procedures for setting up a new method of measurement recommend that the measure undergoes validation after the design phase. Internal validity of the measure is tested

during the design phase using statistical methods described in detail in Chapter 4. Subsequently, external validity is demonstrated by the degree to which the procedures for creating the measure are applicable to a different situation. This can be achieved through systematic replication [Barlow & Hersen 1984], which is an extension of procedures undertaken in setting up the measure to a validation dataset arising from a routine care setting. Such data are preferred because the measure is intended for characterisation of routine care.

2.2. Data for designing the process of care measure

The Kenyan district hospitals data come from 12,036 admission episodes of children aged less than 60 months admitted to hospital with acute illnesses included in the inpatient IMCI guidelines, namely: malaria, pneumonia, meningitis, malnutrition, diarrhoea with or without dehydration, neonatal sepsis, prematurity and/or low birth weight. The data were collected retrospectively from paediatric admission records in 8 first-level referral hospitals in Kenya whose characteristics are summarised in Table 2.2-1. The hospitals each had between 1,100 and 4,500 paediatric admissions and 1,200 deliveries annually and were deliberately selected to be representative of the variety of rural Kenyan hospitals.

Table 2.2-1: Characteristics of hospitals in the Kenyan district hospitals study

Hospital	Malaria transmission setting	Antenatal HIV prevalence (High=>10%, Mod=5-10%)	Infant mortality rate per 1,000	Catchment population with income < $2 a day	Paediatric admissions per year	Paediatrician and Medical Officer interns
H1	Intense	High	>100	50-70%	3,500	No
H2	Highland	High	~70	50-70%	5,000	Yes
H3	Low	Moderate	~40	~35%	3,300	No
H4	Arid	Moderate	~70	50-70%	1,700	No
H5	Intense	High	>100	50-70%	2,500	No
H6	Arid	Moderate	~70	50-70%	1,100	No
H7	Highland	High	>100	50-70%	4,500	Yes
H8	Low	Moderate	~40	~35%	1,800	No

H1 - H4 are intervention hospitals, H5 - H8 are controls

Adapted from Opondo et. al 2009

The admissions occurred over a 40-month period between December 2005 and March 2009 inclusive, during the study described briefly in section 1.4.2 [Ayieko *et al.* 2011]. Four 6-monthly surveys were undertaken at each hospital, the first one focusing on

admissions in the 9-month period preceding it. Two additional surveys were conducted at the intervention group of hospitals after the active intervention had ended to explore any potential residual effects. The target sample size per hospital at each survey was 400. In many of the smaller hospitals this was often not achieved even when all paediatric records meeting the inclusion criteria were included in the sample. Table 2.2-2 shows the distribution of admission episodes across hospital and time and the number of admitting clinicians per hospital. Variables collected from the admission records were: children's age, sex, weight, vaccination status; history of current episode of illness including presenting symptoms and length of illness; clinical signs including the state of airway, breathing, circulation, neurological and gastrointestinal function, nutritional status; admission diagnosis; laboratory tests and their results if available; treatments prescribed (drugs, fluids, blood, oxygen, nutritional support) including their doses, routes, frequencies and durations of administration; date of end of admission and whether they were alive, dead or referred for further treatment, and whether a check-up date was scheduled. The data collection tool is shown in Appendix A.5.

The data were abstracted over 1 to 2 week periods during the 6-monthly hospital visits. Of the 12,036 admission episodes, 2,450 were abstracted at the baseline survey followed by 2,119 and 2,176 in the next 2 follow-up surveys. Data for the study's main end-point were collected during the fourth survey in which 2,714 records were abstracted. A total of 2,577 records were abstracted in the post-intervention period. Linked to each admission episode were unique ID numbers which identified admitting clinicians; these were further linked to their individual characteristics such as age, gender, qualification, years of practice and trainings attended. There were 557 clinicians uniquely identified across the 8 hospitals. There was a separate group of clinicians who were not identified; these were mostly trainees who spent a minimum amount of time in the paediatric wards and only attended to a small proportion – 1.6% (194) – of admissions. That each child's admission episode is linked to the admitting clinician, several observations are linked to each clinician and several clinicians are linked to a hospital indicate that these data are hierarchical; therefore hierarchical analytical techniques are required for statistical inference.

Table 2.2-2: Number of episodes at each hospital across surveys, and admitting clinicians per hospital

	Survey							
Hospital	Baseline	1st follow-up	2nd follow-up	End-point	1st post-intervention	2nd post-intervention	**Total (percentage)**	**No. of clinicians***
H1	327	315	309	349	333	366	**1,999 (16.6%)**	**101**
H2	334	342	356	351	347	323	**2,053 (17.1%)**	**55**
H3	288	277	305	362	310	345	**1,887 (15.7%)**	**164**
H4	342	219	299	351	216	336	**1,763 (14.6%)**	**29**
H5	142	153	175	366			**836 (6.9%)**	**32**
H6	331	224	231	211			**997 (8.3%)**	**29**
H7	348	334	250	373			**1,305 (10.8%)**	**83**
H8	338	255	251	351			**1,195 (9.9%)**	**64**
Total	**2,450**	**2,119**	**2,176**	**2,714**	**1,206**	**1,370**	**12,035†**	**557**

H1 - H4 are intervention hospitals; H5 - H8 are controls and were not part of the last two surveys; * Number identified – unidentified clinicians were lumped together into a single separate group in each hospital. †One observation is missing a survey indicator

Processes of care undertaken during the initial 24 hours of admission – the target of the intervention – were extracted from these admission records using the data collection tool. Processes examined included documentation of clinical signs and symptoms, diagnostic tests, illness diagnosis and severity classification, drug prescription, and post-admission care. Outcomes data collected included status at the end of hospital stay (discharged alive, dead, referred to another hospital or absconded) and date of discharge from which duration of admission was calculated. Most of the data were categorical, indicating the presence, absence or missingness of key features of illness from the case records; there were however a few continuous variables, including age, weight, height, temperature, duration of symptoms, respiratory rate and heart rate.

The process of care for malaria, pneumonia and diarrhoea are the focus of this work for reasons discussed in section 1.4.4. Figure 2.2-1 shows that these three diseases were responsible for 8,090, 5,231 and 2,543 admission diagnoses respectively, of varying severity across the six surveys conducted over the course of the study.

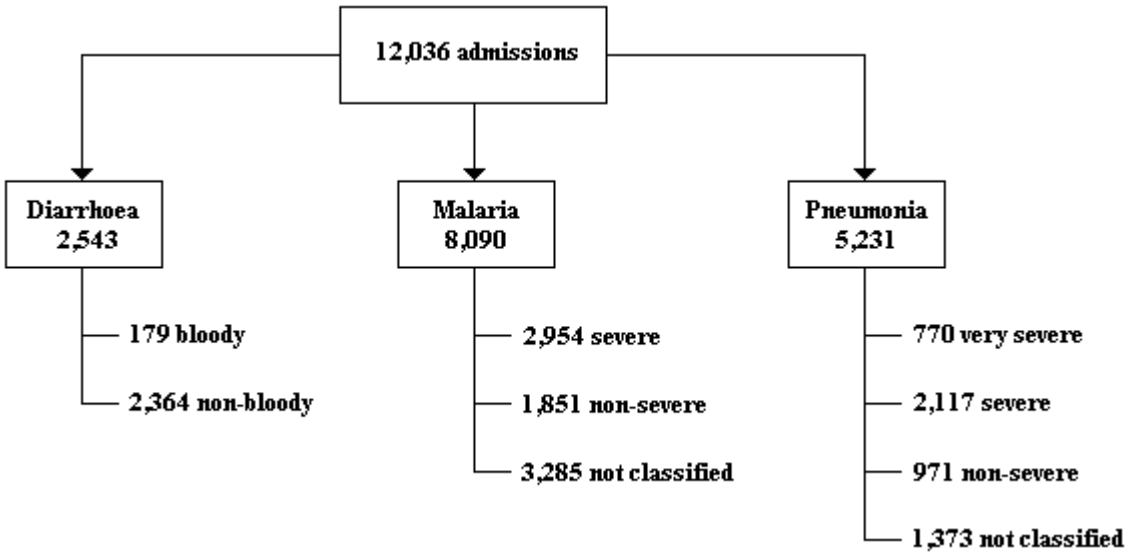


Figure 2.2-1: Summary of diagnoses and illness severity classifications observed in the data for designing the process of care measure

The diagnoses are not mutually exclusive within patient

There were 6,150 (51.1%) episodes with only one of the three diseases diagnosed. Another 4,188 (34.8%) episodes had two diseases diagnosed; of these 2,832 were of malaria and pneumonia, 1,142 of malaria and diarrhoea, and 223 of pneumonia and diarrhoea. Only 466 (3.7%) episodes had all three diseases diagnosed. There were 1,252 (10.4%) episodes which did not feature any of the three diseases – this means only 10,784 of the 12,036 admission episodes were relevant to this work. The largest

proportion of the admission episodes were observed at the baseline survey (2,450 – 20.4%) and fourth survey which was also the main study end-point (2,714 – 22.6%). In the first follow-up (second) and second follow-up (third) survey 2,119 (17.6%) and 2,176 (18.1%) episodes respectively were observed. The fifth and sixth surveys were undertaken to examine whether any intervention effects in the main study could be sustained upon withdrawal of the intervention; they were conducted in the intervention hospitals only. For this reason the number of admissions in this phase of the study was much lower – 1,206 (10.0%) and 1,370 (11.4%) respectively – than each of the four previous surveys. Table 2.2-3 details the distribution of disease episodes over time.

Table 2.2-3: Number of episodes of each disease/comorbidity across surveys

	Survey						
Disease	Baseline	1st follow-up	2nd follow-up	End-point	1st post-intervention	2nd post-intervention	**Total**
None*	262	233	254	234	122	147	**1,252**
Malaria	953	646	581	750	312	437	**3,679**
Pneumonia	304	336	350	419	157	173	**1,739**
Diarrhoea**	108	102	111	208	99	104	**732**
Malaria+Pneumonia	548	494	600	579	333	269	**2,823**
Malaria+Diarrhoea	225	172	155	328	102	159	**1,141**
Pneumonia+Diarrhoea	13	55	35	64	25	31	**223**
Malaria+Pneumonia+ Diarrhoea	37	81	90	132	56	50	**446**
Total	**2,450**	**2,119**	**2,176**	**2,714**	**1,206**	**1,370**	**12,035**†

*May include other diseases than the three considered here; **Diarrhoea with or without dehydration at each mention; †One observation is missing a survey indicator

Mortality was similar across groups over time, albeit nominally higher in the intervention hospitals in three of the four time points when data was available from both groups as shown in Table 2.2-4. The outcomes for 295 (2.5%) children were missing.

Table 2.2-4: Percentage (95% confidence intervals)[n] of deaths in the intervention and control hospitals across surveys

Survey	**Control**	**Intervention**	**Overall**
Baseline	5.7 (4.3 – 7.0) [1,111]	7.0 (5.6 – 8.4) [1,274]	**6.4 (5.4 – 7.4) [2,385]**
1st follow-up	7.1 (5.4 – 8.7) [935]	8.2 (6.6 – 9.9) [1,094]	**7.7 (6.5 – 8.8) [2,029]**
2nd follow-up	8.9 (7.0 – 10.8) [858]	8.8 (7.2 – 10.4) [1,228]	**8.8 (7.6 – 10.0) [2,086]**
End-point	6.8 (5.4 – 8.2) [1,276]	10.5 (8.9 – 12.1) [1,388]	**8.7 (7.7 – 9.8) [2,664]**
1st post-intervention	—	10.9 (9.1 – 12.6) [1,206]	**10.9 (9.1 – 12.6) [1,206]**
2nd post-intervention	—	8.8 (7.3 – 10.3) [1,370]	**8.8 (7.3 – 10.3) [1,370]**
Overall	**7.0 (6.2 – 7.8) [4,180]**	**9.1 (8.4 – 9.7) [7,560]**	**8.3 (7.8 – 8.8) [11,740]**

Only data collected over the active intervention phase of the study – 9,459 cases records in the first four surveys and specifically the 8,476 with at least one of the three diseases – are of interest to this work. Over this period there was a significant difference between the two groups of hospitals in the proportions of admissions attributed to the three

diseases. There were not only more admissions in the intervention than control hospitals at each time-point but also more episodes featuring at least one of the three diseases. At the baseline and first follow-up survey 89.5% (95% CI 87.8% to 91.1%) and 89.6% (95% CI 87.8% to 91.4%) respectively of episodes in the intervention group, and 89.1% (95%CI 87.3% to 90.9%) and 88.3% (95% CI 86.3% to 90.3%) respectively of episodes in the control group were attributed to at least one of the three diseases. By the second follow-up and main end-point surveys these proportions were significantly higher in the intervention group – 91.4% (95% CI 89.9% to 93.0%) and 93.8% (95% CI 92.6% to 95.1%) respectively – than the control group – 84.0% (95% CI 81.6% to 86.4%) and 88.7% (95% CI 87.0% to 90.4%) respectively.

Most children had a diagnosis of malaria or pneumonia or both; diarrhoea alone or in combination with another disease was less commonly diagnosed. Children in the intervention group were more likely to have at least one of the three diseases diagnosed, as shown in Table 2.2-5. This was most likely because the intervention encouraged clinicians to record a diagnosis for any child whose signs and symptoms were consistent with the criteria for that diagnosis laid out in the guidelines. Overall, there were 8.4% more children the intervention arm than the control arm in this group of children.

Table 2.2-5: Number (percentage) of episodes of each disease and disease combination in the intervention and control hospitals in the first four surveys

Disease(s)	**Control**	**Intervention**	**Total**
None	531 (54.0)	452 (46.0)	**983 (10.4)**
Malaria	1,414 (48.3)	1,516 (51.7)	**2,930 (31.0)**
Pneumonia	738 (52.4)	671 (47.6)	**1,409 (14.9)**
Diarrhoea	252 (47.6)	277 (52.4)	**529 (5.6)**
Malaria + Pneumonia	822 (37.0)	1,399 (63.0)	**2,221 (23.5)**
Malaria + Diarrhoea	404 (45.9)	476 (54.1)	**880 (9.3)**
Pneumonia + Diarrhoea	83 (49.7)	84 (50.3)	**167 (1.8)**
Malaria + Pneumonia + Diarrhoea	89 (26.2)	251 (73.8)	**340 (3.6)**
Total	**4,333 (45.8)**	**5,126 (54.2)**	**9,459 (100)**

Demographic characteristics and other features of the admission episodes, shown in Table 2.2-6, indicate that the groups were balanced with respect to most exposure and outcome variables except mortality which was slightly higher in the intervention group.

Table 2.2-6: Demographic characteristics and characteristics of the admission episodes in the intervention and control hospitals in the first four surveys

Characteristic	**Control**	**Intervention**	**Overall**
Age in months, median (IQR)	12 (7 – 24)	13 (7 – 24)	**12 (7 – 24)**
Gender, percent male (95% CI)	45.6 (44.1 – 47.2)	44.8 (43.6 – 46.0)	**45.1 (44.1 – 46.1)**
No. of diseases, median (IQR)	1 (1 – 2)	1 (1 – 2)	**1 (1 – 2)**
Duration of admission, median (IQR)	3 (2 – 5)	3 (2 – 6)	**3 (2 – 5)**
Mortality, percent (95% CI)	7.0 (6.3 – 7.8)	9.1 (8.4 – 9.7)	**8.3 (7.8 – 8.8)**

2.3. *Source of data for external validation*

Two datasets are available for external validation. One is a cross-sectional survey of 22 training hospitals in Kenya undertaken in 2012 by the Ministry of Health [MoH 2013], and the other is a retrospective cross-section of paediatric admissions in 7 hospitals. The aim of the 22-hospital survey was to evaluate routine service delivery in maternal, neonatal, paediatric, internal medical and surgical units of the surveyed hospitals. Data on process of care were collected from case record forms which were identified from inpatient registers starting from 31st May 2012 backwards until the desired sample of approximately 60 records per unit per hospital was achieved. The 7-hospital cross-sectional data were collected in parallel with a trial seeking to demonstrate non-inferiority of oral amoxicillin when compared to injected penicillin for the treatment of severe pneumonia. They span the period between September 2011 and August 2013 and include all children with non-surgical diagnoses admitted to the paediatric wards during the trial other than those randomised to either arm of the study. The purpose of collecting observational data in this trial was to determine whether children recruited into the trial differed systematically from other children receiving routine care.

Together these two validation datasets count for 12,772 admission episodes over a 2½ year period not covered by the Kenyan district hospitals data. They are suitable for validating the proposed measure for a number of reasons. Firstly, they are broadly representative of the variety of practice settings which exist in the Kenyan health system compared to the relatively smaller sample of four hospitals per arm in the Kenyan district hospitals study which were systematically selected for inclusion into the study based on considerations of feasibility. There is however some overlap in the contribution of the various hospitals to the three datasets: 6 of the 8 hospitals in the district hospitals study were also included in the MoH survey of hospitals, and 4 of the 22 hospitals in the MoH survey were among the 7 sites that were part of the study evaluating all hospital admissions linked to the pneumonia study. This is illustrated in Appendix A.6. Table 2.3-1 details the numbers of admission episodes and admitting clinicians in each of the two datasets.

Secondly, the care provided to children in the validation datasets can be considered to be ‘routine’ because it was not subject to the influences of an external intervention specifically aimed at changing the quality of care. This is in contrast with the district

hospitals study which involved provision of different intensities of an intervention to both groups of hospitals, which potentially altered the environment of care making it less representative of the process of care in a routine setting. For example one component of the intervention was the provision of structured admission record forms to clinicians to facilitate documentation of the process of care. Although similar job aids are provided in many hospitals in Kenya their actual use is varied. One study established that when structured admission record forms were provided to clinicians in the absence of any other supervision or incentive, adoption ranged from 50% to 84% [Mwakyusa *et al.* 2006]. In the district hospitals study only 34 out of 2,450 admission episodes (1.4%), all from the same hospital, were made on a structured admission form at baseline before the intervention had a chance to catch on.

To summarise, the validation data represent a different case mix of the three diseases, as shown in Figure 2.3-1 and Table 2.3-2, different mortality risks – 6.0%, 95% CI 5.5 to 6.4 in the pneumonia study, and 4.7%, 95% CI 3.7 to 6.0 in the MoH survey data – and different demographic characteristics, presented in Table 2.3-3, as compared to the district hospitals data. Nevertheless there is sufficient overlap between the variables collected in them and those in the district hospitals data to allow for determination of whether the methods developed for measuring the quality of process of care in the latter situation are applicable to a more general setting.

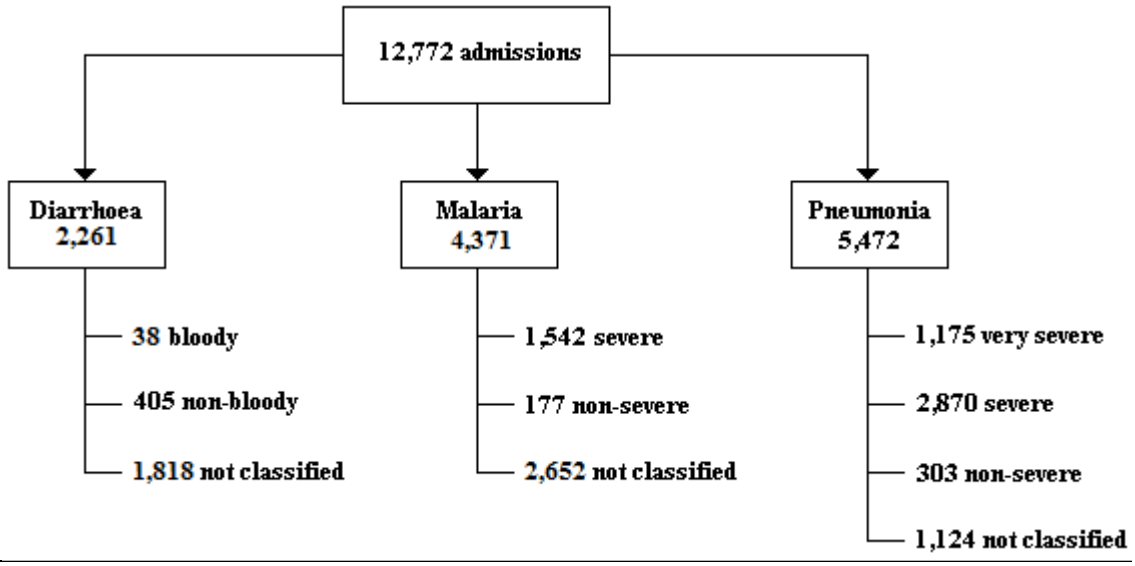


Figure 2.3-1: Summary of diagnoses and illness severity classifications observed in the data for external validation of the process of care measure

The diagnoses are not mutually exclusive within patient and exclude diseases not of interest to this work.

Table 2.3-1: Number of admission episodes and clinicians across hospitals in the validation datasets

Hospital ID (ID in DH dataset)	No. of episodes	No. of clinicians	Dataset	Hospital ID (ID in DH dataset)	No. of episodes	No. of clinicians	Dataset
1	1,952	70	Pneumonia trial	14	2,520	60	Pneumonia trial
1	63	10	MoH survey	15	60	30	MoH survey
2 (H5)	59	11	MoH survey	16	59	15	MoH survey
3	541	31	Pneumonia trial	17	60	10	MoH survey
4	60	20	MoH survey	18	2,181	122	Pneumonia trial
5	1,756	73	Pneumonia trial	18	75	11	MoH survey
5	60	19	MoH survey	19	60	9	MoH survey
6	60	13	MoH survey	20	59	20	MoH survey
7 (H1)	60	21	MoH survey	21	61	17	MoH survey
8 (H6)	60	5	MoH survey	22	60	10	MoH survey
9 (H7)	60	21	MoH survey	23	447	50	Pneumonia trial
10	61	18	MoH survey	23	60	16	MoH survey
11 (H2)	2,037	28	Pneumonia trial	24	60	8	MoH survey
12	60	14	MoH survey	25	60	16	MoH survey
13 (H3)	60	30	MoH survey	*1 observation in the pneumonia trial dataset is missing its hospital ID			

Table 2.3-2: Number (percentage) of episodes of each disease and disease combination in the two validation datasets

Dataset	None	Malaria	Pneumonia	Diarrhoea	Malaria and Pneumonia	Malaria and Diarrhoea	Pneumonia and Diarrhoea	Malaria, Pneumonia and Diarrhoea	**Total**
POD*	3,956 (34.6)	1,165 (10.2)	3,846 (33.6)	1.381 (12.1)	263 (2.3)	71 (0.6)	729 (6.4)	24 (0.2)	**11,435 (100)**
PSS**	535 (40.0)	151 (11.3)	559 (41.8)	34 (2.5)	36 (2.7)	7 (0.5)	13 (1.0)	2 (0.2)	**1,337 (100)**

*POD = pneumonia trial observational data; **PSS = Ministry of Health survey data

Table 2.3-3: Demographic characteristics and characteristics of the admission episodes in the two validation datasets

Characteristic	Pneumonia trial	MoH survey	Overall
Age in months, median (IQR)	14 (7 – 34)	16 (8 – 30)	**14 (7 – 33)**
Gender, percent male (95% CI)	55.5 (54.6 – 56.4)	58.4 (55.7 – 61.0)	**55.8 (54.9 – 56.7)**
No. of diseases, median (IQR)	1 (0 – 1)	1 (0 – 1)	**1 (0 – 1)**
Duration of admission, median (IQR)	3 (2 – 6)	3 (2 – 6)	**3 (2 – 6)**
Mortality, percent (95% CI)	6.0 (5.5 – 6.4)	4.7 (3.7 – 6.0)	**5.8 (5.4 – 6.3)**

Chapter 3 – Scale Construction and Item Selection

3.1. Introduction

Measuring a construct such as quality of care involves assigning numbers or symbols to items related to it in a way that corresponds to their relationship with the construct [Rosenthal & Westen 2003]. For this reason constructing a scale for measurement begins with the identification of a set of such items. According to the domain sampling theory [Ghiselli *et al.* 1981, Kline 1960] there are many possible items which could be sufficiently related to any construct to justify their use in measurement, but it is not always possible to identify them all; selection therefore seeks to identify items which are most relevant for the intended application. Selected items are then summarised into a measure, often by adding up item scores. This summation may be weighted if there is good evidence of greater importance of some items over others; an unweighted sum of scores may however be preferred for its simplicity and intuitiveness. Subsequently the functional and numeric properties of the measure are tested and refined until a satisfactory measure is identified (Figure 3.1-1). These steps outline the approach to development of a novel measure that is adopted in this thesis.

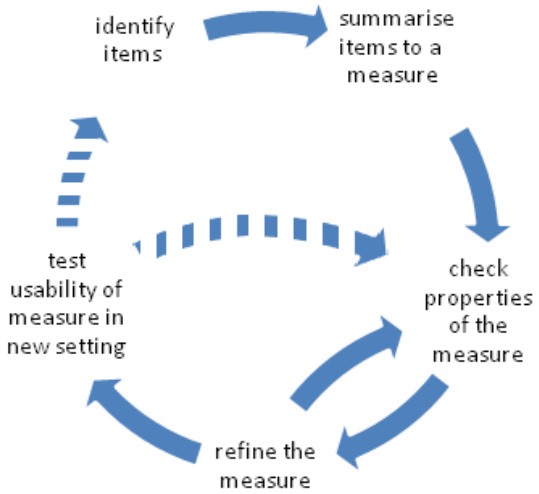


Figure 3.1-1: Generic steps in constructing a measure
Dashed arrows are optional loops through which candidate measures are modified

Functionally a measure and all its component parts should be seen to measure what it is intended to, and this constitutes its internal validity [Brewer 2000, Cozby 1993]. The

measure should also be externally valid, that is, applicable to a broad range of new settings which are sufficiently similar to those under which it was developed and to which it is intended to be used [Steckler & McLeroy 2008]. The measure should also be reliable; this refers to the consistency of measurement by its component parts or upon repeated use across time or by different raters [McDowell & Newell 2006]. A brief discussion of these concepts is presented in Appendix A.7.

Numerically a measure should ideally exhibit normality, approximate normality or normality upon suitable transformation. This ensures that the intuitive measures of central tendency and dispersion such as means and variances, and the simpler and more powerful parametric statistical methods are valid for inference. The components of an observed measure, X, are theoretically related to the quantity of a construct's attribute being measured, T, by:

$$X_i \sim T_i + e_i$$

where e is normally distributed with a mean of zero and standard deviation of 1 and uncorrelated with T and itself across i. Under these circumstances X should be normally distributed or at least tend towards normal following the central limit theorem. The numerical properties of a measure are thus largely determined by its level of measurement, a brief discussion of which is presented in Appendix A.8.

3.2. Approaches to measuring health-related constructs

There is precedence to the application of the generic approach illustrated in Figure 3.1-1 to the development of measures of health-related constructs. Many of these are well established measures with demonstrable relevance, reliability and validity which are still in use to date in their original or improved forms. Three specific applications of two broad examples are described here in brief.

3.2.1. Example 1: prognostic scores

The Glasgow Coma Scale (GCS) for evaluating and monitoring level of consciousness in individuals with brain trauma has been used successfully for almost thirty years, and attempts to replace it with other measures have mostly been unsuccessful due to its simplicity and practicality [Nye *et al.* 2012, Zuercher *et al.* 2009, Wijdicks 2006]. The GCS has three components, namely motor response, verbal response and eye opening.

Six grades of motor response are defined representing six distinct but ordered levels of motor activity. Similarly five levels of verbal response and four of eye opening are defined. Each component is scored separately as the patient's level of response on that component and an overall score is created by summing the component scores. Higher scores represent better neurological function therefore less severe injury (Table 3.2-1).

Table 3.2-1: Items scored in the Glasgow Coma Scale (GCS) and the Acute Physiology and Chronic Health Evaluation (APACHE) score

Glasgow Coma Scale (GCS)	**Acute Physiology and Chronic Health Evaluation (APACHE II) items and their scores for derangement from normal**
Grades of Best Motor Response 6 Carrying out requests 5 Localising response to pain 4 Withdrawal to pain 3 Flexor response to pain 2 Extensor posturing to pain 1 No response to pain	Age 0-6 Haematocrit 0-4 White blood cell count 0-4 Rectal temperature 0-4 Mean arterial blood pressure 0-4 Heart rate 0-4 Respiratory rate 0-4 Serum sodium 0-4 Serum potassium 0-4 Serum creatinine 0-4 Serum bicarbonate 0-4 Oxygen saturation 0-4 Alveolar-arterial oxygen concentration gradient 0-4 Arterial pH 0-4 History of immunocompromisation or organ failure 0-5 Glasgow Coma Scale score 0-15
Grades of Best Verbal Response 5 Oriented 4 Confused conversation 3 Inappropriate speech 2 Incomprehensible speech 1 No verbal response	
Eye Opening 4 Spontaneous eye opening 3 Eye opening in response to speech 2 Eye opening in response to pain 1 No eye opening	**Estimated mortality risk for total scores:** 0-4: 4%, 5-9: 8%, 10-14: 15%, 15-19: 25%, 20-24: 40%, 25-29: 55%, 30-34: 75%, >34: 85%.
Interpretation of total score: <9: severe injury, 9-12: moderate injury, 13-15: minor injury	

A similar approach is taken in the Acute Physiology and Chronic Health Evaluation (APACHE) score for classifying severity of illness for patients admitted for intensive care [Knaus *et al.* 1985]. Items indicative of acute and chronic health status are assigned scores mostly between 0 and 4 to represent their derangement from normal levels or levels associated with the best prognosis. For example, very high and very low heart-rates are scored 4, intermediately high and low ones 2 and normal rates scored 0. The item scores are summed up to a total APACHE score; higher scores represent increased risk of mortality.

There have been several studies of these associations which have shown high agreement between score-predicted and observed mortality [Minne *et al.* 2008, Zimmerman *et al.* 2006, Markgraf *et al.* 2000, Vassar *et al.* 1999, Zimmerman *et al.* 1998, Beck *et al.* 1997, Murphy-Filkins *et al.* 1996, Zhu *et al.* 1996]. Where agreement has been poor

attempts have been undertaken to refine and improve the scores. Refinements to the original APACHE score for example have produced the APACHE II and APACHE III scores which have even better accuracy in predicting mortality than the original score [Mann *et al.* 2012, Zali *et al.* 2012]. It is currently possible to predict a patient's mortality risk using their total score – these estimates are shown in Table 3.2-1.

3.2.2. Example 2: quality of life measures

Quality of life measures can be thought of as humanistic outcomes in contrast to the more traditional clinical and economic outcomes commonly reported in clinical research, quality of care and healthcare evaluation literature [Maloney & Chaiken 1999]. They are not only inherently subjective, but also highly confounded by many observable and non-observable factors such as those summarised in Figure 3.2-1. Many quality of life measures are multidimensional constructs, encompassing several items relating to physical and psychological well-being often described in terms of limitation of activities, but may also include items to quantify pain, vitality, cognition and general health perception [Smith *et al.* 2005].

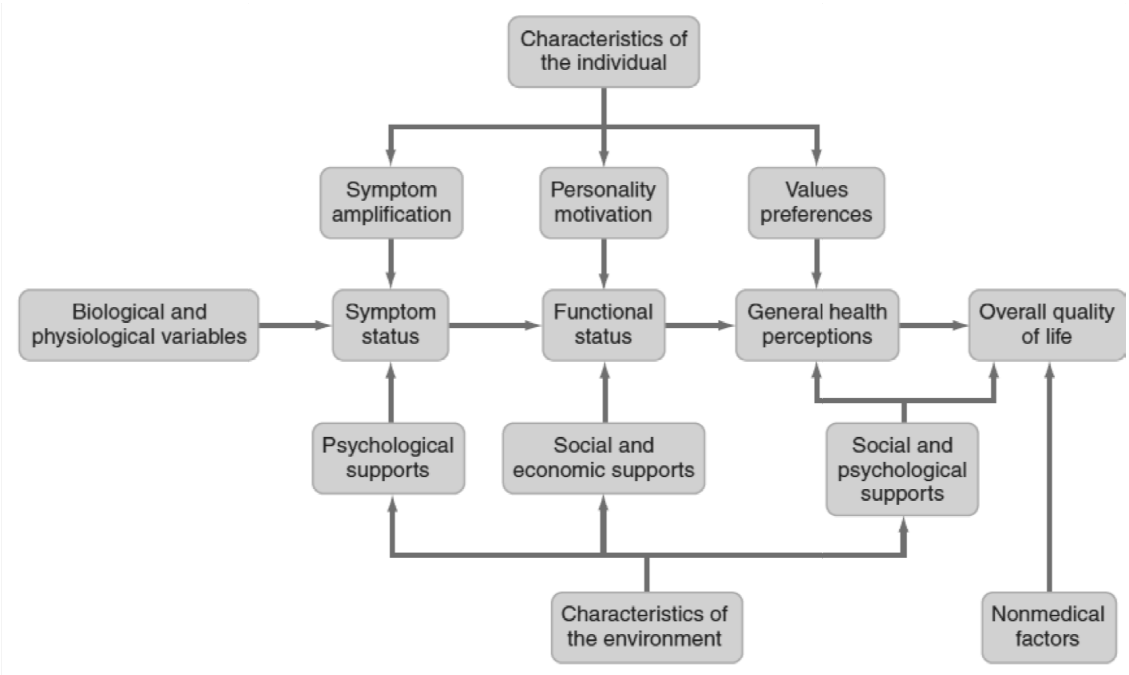


Figure 3.2-1: Determinants of quality of life in a health-related quality of life conceptual model [Wilson & Cleary 1995]

Items contributing to quality of life measured vary widely across measures: the original WHO Quality of Life Group Scale (WHO-QOL) has 15 items [Burckhardt & Anderson 2003], the Nottingham Health Profile (NHP) has 38 items [Hunt *et al.* 1985], and the Sickness Impact Profile (SIP) has 136 items measuring 2 dimensions of well-being [Gilson *et al.* 1975]. There is some overlap of items across the measures. These differences and their subjective nature could be part of the reason quality of life measures do not always yield consistent results even when used together in the same study [Klevsgård *et al.* 2002, Chetter *et al.* 1997]. Additionally, unlike with prognostic scores, it is not always possible to validate a quality of life measure using an objective outcome even when links to such an outcome may be obvious at face value. Such outcomes may simply be non-existent for that measure, or perhaps links with the construct it is thought to measure may be difficult to demonstrate due to complex, poorly-defined or highly confounded relationships.

Nevertheless none of these shortfalls invalidate their use as outcome measures in their own right. In fact quality of life measures have become the main tools for assessing health status according to patients' own experiences, allowing impacts of illnesses or interventions to be examined from patients' points of view [Smith *et al.* 2005]. Utility measures such as Quality Adjusted Life Years (QALYs) and Disability Adjusted Life Years (DALYs) have been derived from health related quality of life measures by combining them with measures of quantity of life. QALYs and DALYs are interpreted as population-averaged preferences of alternative health statuses [Smith *et al.* 2005] and are used in evaluating cost-effectiveness of many health interventions [Sassi 2006, Prieto & Sacristan 2003].

3.3. Item selection for quality of care measurement

3.3.1. General considerations in item selection

There are parallels between the properties of quality of life measures and prognostic scores, and the features of a quality of care measure which could inform the selection of suitable items for the measure. One is that if an objective or desired outcome – such as death or survival, physiological states and duration of hospital stay – is known or thought to be associated with the items to be used for constructing the measure, then item selection can be based on how well the items relate to the objective or desired outcome. This explains why items such as indicators of verbal and motor response,

which are physiologically related to level of incapacity and risk of death, are included in a prognostic score. Ideal as this approach may be in informing item selection for the quality of care measure, it is not always possible to clearly demonstrate associations between indicator items and an objective outcome, and it may simply be sufficient that there is a reasonable *expectation* that the candidate items relate to what could plausibly lead to desirable outcomes. Alternatively, suitable items for the proposed measure may be ones that relate to a targeted or preferred standard of care, just as QALYs and DALYs relate to preferred health statuses.

In this thesis it is the standards of care described by clinical practice guidelines for management of acute childhood illnesses which will inform item selection. This is a hybrid of the approaches to item selection for prognostic scores and quality of life measures: it is similar to the prognostic scores in that it implicitly links the items contributing to the measure – derived from guideline recommendations – to the best possible outcomes. It is also similar to the quality of life measures by focusing on items that measure how well what is done at the point of care is aligned with what ought to be done, in so doing linking the measure to an established desired level of care, just as items in quality of life measures relate to desired levels of physical and mental well-being. It thus provides a good starting point to flagging weaknesses in the process of care which if reduced or eliminated could improve outcomes and overall care. Importantly, this approach solely focuses on appropriateness of processes at the point of care. It does not attempt to describe or quantify other contextual factors and health system complexities known to influence quality of care such as (lack of) resources, institutional and government policies, and organizational culture. In so doing it potentially represents a major leap forward in a low-income setting where routine quality of care measurement and reporting is virtually non-existent making it impossible to examine causes of variation in outcomes across places.

3.3.2. Guideline-defined standards of care

In Kenya, standards of in-patient care for children are defined in the Basic Paediatric Protocols [MoH 2010, MoH 2007, MoH 2006]. These are a set of practice guidelines for health care workers who provide care to hospitalized children. The guidelines aim to steer care towards practices which are linked to good outcomes. They provide guidance on key diagnostic signs of illness to be documented during initial clinical assessment,

criteria for distinguishing illness severity based on history, signs and symptoms and diagnostic test results, and recommendations on treatments appropriate for various illnesses and severity classifications (Figure 3.3-1, Figure 3.3-2 and Figure 3.3-3).

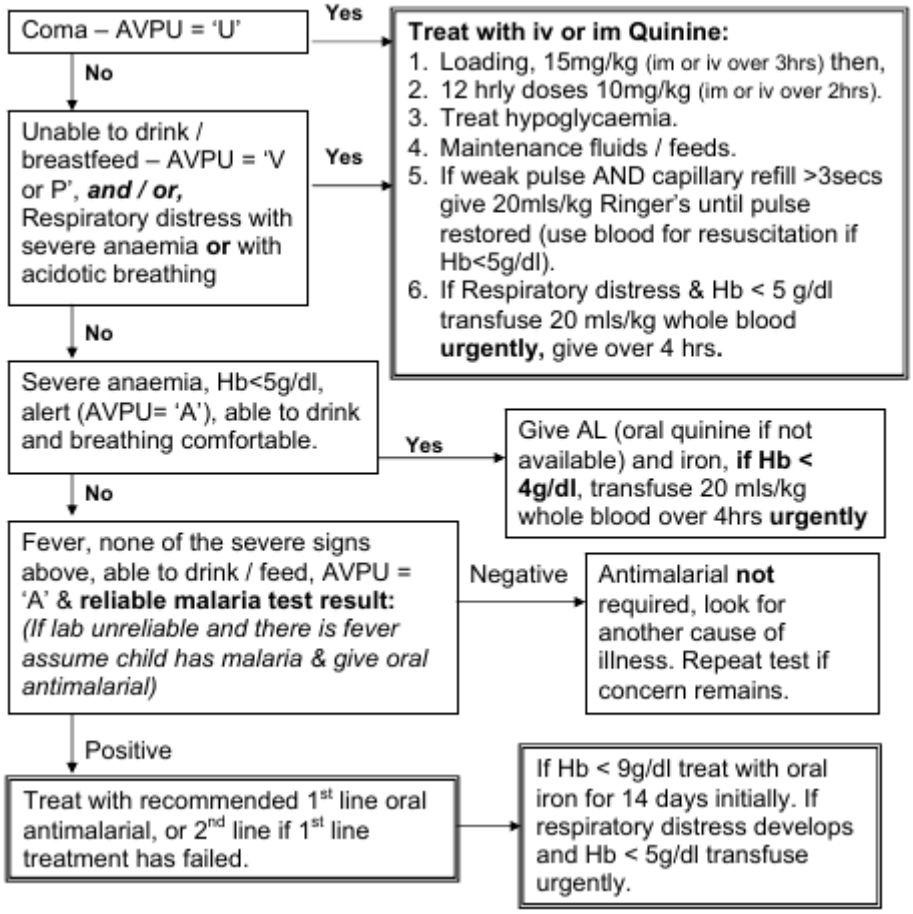


Figure 3.3-1: Guidelines on clinical management of malaria from the Basic Paediatric Protocols

AVPU = alert, responsive to voice, responsive to pain, unconscious; a scale for characterising a patient's level of (un)consciousness
AL = artemether-lumefantrin, a fixed dose combination antimalarial drug that is the first-line of treatment for non-severe malaria at the time of publication of these guidelines

[MoH 2010]

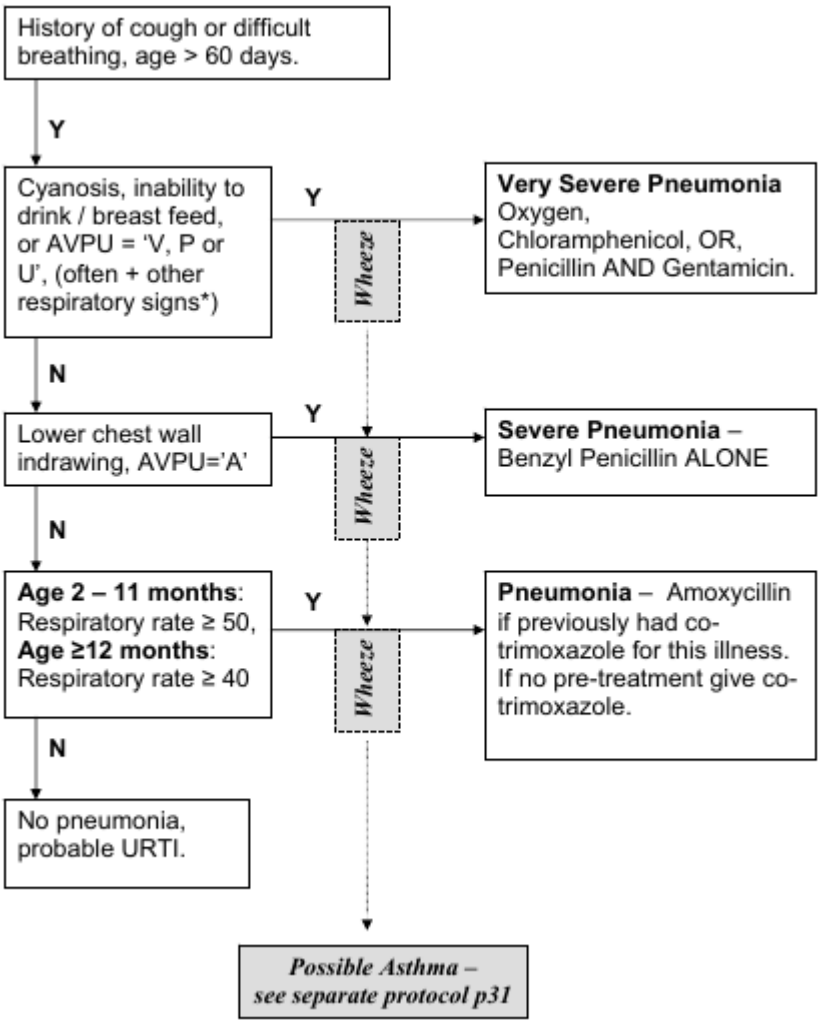


Figure 3.3-2: Guidelines on management of acute respiratory infections and pneumonia from the Basic Paediatric Protocols

AVPU = alert, responsive to voice, responsive to pain, unconscious, a scale for characterising a patient's level of (un)consciousness
URTI = upper respiratory tract infection

[MoH 2010]

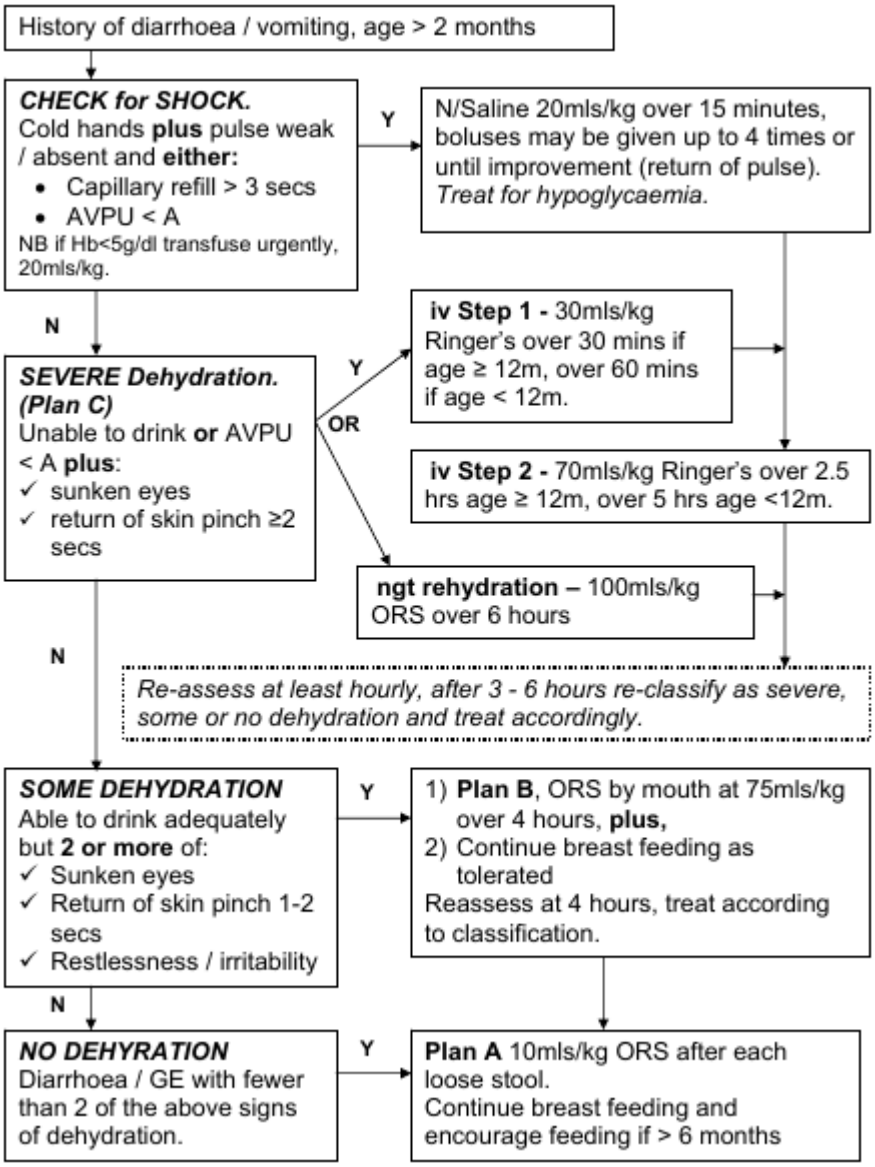


Figure 3.3-3: Guidelines on management of diarrhoea/dehydration from the Basic Paediatric Protocols

AVPU = alert, responsive to voice, responsive to pain, unconscious, a scale for characterising a patient's level of (un)consciousness
GE = gastroenteritis
ORS = oral rehydration salt solution

[MoH 2010]

Directions on the use of recommended drug treatments including dosages, routes, frequencies and durations of treatment are also given in the protocols (Appendix A.9). From these protocols three distinct steps in the process of care can be defined which not only represent different dimensions of process but also distinct competencies. An initial assessment step encompasses the documentation of signs and symptoms key to identifying illnesses and determining their severity. It may be guided by tools (such as a

paper or computer-based structured assessment tool) and may simply involve following rules in checking off signs and symptoms.

This step is followed by the diagnosis phase which sets out algorithms on how the recorded clinical signs and symptoms can be used to identify the presenting illness. There is also guidance on classifying illness severity based on the presence or absence of danger signs. From a process of care perspective it represents a cognitively more complex set of tasks than assessment and is highly dependent on a complete assessment. The third step is the treatment step. Guidance on treatment steers the choice of correct drug(s) for treating the diagnosed illness and severity classification and their correct use. Accurate treatment is also a cognitively complex task as it requires integrating the drug use and dosing information presented in guidelines and calculation of very specific dosages. It is less dependent on completeness of information from other domains since once a choice is made to follow any course of treatment then that treatment ought to be provided as recommended. The guidelines thus provide a starting point for identifying suitable quality 'items' to be used for constructing the quality of care measure.

3.3.3. Domains in the proposed quality of care measure

Many measures found in the literature group items into domains: these are abstract higher-level units representing distinct dimensions of the construct of interest which are themselves of interest in their own right. Indeed domain-level measures are simply distinct elements of an overall summary measure which quantify how its different conceptual parts contribute to the measure. For example, the original 15-item WHO-QOL considered quality of life to span five domains named material and physical well-being, relationships with other people, social, community and civic activities, personal development and fulfilment, and recreation [Burckhardt & Anderson 2003]. The Nottingham Health Profile quantifies stress related to potentially disabling conditions through six domains namely pain, physical mobility, emotional reactions, energy, social isolation and sleep [Klevsgård *et al.* 2002].

In a similar fashion indicator items identified from guideline-recommended phases of process of care can be grouped into distinct domains of generic relevance to the clinical process. In this thesis three domains are proposed, and nomenclature linked to conceptual labels of sets of guideline-recommended clinical processes undertaken at each domain are assigned to them. These are: assessment of signs and symptoms of

disease ('assessment'), diagnosis of disease and classification of severity ('classification'), and treatment. Grouping items into domains in this manner should allow for the use of different items in scores pertinent to each disease (disease-specific measures) but linkable to common domains for scoring purposes (domain-specific measures). Conceptually domains also allow combination of information from different disease episodes within patients (admission or episode-specific measures) and across patients (patient-, clinician-, department- and hospital-level measures). Doing this effectively provides an efficient and practical multilevel approach to measurement with the key capabilities required of a measure of quality of care for routine use: 1) domain-specific measures allow for the identification of 'problem areas' in the process of care; 2) disease-specific measures provide the capability to identify illnesses for which care is poor; 3) patient-specific measures enable the examination of each individual's admission experience and, potentially, linking of care to outcomes with adjustment for important individual-level factors such as co-morbidities; and 4) measures aggregated at clinician, departmental and hospital level enable identification of good/poor areas of care. Figure 3.3-4 provides an outline of the design of the measure proposed in this thesis which has not been reported in literature previously.

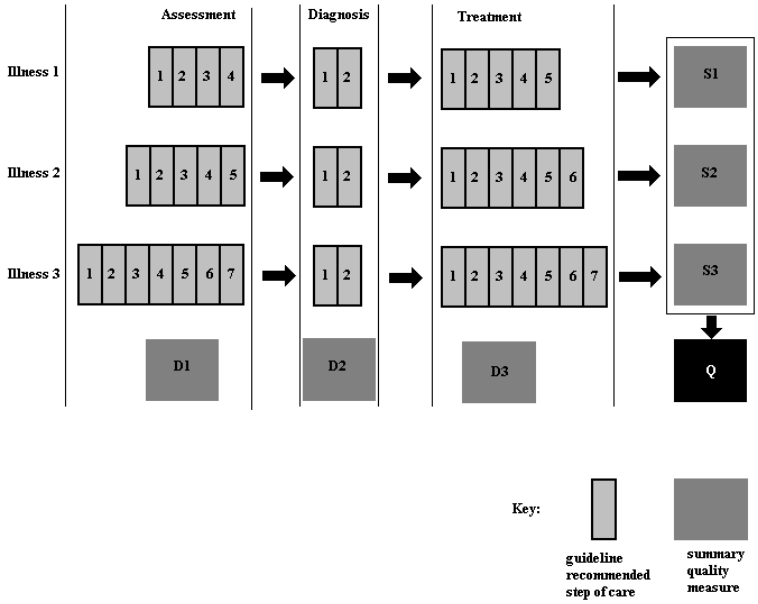


Figure 3.3-4: Outline of the proposed measure showing various levels of summary up to the individual level.

Items contributing to the measure and which may be unique for each illness are labelled 1 through 7 while D1 through D3 represent domain-level aggregate measures for each illness or combination of illnesses. S1 through S3 are summary measures for each illness. Q is the overall measure at the individual level, which when averaged for children attended to by the same clinician or at a department or hospital constitute higher-level aggregate measures.

3.4. Summary

A measure of quality of care is proposed to quantify how well the process of care for admitted children follows the recommended standards of care set out in clinical practice guidelines. The measure is constructed by first identifying relevant indicator items for process of care from the Basic Paediatric Protocols – these are the established standards of paediatric inpatient care for children admitted to Kenyan hospitals suffering from the commonest causes of illness in this age group. The binary scores from these items are then summed into a measure intended to characterise care at the patient level, allowing for adjustment for other important predictors of care at this level as may be required in statistical analysis, and also being aggregated at higher levels for reporting to target audiences for the purposes of quality assurance and improvement. The construction of the proposed measure follows a widely accepted approach to creating a novel metric for a multidimensional construct as illustrated with some examples of measures of other health related constructs. Nevertheless there is a need to demonstrate that the same approach, when applied to a measure of quality of care, possesses the desired functional properties – good reliability and validity – which positively impact on its suitability for the intended applications. Chapter 4 describes the statistical methods for exploring and testing these properties.

Chapter 4 – Statistical Methods

4.1. Introduction

This Chapter presents an overview of the two main statistical methods underlying the analyses to be applied in testing some characteristics of the proposed measure. The first is factor analysis, a multivariate approach to exploring patterns of variation in the items contributing to a measure. It is needed for exploring the proposed abstraction of items into domains. Related to this are polychoric correlation coefficients for measuring correlations between binary items. The second group of methods is generalised linear models. These are required for examining the association between the proposed measure and objective outcomes of care to establish external validity, and also for assessing agreement between different measures.

4.2. Proposed exploration of item groupings and domains

Item selection and aggregation has to this point relied on logical arguments about how items relate to quality of care and aggregate together in domains, because this approach is arguably important when attempting to design a measure that is meaningful to health workers, managers and policy makers. However there are statistical methods to help the score development process further by not only examining the structure of relationships among items, but also between them and a construct. This can be undertaken even in the absence of a hypothesis about these relationships, and allow for testing the assumptions underlying these relationships. The most commonly reported of these techniques is factor analysis: this is a set of procedures which helps identify relationships between items and group them into *factors* – analogous to the domains previously discussed – representing sub-groups of items which define discrete characteristics within a construct [Nunnally & Bernstein 1994].

4.2.1. Overview of factor analysis

The main assumption in factor analysis is that there exist factors, $F_1, F_2, F_3 \ldots F_m$, which are not directly observed but which manifest themselves through observable items $T_1, T_2, T_3 \ldots T_n$ measured with some errors $e_1, e_2, e_3 \ldots e_n$ that are independent of both the factor and each other and identically distributed with a mean of zero and some variance. Given a factor, F_m, the items measuring it are independent of each other and any

relationship between items is only through the factor (Figure 4.2-1). Models that make this assumption are known as orthogonal factor models while those that relax this assumption are called dependent factor models.

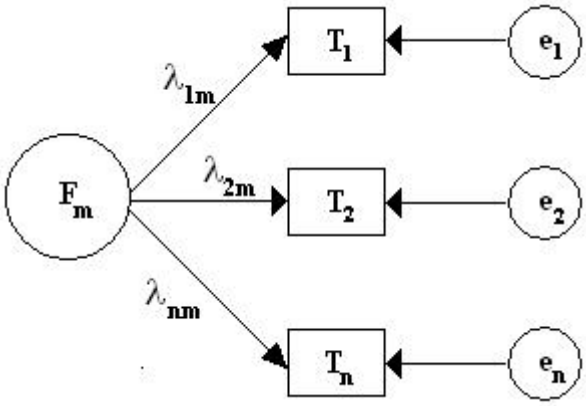


Figure 4.2-1: Path diagram showing the relationship between latent factors, F_m, observed items, T_n and measurement errors, e_n

Factor analysis is about examining correlations between items and factors. The correlation between the factor, F_m and any item, T_n, is measured through the item's *loading* on the factor, λ_{nm}. If the items are normally distributed with zero mean and unit variance then the loadings are equivalent to the regression slope of the corresponding item on the factor or a correlation between the item and the factor, or the proportion of variability in the item explained by the factor. Factor loadings, which range in value from -1 to 1, could be interpreted in a similar fashion as regression coefficients or path coefficients in path analysis, that is, the magnitude and direction of the link between items and factors.

The path diagram in Figure 4.2-1 can be expressed mathematically as:

$$T_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \lambda_{13}F_3 + \ldots \lambda_{1m}F_m + e_1$$

$$T_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \lambda_{23}F_3 + \ldots \lambda_{2m}F_m + e_2$$

$$T_n = \lambda_{n1}F_1 + \lambda_{n2}F_2 + \lambda_{n3}F_3 + \ldots \lambda_{nm}F_m + e_n$$

It can also be expressed in matrix notation as:

$$\boldsymbol{T_{nx1} = \lambda_{nxm}\, F_{mx1} + e_{nx1}}$$

which is equivalent to:

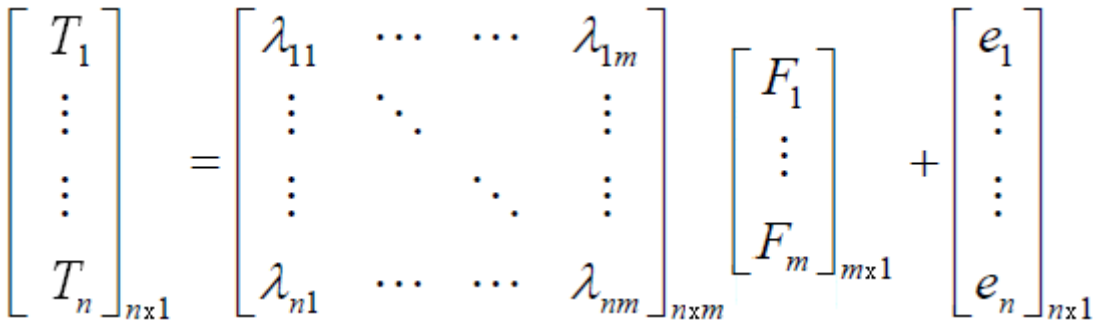


Figure 4.2-2: Matrix notation for factor analysis

The error terms $e_1 - e_n$ are independent of each other so that $E(e_n) = 0$ and $Var(e_n) = \sigma^2_n$.The factors F_m are independent of each other and of the e_n such that $E(F_m) = 0$ and $Var(F_m) = 1$. Each factor is therefore a combination of items, loadings and errors. Every item may load on more than one factor but the maximum number of factors that can be extracted from a set of items is equal to the number of items. In the context of this work this implies that more domains could in theory be identified from the list of quality items and each item can 'belong' to more than one domain. For this reason the final design of the score must not only be consistent with an identifiable factor structure but also be clinically relevant.

The variance of each item can be calculated as a function of all the factors it loads on as:

$$Var(T_n) = \lambda^2_{n1}Var(F_1) + \lambda^2_{n2}\,Var(F_2) + \lambda^2_{n3}\,Var(F_3) + \ldots \lambda^2_{nm}\,Var(F_m) + Var(e_n)$$

$$= \lambda^2_{n1} + \lambda^2_{n2} + \lambda^2_{n3} + \ldots \lambda^2_{nm} + \sigma^2_n$$

The proportion of total variance in an item explained by the factors, termed *communality* of the item, is the sum of loadings of all the factors it loads on ($\lambda^2_{n1} + \lambda^2_{n2} + \lambda^2_{n3} + \ldots \lambda^2_{nm}$). This intuitively suggests that items with high communality load on more factors and are therefore less unique. A related characteristic, the *uniqueness* of an item, is given by 1 minus its communality. This implies that higher communalities correspond to higher factor variances, while higher uniquenesses correspond to higher error variances. The simplest approach to estimating item communalities is the principal components analysis (PCA) method described in section 4.2.3. It accounts for all variance in the correlation matrix, including variances due to measurement error [Nunnally & Bernstein 1994]. Due to this method's dependence on total variance it is important that the items being considered be measured on the same scale. If this is not

the case they may be standardised to have zero means and variances of 1 (giving total variances equal to the number of items), or using correlation coefficients obtained from the correlation matrix of the items. The latter method is usually applied in PCA. An approach which removes measurement error from an item's unique variance in the correlation matrix is the common factor analysis which results in total extracted variances of less than 100%.

4.2.2. Characteristics of item correlations, variances and covariances in factor analysis

Factor analysis makes repeated use of correlations or covariances to examine relationships between items. The covariance between two sequential items T_n and $T_{(n-1)}$ is:

$$Cov(T_n,T_{(n-1)}) = \lambda_{n1}\lambda_{(n-1)1}Var(F_1) + \lambda_{n2}\lambda_{(n-1)2}\ Var(F_2) + \lambda_{n3}\lambda_{(n-1)3}\ Var(F_3)\ +\ldots+\ \lambda_{nm}\lambda_{(n-1)m}$$

$$Var(F_m) = \lambda_{n1}\lambda_{(n-1)1} + \lambda_{n2}\lambda_{(n-1)2} + \lambda_{n3}\lambda_{(n-1)3}\ +\ldots\ +\ \lambda_{nm}\lambda_{(n-1)m}$$

The variances and covariances are arranged into a symmetric matrix implied by the model assumptions: the theoretical variance-covariance matrix. An observed variance-covariance matrix is then obtained from the observed items. Since the factors are unobservable and unitless, it is often assumed for the sake of mathematical convenience that they are standardised; this allows the interchangeable use of correlations and covariances in the estimation procedure. An iterative algorithm is then applied to produce factor estimates and loadings corresponding to a theoretical matrix which is as close to the observed one as possible.

If there are sufficient relationships to enable identification of common underlying factors then this covariance or correlation matrix should be significantly different from an identity matrix, $\boldsymbol{I}$. For this not to be the case would imply no relationship at all between any two items. The Bartlett's test of sphericity is an initial chi-squared test to examine whether a set of items is factorable by checking that the correlation matrix is sufficiently different from an identity matrix [Bartlett 1950]. If N observations on t items have a correlation matrix whose determinant is $|\boldsymbol{D}|$ then the test statistic, which has ½ $t(t-1)$ degrees of freedom, is:

$$\chi^2 = -\left[(N-1) - \left(\frac{2t+5}{6}\right)\right] log_e\ |\boldsymbol{D}|$$

The correlation matrix must also be invertible. If $\boldsymbol{C}$ is a correlation matrix then its inverse, $\boldsymbol{C}^{-1}$, is the matrix such that $[\boldsymbol{C}][\boldsymbol{C}^{-1}] = \boldsymbol{I}$. To be invertible the *determinant* of a matrix must not be zero [Hays 1994]. The determinant is calculated by Laplacian expansion [Doherty *et al.* 1949]. For example, for a 2x2 matrix it is the difference between the products of all elements of the matrix's leading and lagging diagonals. A non-invertible square matrix – a *singular* or *non-positive definite* matrix – has a zero determinant and is not factorable [Fraleigh *et al.* 1994]. Such a situation may arise if items are too highly correlated with each other as may be the case if some items are a combination of others. The dependent items must then be removed from the analysis.

Every factorable n x n correlation matrix is also associated with n values of a scalar, λ, such that:

$$|\boldsymbol{C} - \lambda \boldsymbol{I}| = 0$$

Each λ is an *eigenvalue* of the correlation matrix $\mathbf{C}$ and is associated with an n x 1 column vector, $\mathbf{v}$, such that:

$$\boldsymbol{C}\boldsymbol{v} = \lambda \boldsymbol{v}$$

Similarly each $\mathbf{v}$ is an *eigenvector* of $\boldsymbol{C}$. In factor analysis eigenvectors and eigenvalues correspond to factors and the amount of variance they explain, respectively. Factor loadings are the product of the elements of each eigenvector and the square-root of its eigenvalue.

4.2.3. Factor extraction

Factor analysis begins with an examination of the pattern of correlations between items to identify those that are highly correlated with each other; these are assumed to be the observed manifestations of the same unobserved factors, and the less uncorrelated items are assumed to be influenced by different factors [DeCoster 1998]. The aim of the process is to recreate the correlation matrix of the observed items using a set of factors which is a function of the observed items [Pett *et al.* 2003].

Computationally, factor extraction begins with the creation of an arbitrary row vector – the first trial eigenvector – by adding up the elements of each of the columns of the correlation coefficient. The sum of the squared elements of the trial eigenvector gives the length of the vector. Each of this vector's elements is then divided by the square-root of its length to obtain the first *normalized* trial vector. The sum of squares of the

elements of a normalized vector is always 1. This vector is also an approximation of the eigenvector of the first factor; a better estimate – the second trial eigenvector – is obtained by multiplying it by the original correlation matrix. A second normalized trial vector is obtained from the second trial eigenvector by dividing each of its elements by the square-root of its length. This iterative process is repeated until the sum of squared differences between the final and penultimate normalized trial vectors is less than 0.00001 – at this point convergence has been achieved [Kline 1994]. The final normalized vector is the first eigenvector of the correlation matrix, and the square-root of final trial eigenvector's length is its first eigenvalue. Item loadings on the first factor are the product of the first eigenvector and the square-root of its eigenvalue.

Item loadings on the second factor are obtained following the same iterative process, starting with a residual matrix of the original correlation matrix from which the effect of the first factor has been subtracted. The squared elements of the leading diagonal of a matrix made up of all possible cross-products of factor loadings of the first factor provides an estimate of its effect. The leading diagonal of the residual correlation matrix is no longer made up of 1's but the proportion of remaining variance unexplained by the first extracted factor, and many of the remaining correlations are small or even negative; this presents a problem because to identify any remaining factors there must be sufficient residual correlation between items even after extraction of initial ones. The solution is to identify columns of the matrix for which if the negative correlation coefficients of their elements are inverted then column sums, hence correlation coefficients, are maximised. Later when factor loadings have been estimated the negative signs will be restored to them. The maximised column sums are now the second trial eigenvector, and the process proceeds to convergence as previously described.

The extraction procedure described in the last two paragraphs is called the principle components analysis (PCA). There are other extraction methods which vary slightly from this. For example the principal factor analysis (PFA) method begins by inserting communalities estimated from the squared multiple correlations or highest absolute correlations in a row into the leading diagonal of the correlation matrix to replace the 1's then proceeding like in a PCA; iterative principal factor (IPF) or least squares method is akin to a PFA with re-estimation of communalities after each factor extraction; and maximum likelihood method is an iterative procedure that assumes

normal distribution of factors and uses a likelihood function that maximizes parameters to estimate them.

The maximum number of extractable factors is equal to the number of items. However each successive factor explains only part of the variability remaining after extraction of – and always less overall variability than – the previous ones. Thus if the aim of the analysis is to identify a set of factors responsible for most of the variability in the items (a form of 'data reduction') then factor extraction is stopped when factors explaining a pre-determined amount of variability have been extracted, a pre-planned number of factors have been obtained, or factors with eignevalues greater than 1 have been identified. Ultimately it is the interpretation and intended application of the extracted factors that should be used as a guide on the number to be extracted [Nunnally & Bernstein 1994].

Finally to improve interpretability of the extracted factors, item loadings are jointly 'rotated' so that loadings are highly positive or negative on some factors while very low on others. If the factors are thought to be completely independent then orthogonal rotation techniques are applied. These achieve rotation through maximizing squared loading variances across items by adding over factors (varimax rotation) or across factors by adding over items (quartimax rotation). Alternatively if there is thought to be some relationship between factors then oblique rotations which minimize squared loading covariance between factors (oblimin) or use simplified orthogonal rotation (promax) are applied.

4.2.4. Exploratory and confirmatory factor analysis, and structural equation modelling

Factor analysis could be undertaken to identify and quantify factors relating to observed items, or to test theorized factor models against observed data. These distinct applications, summarised in Table 4.2-1, are referred to as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) respectively. Generally EFA is applied in situations where the underlying construct or domain structure of data is unknown while CFA is used when there is a theory about domain structure to be tested. EFA may be used to generate a theory on domain structure followed by CFA to test the theory. For the latter application it is advisable fit the confirmatory model on different dataset (e.g. a validation dataset) or different halves of the same dataset if there is still a sufficient

number of observations in the split dataset, remembering that factor analysis requires relatively large number of observations compared to many other multivariate methods. EFA and CFA may also be applied in repeated cycles to generate, modify and refine theories about the underlying factor structure in observed data.

Table 4.2-1: Contrasting exploratory and confirmatory factor analysis

Exploratory factor analysis (EFA)	**Confirmatory factor analysis (CFA)**
Steps	
1. Measure construct of interest using suitable items 2. Create the correlation matrix of the items 3. Select a number of factors to retain 4. Extract the factors 5. Rotate the factors 6. Interpret the factors 7. Create factor scores for further analysis if required	1. Define factor model 2. Measure construct of interest using suitable items 3. Create the correlation matrix of the items 4. Fit defined factor model to the data 5. Test goodness of fit of model 6. Compare with alternatively formulated factor models
Uses	
• Identify nature of factors influencing measures • Determine how items group together into domains • Show the dimensions of a construct that a scale is capable of measuring • Identify the most important domains of constructs. • Create factor scores to be used for further analyses	• Test the validity of any given factor model • Compare the fits of alternate factor models to the same data • Compare factor loadings and test their significance • Examine correlation between factors • Examine convergent and discriminant validity of different measures

[DeCoster, 1998]

Structural equation modelling (SEM) is a technique that combines CFA with multiple regression [Ullman 2001]. An SEM has two parts: a *measurement model* – the CFA– and a *structural model* (also referred to as a *path model*) which is essentially a series of linear relationships estimated simultaneously using multiple regression. The regression component in SEM is used to explore associations between observed non-measurement *exogenous* (independent) variables and latent variables estimated through CFA, and also with *endogenous* (dependent) variables in causal hypotheses. Thus SEM extends the hypothesis testing approach in CFA to a multivariate analysis of a theory that poses causal relationships among several variables [Lei & Wu 2007].

4.2.5. Goodness-of-fit assessment for factor analysis models

Model fit assessment begins by working back from the error variances and predicted correlations to obtain the item correlation matrix implied by the extracted factors; this is then compared to the original correlation matrix [Prudon 2013, Hooper *et al.* 2008]. The difference between these two matrices is the residual matrix, which is a function of prediction error of the model and random error arising from the estimation sample [Cudeck & Henly 1991]. If the correlation matrix implied by the factor analysis model is:

$$\mathbf{I} = \mathbf{L}'\,\mathbf{L} + \mathbf{\Psi}$$

where **L** is a matrix of the factor loadings, the diagonals of the matrix **Ψ** bear the variances, and:

$$\hat{\mathbf{I}} = \left(\frac{n-1}{n}\right)\mathbf{S}$$

where **S** as the sample correlation matrix, then a test statistic for goodness of fit – the Bartlett-Corrected Likelihood Ratio Test Statistic – is given by:

$$\chi^2 = \left(n - 1 - \frac{2t + 4m - 5}{6}\right)\log\frac{\left|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}\right|}{\left|\hat{\mathbf{I}}\right|}$$

where t is the number of items, m the number of factors and n the number of observations. The degrees-of-freedom for the χ^2 statistic is:

$$df = \frac{(t+m)^2 - t - m}{2}$$

Under the null hypothesis there is no difference between the matrices and χ^2 should be small; a large value (corresponding to a small p-value) therefore indicates poor fit. However, because χ^2 is a function of n, large samples (approximately $n > 200$) will almost certainly correspond to small χ^2 values, regardless of model fit. For this reason other fit indices are often preferred for larger studies. They either assess how far the model is from a perfect fit (absolute fit indices), how far the model fit is from the best

possible fit for the data (incremental or relative fit indices), or how well the model fits with adjustment for its complexity and sample size (parsimony fit indices).

They include:

- root mean square error of approximation (RMSEA), an absolute fit index which is a modification of the χ^2 which favours large samples:

$$RMSEA = \sqrt{\frac{(\frac{\chi^2}{df} - 1)}{n - 1}}$$

  Smaller RMSEAs indicate better fitting models. Conventionally models with RMSEA equal to or less than 0.06 are considered to be well fitting [Hu & Bentler 1999].

- standardised root mean square residual (SRMSR) which is also an absolute fit index obtained from the sum of the squared residuals comparing the sample correlation matrix to the implied correlation matrix as follows:

$$SRMSR = \sqrt{\frac{1}{r}\sum (S_{ij} - I_{ij})^2}$$

  where r is the number of residuals given by $t(t+1)/2$ [Jöreskog & Sörbom 1988]. Similar to the RMSEA smaller values of SRMSR indicate better fits, but unlike the former this index is not based on the χ^2. The conventional cut-off for a model to be considered well-fitting is SRMSR equal to or less than 0.08 [Hu & Bentler 1999].

- Tucker-Lewis index (TLI) [Tucker & Lewis 1973] which is a comparative fit index. It is the ratio of the χ^2 of the implied correlation matrix to that of a null model that assumes that all items are uncorrelated:

$$TLI = \frac{\frac{\chi^2}{df} null - \frac{\chi^2}{df} implied}{\left(\frac{\chi^2}{df} - 1\right) null}$$

The closer the TLI is to 1 the better the model fit. The conventional cut-off for a well-fitting model is a TLI equal to or greater than 0.90. Some studies have suggested a higher threshold of 0.95 to further reduce the risk of accepting mis-specified models [Bentler 1990, Hu & Bentler 1999]. The TLI is also sensitive to sample size: it sometimes indicates poor fit even when other indices point to good fit when sample sizes are small [Bentler 1990, Kline 2005, Tabachnick & Fidell 2007]

- Comparative fit index (CFI) [Bentler 1990] which is similar to the TLI in its derivation and interpretation:

$$CFI = 1 - \frac{(\chi^2 - df)implied}{(\chi^2 - df)null}$$

- Akaike information criterion (AIC) [Akaike 1974], a parsimony fit index obtained by $2k - 2\log_e(L)$ where k and L are the number of parameters and the maximum likelihood estimate of the model, and a variety of eponymous goodness-of-fit indices which adjust for sample size and model complexity.

The various fit indices give information on different aspects of the fit of the factor analysis model; it is therefore good practice to report at least one of each type [Hooper *et al.* 2008]. It is also worth noting that although it is common practice to use conventional cut-offs of fit indices to decide whether to reject or adopt a model, fit indices have different sensitivities to peculiarities across models, such as sample sizes and degrees of freedom [Fan & Sivo 2007]. For this reason the suggested cut-offs should not be interpreted too strictly. Table 4.2-2 provides some guidance on the interpretation of three commonly reported fit indices.

Table 4.2-2: Suggested criteria for assessing the fit of confirmatory factor analysis models

Fit index	**Good fit**	**Acceptable fit**	**Marginal fit**	**Poor fit**
CFI	> 0.95	0.90 – 0.95	0.85 – 0.89	< 0.85
TLI	> 0.95	0.90 – 0.95	0.85 – 0.89	< 0.85
RMSEA	< 0.05	0.05 – 0.08	0.09 – 0.10	> 0.10

4.3. Application of factor analysis to the development of a measure of process of care

Factor analysis can enrich the score development process in several ways. First, as noted at the beginning of this section, the initial phases of the process have relied on a plausibility approach to determine the properties of a measure that make it sensible and meaningful to its target audience and users. A confirmation of the proposed domains and item relationships may demonstrate the internal consistency of the design of this measure. This can be achieved by performing factor analysis on a test dataset to identify latent factors most responsible for observed variability.

Secondly, the factor extraction process allows for identification of items which do not add value to the score, either because they are redundant due to very high correlation with other items, or because they do not exhibit sufficient variability across individuals or places to be useful in measuring variability in the process of care. This arguably increases the sensitivity of the measure to real and clinically significant changes in the process of care which the measure seeks to identify in the first place.

Lastly, a combined application of exploratory and confirmatory factor analysis can test and refine the proposed score design to yield a more robust measure better supported by data or one which better fits the proposed application. Factor analysis also allows for the testing of various assumptions about how the domains relate to each other, for example, whether data supports the idea that they could be completely independent of each other (as proposed in an orthogonal factor model) or whether there are indeed some relationships between them.

4.4. Comparison of alternate candidate measures

Refining the design of the proposed measure is likely to yield alternative formulations of the measure. Comparisons between the different forms of the measure allows for alternate-form reliability testing, and investigation of potential loss of information that can be taken into account to inform any necessary design adjustments. The choice of method for comparison depends on characteristics of the measures being compared, such as the type of scale of measurement (whether categorical or continuous), multiplicity of comparison (i.e. simultaneous comparison of two or more measures) and whether additional sources of variation are to be accounted for in the comparison.

The Pearson correlation coefficient, r, this is the most commonly reported measure of the relationship between measures or variables in data. It quantifies the strength of a linear relationship between a pair of continuous measures X_i and Y_i [Pearson 1901]. It is given by:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

It is constrained to $-1 \leq r \leq 1$. Values close to these extremes imply strong linear relationship. Positive and negative values imply positive and negative linear relationships (correlations) respectively. Zero values mean there is no correlation between the two measures; however, this is not the same as there being no relationship between them – two measures with a quadratic relationship may have a correlation of zero because the relationship between them can be perfectly non-linear. A suggested interpretation of the absolute values of r is, in increasing quintiles, 'very weak', 'weak', 'moderate', 'strong' and 'very strong' correlations respectively [Evans 1996, Cohen 1992, Taylor 1990]. A Pearson correlation is only valid for measurements on an interval or ratio scale and which are both normally distributed; it is sensitive to deviations from normality including skewness and outliers. A non-parametric equivalent for non-normally distributed interval or ratio scale measures is the Spearman rank correlation coefficient, ρ, which is essentially a Pearson correlation coefficient calculated not on the raw continuous measures X_i and Y_i but on their ranks.

For measures on an ordinal scale, such as binary and ordered categorical measures, polychoric correlation coefficients are the appropriate way of examining linear relationships [Pearson 1901]. These can be used even if the measures being examined have different numbers of categories. Polychoric correlations have one characteristic in common with factor analysis: they are a special case of latent trait modelling. They are based on the assumption that underneath the ordinal scale of a measure is a continuous normally distributed latent trait, and that the levels of the scale represent ordered categories of the trait which depend on some underlying thresholds [Uebersax 2006]. In calculating polychoric correlations this trait need not necessarily be normally distributed; any unimodal and approximately symmetrically distributed trait is sufficient to estimate valid polychoric correlation coefficients.

If X_1 and X_2 are measures of the same outcome on scales 1 and 2, and Y_1 and Y_2 are the underlying latent trait levels distinguished into the observed ordered levels of X_1 and X_2 by some thresholds on Y_1 and Y_2; assuming the true unknown level of the trait is L, then:

$$Y_1 = \beta L + u_1 + e_1 \text{ and } Y_2 = \beta L + u_2 + e_2$$

where u is the observer-specific impression of the trait measured with error e_1, $e_2 \sim N(0, \sigma_e)$ independent and identically distributed, and $var(e_1) = var(e_2)$.

This can be simplified to:

$$Y_1 = \beta_1 L + e_1 \text{ and } Y_2 = \beta_2 L + e_2$$

If L is standardised to be $N(0, 1)$ then $var(Y_1) = var(Y_2) = 1$, and also $\beta_1 = \beta_2 = \beta$, and the polychoric (*tetrachoric* if X_1 and X_2 are binary) correlation coefficient r is given by β^2.

Factor analysis of categorical measures uses polychoric correlation matrices of items to identify patterns of relationships between them when performing the factor extraction procedure described in section 4.2.3.

When exploring sources of variation beyond the individual level, then regression modelling is the preferred approach of comparing alternate measures. A generalized linear model of a measure $\boldsymbol{Y}$ on a reference measure $\boldsymbol{X}$ can be defined as:

$$E(\boldsymbol{Y}) = \mu = f(\boldsymbol{X}\beta)$$

where $E(\boldsymbol{Y})$ is the expectation of Y and β is a maximum likelihood estimate of an unknown parameter which when multiplied by an X and transformed through f – the link function – gives Y. The link function is monotonic (order-preserving) and its choice is influenced by the distribution of Y (Table 4.4-1). More formally for multilevel data:

$$E(\boldsymbol{Y}) = f(\boldsymbol{X}\beta + \boldsymbol{Z}u + e)$$

where $\boldsymbol{Z}$ is a sub-matrix of $\boldsymbol{X}$, $u \sim N(0, \sigma_u)$, $e \sim N(0, \sigma_e)$ and u are completely independent of e. Once β has been estimated using a calibration sample or a random half of the data, it is used to predict $\boldsymbol{Y}$ given the observed $\boldsymbol{X}$'s. The proportion of agreement between predicted and observed $\boldsymbol{Y}$'s provides an estimate of agreement of the two measures. Through repeated sampling of the data by bootstrapping or Markov

Chain Monte Carlo (MCMC) technique, estimates of uncertainty about the calculated rates of agreement can be produced.

Table 4.4-1: Link functions for some distributions of Y

Type of distribution	Name of link function	Expression of link function
Normal	Identity	$X\beta = \mu$
Binomial, Categorical	Logit	$X\beta = \ln\frac{\mu}{1-\mu}$
Poisson	Log	$X\beta = \ln \mu$
Exponential	Inverse	$X\beta = -\mu^{-1}$

The amount variability in predicted ***Y*** explained by ***X*** at each level of clustering is the intraclass correlation (ICC). It is interpreted at the proportional agreement between measurements. It is particularly useful when none of the measures is a gold-standard and the interest is simply in exploring consistency between them. It is calculated as:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

The assumption of independence between σ_u and σ_e is important because it implies that ***X*** and ***Y*** do not introduce additional variability in the underlying construct during measurement. If this assumption does not hold then this extra source of variability must be added to the denominator.

4.5. Association with mortality

4.5.1. Specification of the model

Averting mortality and other adverse outcomes is often the objective of providing care. For this reason mortality rates are often a key measure of the quality of care. However as established in section 1.1.1 a variety of factors often confound the relationship between the quality of care given to an individual and their risk or odds of death. The proposed measure is designed to allow for adjusting for some of these factors and, in so doing, fostering better estimation of the relationship between care given and outcomes of care. The measure focuses on the admission care that may be critical to clinical progress in the first 48 hours of hospital stay for children, a period during which when many deaths occur [Couto *et al.* 2013, Adeboye *et al.* 2010, Campbell *et al.* 2004, Berkley *et al.* 2003, Sodemann *et al.* 1997, Commey *et al.* 1994]. This large number of

events increases the power of analyses which explore the association of the score with mortality adjusted for: individual level factors such as illness severity and number of concurrent diagnoses which potentially modify the risk of mortality; hospital level effects such as resource availability and nursing care, and other cluster-level predictors of mortality that may not be directly observed or quantified; and other known factors which could potentially affect mortality risk such as temporal and group effects.

For the purposes of this work the mortality risk of a child *i* in hospital *j* is defined as the probability of death being the outcome of this episode of hospitalisation: $\Pr(Y_{ij} = 1)$. This probability is predicted in a random effects logistic regression model:

$$\text{logit}[\Pr(Y_{ij} = 1)] = \beta_0 + \Sigma_1^k [\beta_k x_{kij}] + \mu_j$$

where β_0 is the baseline risk of death and $\Sigma_1^k [\beta_k x_{kij}]$ is the sum of effects of k predictors of mortality related to this episode. μ_j represents the hospital-specific random variations which are normally distributed with a zero mean and variance of σ^2_u for children with the same values of the mortality-predicting covariates. It is minimized by covariate sets which are highly predictive of mortality. For any variable, the proportion by which μ_j declines upon adjustment for quality of care provides an estimate of that variable's contribution to predicting mortality. This approach can be used to test the validity of the measure by estimating the magnitude and strength of evidence for better care in averting mortality.

4.5.2. Assessment of model goodness-of-fit

The model has a fixed part and a random part, and for this reason its overall fit is assessed by considering these two parts separately. The fit of the fixed part is assessed in a similar fashion as a linear model fitted using the ordinary least squares (OLS) method: by checking that residuals – the difference between model-predicted and observed values of the outcome – are normally distributed with a mean of zero [Weisberg 2005, Haessel 1978]. Under this assumption outcome is neither underestimated nor overestimated all through its range of possible values.

The logistic regression model yields predicted probabilities of the outcome taking on one of two possible values. For example the predicted probability that $Y_{ij} = 1$ is given by:

$$\hat{\pi} = \frac{e^{\beta_0 + \Sigma_1^k [\beta_k x_{kij}]}}{1 + e^{\beta_0 + \Sigma_1^k [\beta_k x_{kij}]}}$$

Predicted probabilities are only approximate, and the corresponding residuals can never be normally distributed unlike those obtained from a linear model. Nevertheless there should be no association between the predicted probabilities and the levels of the outcome, since such association would imply that the model overestimates or underestimates the outcome for some values of the outcome. A χ^2 statistic can be obtained from the observed ('*o*', represented by y_i) and predicted/expected ('*e*', the product of the sample size, *n*, and the predicted probability π) values of the outcome:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o-e)^2}{e}$$

$$= \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

If there are continuous predictors in the model, then the predicted frequency of some values will be too small for the predictions to follow a χ^2 distribution. A way around this problem is to group the predicted values into *g* percentiles to obtain sufficient numbers of observations in the resulting contingency table [Hosmer & Lemeshow 2000]. An *H* statistic which follows a χ^2 distribution is then obtained by:

$$H = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{N_g \pi_g (1-\pi_g)}$$

where O_g is the number of predicted values in the g^{th} percentiles of risk, E_g is the expected number of predicted values in the g^{th} percentiles of risk, N_g is the total number of observations and π_g is the predicted probability of the outcome in g^{th} percentiles of risk. The main criticisms against this test are its sensitivity to minor misspecifications especially in models based on large samples, and the great influence of *g* – whose selection is often arbitrary – on the test statistic [Allison 2014].

The fit of the random part of the validation model is assessed by exploring how well the assumption of normality of random effects is satisfied. This is achieved by examining a

histogram of cluster level random effects, or a plot of cluster level random effects versus their ranks [Rabe-Hesketh & Skrondal 2008].

4.5.3. Pooling evidence of association from different studies

If y_i is the estimate of the odds of death per unit increase in the score with var_i as its variance and θ is the overall pooled estimate across k studies, then the chi-squared statistic, Q, is calculated as:

$$Q = \sum_i w_i(y_i - \theta)^2$$

where:

$$w_i = \frac{1}{var_i}$$

with *k-1* degrees of freedom. The inverse of the variance of each estimate also represents its weight w_i, in the pooled estimate which is obtained by:

$$\theta = \frac{\sum w_i y_i}{\sum w_i}$$

Larger values of the test statistic correspond to more evidence for the presence heterogeneity across studies. The proportion of variability across studies attributable to heterogeneity, I^2, is calculated as:

$$I^2 = \frac{Q - k + 1}{Q}$$

The estimate of between-study variance τ^2, obtained by:

$$\tau^2 = \frac{Q - k + 1}{\sum w_i - \frac{\sum {w_i}^2}{\sum w_i}}$$

Using this estimate new inverse variance weights are calculated for each study as:

$$w_i^* = \frac{1}{var_i + \tau^2}$$

and an updated overall pooled estimate of the effect of the score on mortality estimated as:

$$\theta^* = \frac{\sum w_i{}^* y_i}{\sum w_i{}^*}$$

with standard error given by:

$$SE(\theta)^* = \frac{1}{\sqrt{\sum w_i{}^*}}$$

The updated overall pooled estimate and its standard error is then used for hypothesis testing.

4.6. Summary

The univariate and multivariate techniques described in this chapter are integral components of the statistical toolkit to be used for exploring the characteristics of the proposed measure of process of care. Factor analysis can be applied in the assessment of whether items selected for the measure aggregate in groupings which are consistent with what is hypothesised. This is useful for establishing construct validity of the measure, and also in identifying redundant items either for modification or removal from the score. Regression methods are useful for examining associations between the proposed measure and other variables, most notably mortality, as another layer of validation. Measures of agreement help in determining the extent to which different measures of the same construct lead to similar conclusions.

Chapter 5 – Construction of the Proposed Measure and Preliminary Validation

5.1. Introduction

In this first results Chapter the proposed measure of quality of care is constructed from process indicator items derived from clinical practice guidelines. Data from the Kenyan district hospitals study, described in detail in section 2.2, are used to illustrate the steps of score construction. The data contain detailed information on patient-level process of care across place (hospital), care provider (clinician) and time (survey), and between two groups distinguished by a quality improvement intervention. Using these data a score is derived from items identified for measurement. Characteristics of the score, including its distribution and differences across diseases, groups and time, are examined to identify and implement necessary improvements to the proposed design.

5.2. Score construction

5.2.1. Domains of the measure

The proposed domains are derived from clinical practice guidelines for management of childhood illnesses in Kenya which recommend that the process of care for a sick child should be a logical and step-wise process informed by a large body of evidence of links to good outcomes. In the initial phase of the process a clinician interviews the child's caretaker to obtain a history of illness, performs a physical examination of the child and may also order any necessary diagnostic tests to confirm the diagnosis of suspected illnesses. Although it is recommended that a uniform set of initial assessment tasks is conducted for all sick children, the core tasks necessary to identify common and important causes of severe illness are fewer and vary from illness to illness.

In the next phase the clinician uses their knowledge and experience to identify the most likely cause of illness and make a classification of its severity based on the history of symptoms, observed clinical signs and results of diagnostic tests. Severity classifications vary across illnesses, with the more severe classifications being associated with higher risks of worse outcomes. In the third phase the clinician charts the most appropriate course of treatment which also varies across illness and severity classification. Variations in clinicians' performance at each of these phases contribute to

overall variation in the process of care. When considered singly these core assessment tasks, diagnostic classifications and recommended treatments make for useful illness-specific indicator items for the process of care [Rowe 2013, Irimu *et al.* 2012, Opondo *et al.* 2011, Osterholt *et al.* 2009].

5.2.2. The basic score

Guidelines on how to provide care are outlined in the Basic Paediatric Protocols and the ETAT+ training course for health workers providing initial care to children admitted to hospital. The development of both components of the intervention has been described in section 1.4.1, the guidelines on assessment of disease signs and symptoms presented in Figure 3.3-1, Figure 3.3-2, Figure 3.3-3, and treatment guidelines shown in Appendix A.9. A list of items corresponding to specific recommendations on how to deliver care was identified from these guidelines. They are listed in Table 5.2-1 and Table 5.2-2. The items were binary, to be scored 1 if they were undertaken as recommended and 0 otherwise. Item scores were aggregated within domains, which were conceptual higher-level units proposed to represent distinct dimensions of the process of care, namely 'assessment', 'diagnosis' and 'treatment' introduced in section 3.3.3.

Table 5.2-1: Items in the assessment and diagnosis domains of the basic process-of-care score

Domain	**Disease**		
	Malaria	**Pneumonia**	**Diarrhoea/dehydration**
Assessment *Each item scored 1 if documented (present, absent, quality or quantity) and 0 otherwise*	1. Fever 2. Convulsions 3. Acidotic breathing 4. Pallor 5. (In)ability to drink or breastfeed 6. Level of consciousness (AVPU) 7. Indrawing 8. Blood test for malaria	1. Cough 2. Difficult breathing 3. Central cyanosis 4. (In)ability to drink or breastfeed 5. Level of consciousness (AVPU) 6. Grunting 7. Indrawing 8. Respiratory rate	1. Diarrhoea 2. Vomiting 3. Capillary refill 4. (In)ability to drink or breastfeed 5. Level of consciousness (AVPU) 6. Sunken eyes 7. Return of skin pinch 8. Character of pulse
Diagnosis *Item score is 1 if a relevant severity classification is indicated, 0 otherwise*	1. Classification: severe or non-severe	1. Classification: very severe, severe or non-severe	1. Classification: shock, severe, some or none

Table 5.2-2: Items in the treatment domain of the basic process-of-care score

Domain	**Disease**		
	Malaria	**Pneumonia**	**Diarrhoea/dehydration**
Treatment *Score items depend on severity classification* *'Drug' score is 1 if correct (singly or in recommended combinations where applicable) according to guidelines for indicated severity classification* *'Route', 'dose', 'duration' and 'frequency' each score 1 if correct (singly and in combination where applicable) for choice of drug(s) according to guideline recommendations for their use, 0 otherwise*	*Severe malaria*: 1. Drug: quinine (loading and maintenance) 2. Route: IV or IM 3. Dose:20mg/kg loading, 10mg/kg maintenance +/- 20% 4. Frequency: twice daily 5. Duration: Stat for loading dose and any duration for maintenance dose *Non-severe malaria*: 1. Drug: artemether-lumefantrine or quinine 2. Route: oral 3. Dose: 5-14.9kg – 1 tab; 15-24.5kg – 2 tabs; 25-34.9kg – 3 tabs; 35kg+ - 4 tabs 4. Frequency: twice daily for AL and thrice daily for quinine 5. Duration: any duration specified	*Very severe pneumonia*: 1. Drug: penicillin and gentamicin and oxygen 2. Route: IV or IM 3. Dose: Penicillin 50,000iu/kg, gentamicin 7.5mg/kg (both +/- 20%) 4. Frequency: Penicillin x4, Gentamicin x1, oxygen any specified 5. Duration: any specified *Severe pneumonia*: 1. Drug: Penicillin only (no gentamicin) 2. Route: IV or IM 3. Dose: 50,000iu/kg +/- 20% 4. Frequency: x4 5. Duration: any specified *Non-severe pneumonia*: 1. Drug: Amoxicillin or cotrimoxazole 2. Route: oral 3. Dose: Amoxicillin 25mg/kg, cotrimoxazole 24mg/kg +/-20% 4. Frequency: Amoxicillin x3, cotrimoxazole x2 5. Duration: any specified	*Shock*: 1. Drug: normal saline or Ringer's lactate/Hartmann's solution 2. Dose: volume/timex4 within +/-20% of 20ml/kg 3. Frequency: at least 1 in an hour *Severe dehydration*: 1. Drug: Ringer's or ORS 2. Dose: total vol/time within +/- 20% of 30ml/kg + 70mg/kg in 3 hours for >1yr or in 6 hours for < 1yr of Ringer's or total vol/time within +/- 20% of 100ml/kg in 6 hours. 3. Frequency: step 1/2 used *Some dehydration*: 1. Drug: ORS 2. Dose: vol/timex4 within +/-20% of 75ml/kg 3. Frequency: at least 1 in an 24 hours *No dehydration*: 1. Drug: ORS 2. Dose: 10ml/kg +/-20% 3. Frequency: any specified

For the assessment domain, a list of signs and symptoms necessary to identify the disease and classify its severity were derived from the guidelines. The assessment score was the number of signs and symptoms documented by the admitting clinician. The diagnosis score was a binary indicator of whether the clinician made a valid classification of the severity of illness, that is, one of the recognised severity classifications in the guidelines. This is because it was not uncommon to find in practice a variety of other classifications which were neither meaningful within the context of the guidelines nor useful for determining the appropriate course of treatment.

An additional step would have been to consider the correctness of the diagnosis and its severity classification; however this was abandoned because it would have potentially led to a 'double-penalty' against less complete assessments whose correctness would have been difficult to ascertain. Furthermore, the judgement of illness severity given by the clinician who examined a sick child was considered to be more reliable and credible to other practitioners than the retrospective judgement of a third-party relying on documented signs and symptoms. This indicator of correctness of severity classification would have at best been subjective; it was therefore excluded from the set of score items.

For the treatment domain, the guidelines explicitly set out recommendations for choice of drugs in each disease and severity classification, recommendations on dosages, route and frequency, and specifications for durations of treatment. Indicators for each of these recommendations were thus defined. Drug use (dose, route, frequency and duration) was scored conditional on drug choice since, just as it was difficult to objectively judge the correctness of a diagnosis, it was hard to definitively determine the appropriateness of treatment selection from the case record alone. It was nevertheless possible to determine whether the treatment was used correctly. Deviations of up to 20% of recommended dosages, which are within therapeutically safe dose ranges for all the drugs used, were considered to be correct to allow for small but clinically insignificant differences between prescribed and guideline recommended drug doses.

Out of the 19 signs and symptoms necessary for identifying and classifying the severity of illness, only two signs – ability to drink or breastfeed and level of consciousness – were common across all three diseases considered. Five treatment indicators were defined to score the treatment of malaria and pneumonia but only three of these – drug choice, dose and frequency – were applicable to diarrhoea/dehydration since the other two (drug route and duration) were implicitly implied by the drug choice and frequency of dosing.

The results were a 9-point (range 0 to 8) assessment score for all three diseases, a 2-point (binary) diagnosis score and a 6-point (range 0 to 5) treatment score for malaria and pneumonia and a 4-point (range 0 to 3) treatment score for diarrhoea/dehydration. The overall process-of-care score – referred to as the basic score because it was an

arithmetic sum of the binary item scores – was a 15-point score (range 0 to 14) for malaria and pneumonia and a 13-point score (range 0 to 12) for diarrhoea/dehydration.

5.2.3. Characteristics of the basic score

Differences in the basic score between the intervention and control group were examined to establish the sensitivity of the score to changes in quality of care that have previously been documented. There was a wide variation in the proportion items achieved across the diseases and time points as shown in Table 5.2-3. At baseline this ranged from 0% for characterisation of pulse to 98.2% for documentation of the presence of diarrhoea in children diagnosed with diarrhoea or dehydration.

In malaria, documentation of the presence of acidotic breathing was poorest at baseline (0.1%), and fever was the most documented sign (86.7%). Level of consciousness was the least documented sign in children with pneumonia (0.8%), and presence of a cough which is a key diagnostic sign of pneumonia was most documented with 88.1% of records having it recorded. At the main study end-point all indicators showed improvement, ranging from 0.4% increase in the frequency of administration of correct drugs for treatment of diarrhoea/dehydration to 79.6% improvement in documentation of the presence of sunken eyes.

Distributions of the basic score across time and groups (Figure 5.2-1, Figure 5.2-2 and Figure 5.2-3) were consistent with what was expected in the presence of improvements in process of care and the heterogeneity across hospitals revealed in the tabulated results: the baseline scores were normally distributed; they were similar across groups as expected; there was also a notable shift towards higher scores in the endpoint score distributions, as would have been expected if the intervention resulted in clinicians undertaking more of the guideline-recommended processes of care.

These confirmed findings which had been demonstrated previously using methods which examined a smaller number of item-specific process-of-care indicators than were included in this score [Ayieko *et al.* 2011].

Table 5.2-3: Percentage of children for whom quality items from the basic score were achieved across all hospitals in the baseline and endpoint surveys

Item in process-of-care score for disease	Percentage of children in which item was achieved					
	Malaria		Pneumonia		Diarrhoea/ Dehydration	
	Baseline n=1,763	End n=1,789	Baseline n=902	End n=1,194	Baseline n=383	End n=732
Assessment						
Fever indicated as present or absent	86.73	95.42				
Convulsions indicated as present or absent	29.33	78.54				
Acidotic breathing indicated as present or absent	0.11	69.70				
Degree of pallor indicated as '0', '+' or '+++'	82.64	90.67				
Laboratory test for malaria requested	79.92	84.24				
Indrawing indicated as present or absent	7.77	72.61	14.19	86.85		
Cough indicated as present or absent			88.14	97.15		
Difficult breathing indicated as present or absent			52.44	89.03		
Central cyanosis indicated as present or absent			24.28	84.51		
Grunting indicated as present or absent			3.77	81.07		
Respiratory rate recorded			16.52	60.39		
Diarrhoea indicated as present or absent					98.17	98.91
Vomiting indicated as present or absent					22.45	87.16
Capillary refill indicated as 'x', '<2s', '2-3s' or '>3s'					0.26	69.54
Sunken eyes indicated as present or absent					5.48	85.11
Return of skin pinch indicated as 'immediate', '1-2s' or '>2s'					5.74	80.46
Character of pulse indicated as 'normal' or 'weak'					0.00	47.68
Ability to drink or breastfeed reported in the affirmative or negative	2.95	69.31	2.22	79.98	2.87	82.10
Level of consciousness classified as 'alert', 'responsive to voice', 'responsive to pain' or 'unconscious'	1.99	60.26	0.78	66.83	1.83	67.90
Diagnosis						
Illness severity classification made consistent with guidelines	7.32	65.18	9.87	81.41	24.28	62.70
Treatment						
Choice of drug for treatment is consistent with illness severity as recommended in guidelines	6.01	40.80	1.11	22.95	40.47	61.61
Drug administered via recommended route*	0.23	44.89	8.20	63.74		
Dose of drug correct for child's body weight according to dosing schedule	1.25	12.86	3.99	45.73	26.89	44.26
Drug administered at recommended daily frequency	5.56	46.12	8.65	61.89	67.62	68.03
Any duration of treatment specified*	6.01	31.86	7.76	54.10		

*Route of drug administration is implicit in the choice of drug for diarrhoea, and duration of treatment need not be explicitly specified

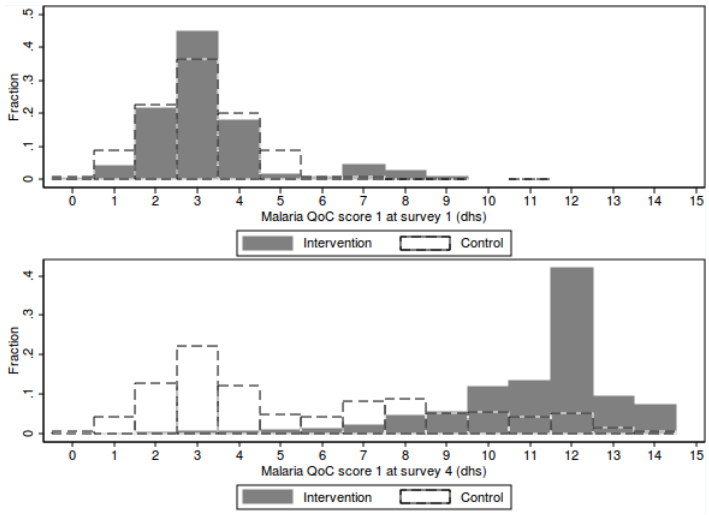


Figure 5.2-1: Distribution of the basic process-of-care score for malaria comparing baseline and main endpoint scores in the intervention and control hospitals

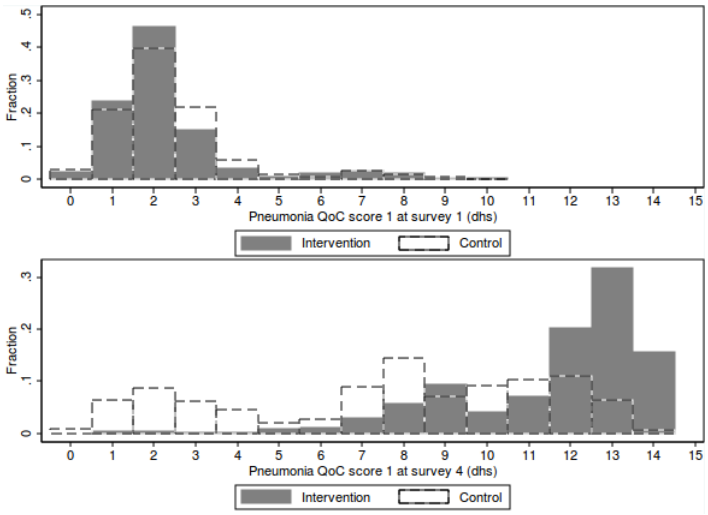


Figure 5.2-2: Distribution of the basic process-of-care score for pneumonia comparing baseline and main endpoint scores in the intervention and control hospitals

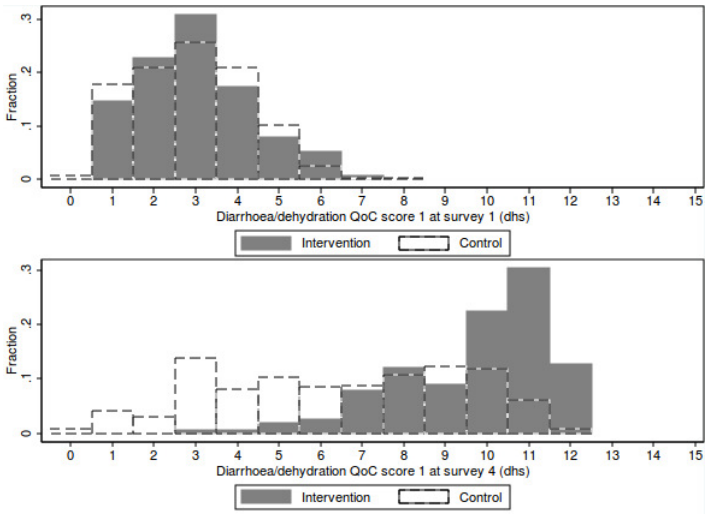


Figure 5.2-3: Distribution of the basic process-of-care score for diarrhoea/dehydration comparing baseline and main endpoint scores in the intervention and control hospitals

5.2.4. Shortfalls of the basic score

The 15- and 13-point scales of the basic score failed to meet some desired properties of the process-of-care measure. First, the scales were heavily weighted towards the assessment domain, which contributed more than half of the items on the scale. Consequently it was possible to achieve a mid-range score, interpretable as 'average' care, by simply completing the clinical assessment even if diagnosis and treatment were not performed according to guideline recommendations. A case in point was a group of 313 out of the 507 children with malaria and process-of-care scores of 8 out of 14 (which was approximately at the median of the distribution) for whom all 8 assessment tasks were completed but neither was a valid severity classification of their illness made nor was there a record of any recommended treatment given.

Secondly, there was a high degree of correlation between several items in this scale. This was assessed using tetrachoric correlation coefficients, which are the preferred measure of correlations between binary items, since Pearson correlation coefficients are valid only for normally distributed variables for reasons explained in section 4.4. Tetrachoric correlation coefficients between items in this scale are shown in Table 5.2-4, Table 5.2-5, Table 5.2-6 and Table 5.2-7. Correlations at baseline – before the intervention had had a chance to change practice and potentially alter the relationships between items in the scale – were also examined and are presented in Appendix A.10.

Table 5.2-4: Tetrachoric correlation matrix of assessment items in the basic malaria process-of-care score

	fever	convulsions	acidotic breathing	pallor	(in)ability to drink	level of consciousness	indrawing	malaria test
fever	1.00							
convulsions	0.64	1.00						
acidotic breathing	0.64	0.94	1.00					
pallor	0.30	0.44	0.52	1.00				
(in)ability to drink	0.62	0.91	0.99	0.52	1.00			
level of consciousness	0.58	0.87	0.92	0.54	0.92	1.00		
indrawing	0.61	0.89	0.99	0.49	0.97	0.90	1.00	
malaria test	0.20	0.23	0.25	0.13	0.25	0.30	0.22	1.00

Table 5.2-5: Tetrachoric correlation matrix of assessment items in the basic pneumonia process-of-care score

	cough	difficult breathing	central cyanosis	(in)ability to drink	level of consciousness	grunting	indrawing	resp. rate
cough	1.00							
difficult breathing	0.55	1.00						
central cyanosis	0.52	0.64	1.00					
(in)ability to drink	0.68	0.79	0.89	1.00				
level of consciousness	0.58	0.71	0.82	0.90	1.00			
grunting	0.65	0.80	0.92	0.98	0.91	1.00		
indrawing	0.59	0.78	0.86	0.95	0.87	0.97	1.00	
resp. rate	0.39	0.38	0.45	0.59	0.54	0.59	0.58	1.00

Table 5.2-6: Tetrachoric correlation matrix of assessment items in the basic diarrhoea/dehydration process-of-care score

	diarrhoea	vomiting	capillary refill	level of consciousness	(in)ability to drink	sunken eyes	skin pinch	pulse
diarrhoea	1.00							
vomiting	0.61	1.00						
capillary refill	0.36	0.91	1.00					
level of consciousness	0.37	0.84	0.86	1.00				
(in)ability to drink	0.42	0.94	0.93	0.88	1.00			
sunken eyes	0.46	0.91	0.94	0.88	0.97	1.00		
indrawing	0.42	0.91	0.94	0.87	0.95	0.97	1.00	
pulse	0.41	0.80	0.79	0.74	0.83	0.84	0.82	1.00

Table 5.2-7: Tetrachoric correlation matrix of treatment items in the basic process-of-care score

		Malaria					Pneumonia					Diarrhoea/Dehydration		
		drug	**route**	**dose**	**freq.**	**dur.**	**drug**	**route**	**dose**	**freq.**	**dur.**	**drug**	**dose**	**freq.**
Malaria	**drug**	1.00												
	route	0.80	1.00											
	dose	0.26	0.64	1.00										
	freq.	0.85	0.93	0.75	1.00									
	dur.	0.12	0.27	0.91	0.58	1.00								
Pneumonia	**drug**						1.00							
	route						0.75	1.00						
	dose						0.33	0.85	1.00					
	freq.						0.68	0.97	0.90	1.00				
	dur.						0.65	0.93	0.81	0.94	1.00			
DnD	**drug**											1.00		
	dose											0.66	1.00	
	freq.											0.44	0.19	1.00

The correlations between 10 pairs of assessment items for malaria, 11 pairs for pneumonia and 19 pairs for diarrhoea/dehydration were ‘very strong’ (tetrachoric correlation coefficients equal to or greater than 0.80) according to the criteria suggested by Evans (1996), as were half of the correlations between treatment item pairs for malaria and pneumonia. Furthermore the malaria assessment and treatment item tetrachoric correlation matrices were not positive (semi)definite: this means there was a high degree of linear dependency between multiple items. These characteristics implied that there was a lot of redundancy between these items: some items did not add any discriminative value to the score, and collapsing them into fewer items, preferably in a systematic process that maintained clinical relevance of the resulting items, could achieve a more balanced scale.

Thirdly, differences in scale ranges across diseases (0–14 for malaria and pneumonia, 0–12 for diarrhea/dehydration) also made it impossible to directly compare their process of care. Various forms of transformation or standardization of the original scores into new scales have been proposed to remedy this [Colman *et al.* 1997]. However these solutions also potentially introduce a new problem: a score produced from summation of indicator items is an ordered categorical outcome whose very intuitive interpretation or unit of measure – an important characteristic for reporting it to its end users – is a count of indicator items achieved. Mathematical transformations, which are only valid for continuous outcomes, would distort and obscure this meaning and therefore do not make for a suitable solution for obtaining a common scale that allows for the desired direct comparison across diseases while preserving the very specific meaning of the scale.

5.2.5. The modified score

The assessment items in the basic score were modified by collapsing them into discrete clinical decision points that constitute the desired processes outlined in the guidelines. These items, listed in Table 5.2-8 and Table 5.2-9, were also designed with the aim of making them generic to the process of care of all three diseases considered, and arguably most other acute childhood illnesses. To this end, assessment items were grouped into three modified items: (1) primary assessment signs required to diagnose the disease of interest; (2) secondary assessment signs necessary to distinguish between disease severity classifications; (3) a third modified item representing complete documentation of all required assessment signs.

For example, for malaria the primary assessment sign was fever. Secondary signs depended on severity. According to guidelines severe malaria was the correct diagnosis for a child who, in addition to fever, presented with at least one danger sign – convulsion, acidotic breathing, inability to drink or breastfeed, altered consciousness or pallor with respiratory distress indicated by grunting or indrawing. Fever in the absence of any danger sign was to be classified as non-severe malaria. A clinician was required to completely exclude the presence of a danger sign to correctly diagnose non-severe malaria. For this reason a complete secondary assessment for non-severe malaria meant documentation of all the danger signs, and this was a higher threshold of performance than was set for severe malaria where the clinician simply had to document fever and the presence of at least one danger sign. The indicator for complete assessment then distinguished between process of care that met the minimum threshold for good care from those that went the extra mile and undertook all tasks as recommended in the guidelines. This approach was extended across the various severity classifications of pneumonia and diarrhoea/dehydration.

For diagnosis, the binary indicator of whether a relevant severity classification was made was retained unchanged. However for treatment two modified items were generated: (1) selection of any broadly relevant drug for treatment of the disease diagnosed, and (2) correct use of selected drug which included correct dose, appropriate route of delivery, frequency and duration where applicable. As with the basic score, the resulting process of care score – named the *modified* score because of its adaptation from the basic one – was a sum of the modified indicator scores for each individual.

Table 5.2-8: Items in the assessment and diagnosis domains of the modified process-of-care score

Domain	Disease		
	Malaria	**Pneumonia**	**Diarrhoea/dehydration**
Assessment *Each grouped item scored 1 if all of its elements are documented (present, absent, quality or quantity) and 0 otherwise*	1. Primary signs: fever 2. Secondary signs: convulsions or acidotic breathing or (in)ability to drink/breastfeed or AVPU, or pallor in the presence of grunting or indrawing if severe, or convulsions and acidotic breathing and (in)ability to drink/breastfeed or AVPU, or pallor and grunting and indrawing if non-severe 3. Complete assessment: all signs documented	1. Primary signs: cough or difficult breathing 2. Secondary signs: central cyanosis or (in)ability to drink/breastfeed or AVPU or grunting or acidotic breathing if very severe, or central cyanosis and (in)ability to drink/breastfeed or AVPU, and grunting and acidotic breathing if severe, or central cyanosis and (in)ability to drink/breastfeed or AVPU, and grunting and acidotic breathing and respiratory rate if non-severe. 3. Complete assessment: all signs documented	1. Primary signs: diarrhoea and/or vomiting 2. Secondary signs: capillary refill or AVPU or (in)ability to drink/breastfeed, and pulse if shock, or capillary refill and AVPU or (in)ability to drink/breastfeed and sunken eyes and skin pinch and pulse if severe, some or no dehydration 3. Complete assessment: all signs documented
Diagnosis *Item score is 1 if a relevant severity classification is indicated, 0 otherwise*	1. Classification: severe or non-severe	1. Classification: very severe, severe or non-severe	1. Classification: shock, severe, some or none

Table 5.2-9: Items in the treatment domain of the modified process-of-care score

Domain	Disease		
	Malaria	**Pneumonia**	**Diarrhoea/dehydration**
Treatment *'Drug' score is 1 if correct (singly or in recommended combinations where applicable) according to guidelines for indicated severity classification* *'Correct use' scores 1 if dose, route, frequency and duration whichever applicable, of selected drug(s) are correct following guideline recommendations for their use, 0 otherwise*	Severe malaria: 1. Drug: quinine (loading and maintenance) 2. Correct use: Route is IV or IM *and* dose is 20mg/kg loading, 10mg/kg maintenance +/- 20% *and* frequency is twice daily *and* duration is stat for loading dose and any duration for maintenance dose Non-severe malaria: 1. Drug: artemether-lumefantrine or quinine 2. Correct use: Route is oral *and* dose is 5-14.9kg – 1 tab; 15-24.5kg – 2 tabs; 25-34.9kg – 3 tabs; 35kg+ - 4 tabs, *and* frequency is twice daily for AL and thrice daily for quinine *and* duration is any duration specified	Very severe pneumonia: 1. Drug: penicillin and gentamicin and oxygen 2. Correct use: Route is IV or IM *and* dose is penicillin 50,000iu/kg, gentamicin 7.5mg/kg (both +/- 20%) *and* frequency is penicillin x4, gentamicin x1, oxygen any specified *and* duration is any specified Severe pneumonia: 1. Drug: Penicillin only (no gentamicin) 2. Correct use: Route is IV or IM *and* dose is 50,000iu/kg +/-20% *and* frequency is x4 *and* duration is any specified Non-severe pneumonia: 1. Drug: Amoxicillin or cotrimoxazole 2. Correct use: Route is oral *and* dose is Amoxicillin 25mg/kg, cotrimoxazole 24mg/kg +/-20% *and* frequency is Amoxicillin x3, cotrimoxazole x2 *and* duration is any specified	Shock: 1. Drug: normal saline or Ringer's lactate/Hartmann's solution 2. Correct use: Dose is volume/timex4 within +/-20% of 20ml/kg *and* frequency is at least 1 in an hour Severe dehydration: 1. Drug: Ringer's or ORS 2. Correct use: Dose is total vol/time within +/- 20% of 30ml/kg + 70mg/kg in 3 hours for >1yr or in 6 hours for < 1yr of Ringer's or total vol/time within +/- 20% of 100ml/kg in 6 hours *and* frequency is step 1/2 used Some dehydration: 1. Drug: ORS 2. Correct use: Dose is vol/timex4 within +/- 20% of 75ml/kg *and* frequency: at least 1 in an 24 hours No dehydration: 1. Drug: ORS 2. Correct use: Dose is 10ml/kg +/-20% *and* frequency is any specified

5.2.6. Characteristics of the modified score

The modified score was a 7-point score (range 0 – 6) across all three diseases, being the sum of six binary items contributing to it. It was therefore possible to directly compare process of care across the diseases using the same indicators. These items were grouped in the same three domains as in the basic score. The assessment domain was a 4-point score (range 0 – 3), diagnosis a binary score as in the basic score and the treatment domain a 3-point score (range 0 – 2).

There was a very wide variation in the proportion of modified score items achieved across the diseases at the various time points similar to the basic score (Table 5.2-10). At baseline this ranged from 0% complete assessment in all three diseases to 98.4% documentation of the primary signs of diarrhoea/dehydration. Documentation of primary signs of all diseases was over 85% both at baseline and at the main study end-point; however secondary signs were poorly documented at baseline in all diseases. As with the basic score all indicators showed improvement, ranging from an absolute improvement of as low as 0.6% in the primary signs of diarrhoea to 63.6% improvement in documentation of secondary signs of pneumonia.

Table 5.2-10: Percentages of children for whom quality items from the modified score were achieved across all hospitals in the baseline and endpoint surveys

Item in process-of-care score for disease	**Percentage of children in which item was achieved**					
	Malaria		**Pneumonia**		**Diarrhoea/ Dehydration**	
	Baseline n=1,763	**End n=1,789**	**Baseline n=902**	**End n=1,194**	**Baseline n=383**	**End n=732**
Primary signs/symptoms of illness documented	86.73	95.42	93.90	98.16	98.43	99.04
Secondary signs/symptoms of illness documented	35.22	83.23	0.11	63.74	0.00	29.37
Complete documentation of signs/symptoms	0.00	47.12	0.00	41.79	0.00	37.98
Illness severity classification made consistent with guidelines	7.32	65.18	9.87	81.41	24.28	62.70
Choice of drug for treatment is consistent with illness severity as recommended in guidelines	6.01	40.80	1.11	22.95	40.47	61.61
Selected drug used as recommended in guidelines	0.17	29.40	32.26	60.97	32.64	44.81

Despite the reduction in the range of its scale, distributions of the process of care scores were still normal, more so at baseline and post-intervention in the control group where there was little heterogeneity of the score across hospitals (Figure 5.2-4, Figure 5.2-5 and Figure 5.2-6). The intervention hospitals had, as a group, higher scores than the controls, signalling an intervention effect similar to what was observed with the basic score.

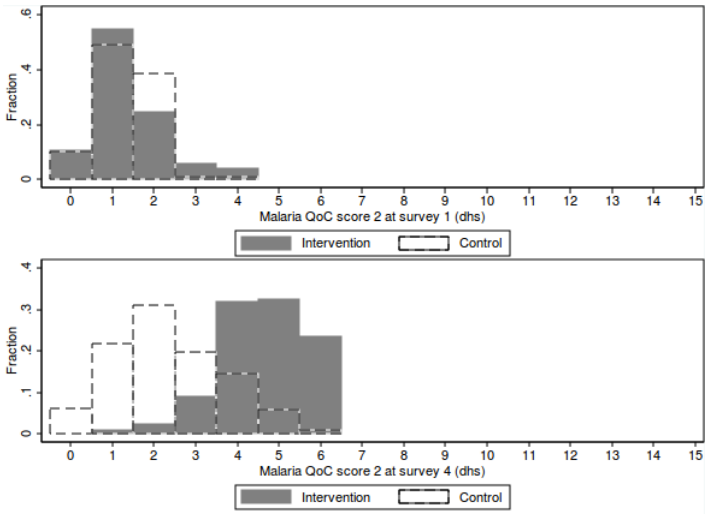


Figure 5.2-4: Distribution of the modified process-of-care score for malaria comparing baseline and main endpoint scores in the intervention and control hospitals

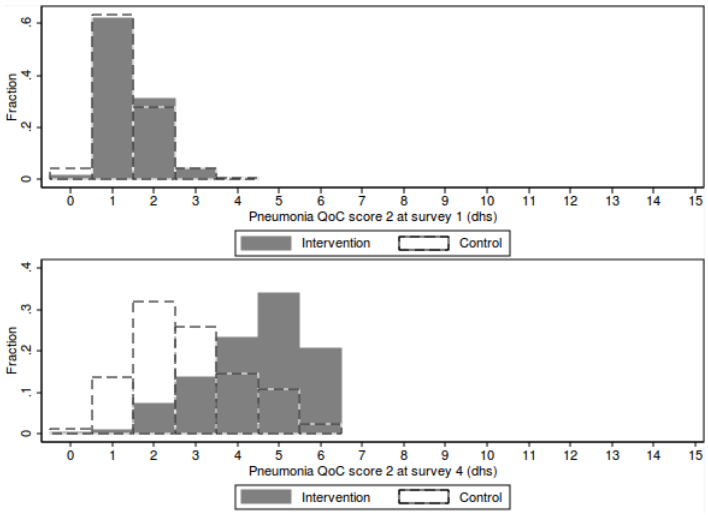


Figure 5.2-5: Distribution of the modified process-of-care score for pneumonia comparing baseline and main endpoint scores in the intervention and control hospitals

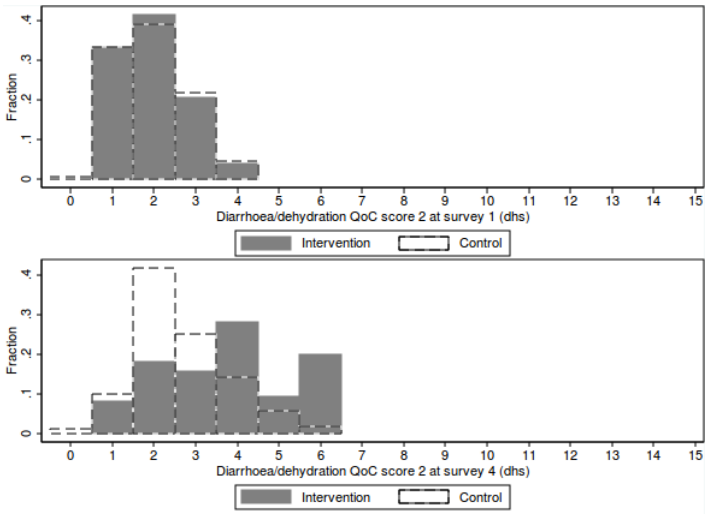


Figure 5.2-6: Distribution of the modified process-of-care score for diarrhoea/dehydration comparing baseline and main endpoint scores in the intervention and control hospitals

Tetrachoric correlation coefficients of items in the same domain ranged between 0.34 to 0.62 ('weak' to 'moderate' according to Evans' criteria) showing much less of the co-dependence between items in the basic score (Table 5.2-11). Between-domain correlation of items was also in this range. However there was a perfect correlation between the classification indicator and the drug choice indicator in malaria and pneumonia, and although this was expected – the choice of drug depended on severity classification – it may suggest that there still remained some redundancy between items.

Table 5.2-11: Tetrachoric correlation matrix of items in the modified process-of-care score

Complete assessment indicator has been excluded because it is perfectly correlated with the other two assessment items by design. 'pri.', 'sec.' and 'class.' are primary signs, secondary signs and illness severity classification indicator items respectively

		Malaria					**Pneumonia**					**Diarrhoea/Dehydration**				
		pri.	**sec.**	**class.**	**drug**	**use**	**pri.**	**sec.**	**class.**	**drug**	**use**	**pri.**	**sec.**	**class.**	**drug**	**use**
Malaria	**pri.**	1.00														
	sec.	0.60	1.00													
	class.	0.56	0.83	1.00												
	drug	0.46	0.66	1.00	1.00											
	use	0.24	0.49	0.48	0.45	1.00										
Pneumonia	**pri.**						1.00									
	sec.						0.62	1.00								
	class.						0.49	1.00	1.00							
	drug						0.23	0.57	1.00	1.00						
	use						0.15	0.28	0.28	0.51	1.00					
DnD	**pri.**											1.00				
	sec.											0.34	1.00			
	class.											0.15	0.87	1.00		
	drug											0.11	0.24	0.27	1.00	
	use											0.21	0.24	0.26	0.60	1.00

5.2.7. The combined score

The basic and modified scores allowed for the measuring of process of care in each of the three diseases separately. However it was quite common for a child to be diagnosed with more than one disease at each episode of hospital admission. Specifically, of the 12,036 admission episodes over the course of this study 6,150 (51.1%) had only one of the three diseases diagnosed, 4,188 (34.8%) had two and 446 (3.7%) had all three. The children represented by the remaining 1,252 admission episodes (10.4%) had other diseases than the three which were the focus of this thesis.

Measuring the patient-level process of care for the 4,634 children with more than one disease thus required some suitable combination of the disease-specific scores. To maintain a consistent scale across the various possible co-morbidity patterns and to avoid reintroducing the problem of non-comparable scores, the combined score was structured around the modified score since the latter already had equivalent items in each domain across the three diseases. An intuitive approach to combining the item scores for an admission episode was to use the arithmetic mean of the disease-specific scores for each item.

For example a child with all three diseases for whom the primary assessment score was 1 in two diseases and 0 in the third had a combined primary assessment score of $^{2}/_{3}$. Similarly a child with malaria and pneumonia for whom only the latter disease had a valid severity classification had a combined severity classification score of ½. As in the basic and modified scores the overall combined score was the sum of the item scores. However unlike the modified score this approach created non-integer scores which no longer represented a count of guideline-recommended process-of-care tasks completed by the clinician (Figure 5.2-7). Thus a score of $4^{1}/_{3}$, which implied four-and-a-third discrete assessment tasks completed for a child with all three diseases, could have instead been the result of an assessment score of $1^{2}/_{3}$, a diagnosis score of 1 and a treatment score of $1^{2}/_{3}$.

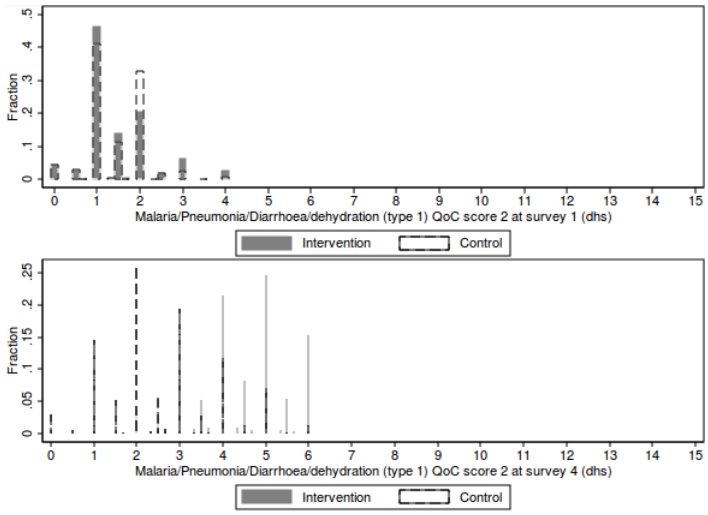


Figure 5.2-7: Distribution of the combined process-of-care score created from the arithmetic mean of disease-specific item scores, comparing baseline and main endpoint scores in the intervention and control hospitals

Thus an alternative approach to combining item scores was based on defining combined items which depended on the diseases a child had been diagnosed with. In this approach a combined item was assigned a score of 1 if the equivalent items in each of the diagnosed diseases had scored 1, and zero otherwise – an all-or-none combination of disease-specific item scores. For example if a child had malaria and pneumonia then the combined primary assessment score was 1 if primary assessment items for both malaria and pneumonia (presence of fever documented, and presence of cough or difficult breathing documented) scored 1; if only one or none of them were documented then the combined score was zero. Although this approach made it more difficult to achieve each level of the score, it intuitively reflected the clinical reality that multiple diagnoses increased the number of guideline-recommended tasks required to effectively manage illness thereby increasing the difficulty in management of illness, and is often associated with higher risk of mortality [Fenn *et al.* 2005].

The combined score retained the main characteristics of the modified score: it was a 7-point score (range 0 – 6); the assessment component was a 4-point score (range 0 – 3), the diagnosis one a binary score and the treatment domain a 3-point score (range 0 – 2). The proportion of items achieved across the diseases at the various time points also varied widely. At baseline it ranged between 0 to 15% in all but one item, but varied widely between 26% and 87% across all items at the endpoint (Table 5.2-12).

All indicators showed the expected change over time, ranging from an absolute increase of 9.8% in the documentation of primary signs of illness, to 50.3% improvement in documentation of illness severity classifications. Also worth noting was the increased denominator at each time-point, a consequence of the combined score now representing the process-of-care measure of a child's admission episode rather than being a disease-specific measure.

Table 5.2-12: Percentage of children for whom quality items from the combined score were achieved across all hospitals in the baseline and endpoint surveys

Item in process-of-care score for disease	**Percentage of children in which item was achieved**	
	Baseline (n=2,450)	**End (n=2,714)**
Primary signs/symptoms of illness documented	77.59	87.36
Secondary signs/symptoms of illness documented	15.02	51.69
Complete documentation of signs/symptoms	0.00	33.86
Illness severity classification made consistent with guidelines	7.80	58.14
Choice of drug for treatment is consistent with illness severity as recommended in guidelines	4.65	26.23
Selected drug used as recommended in guidelines	5.22	28.26

Its distribution also had a similar appearance and changes over time to those of disease-specific scores in Figure 5.2-4, Figure 5.2-5 and Figure 5.2-6, namely approximately normal, superimposed distributions of the two groups at baseline followed by a shift in the intervention group of hospitals towards higher scores at the main end-point. This is illustrated in Figure 5.2-8.

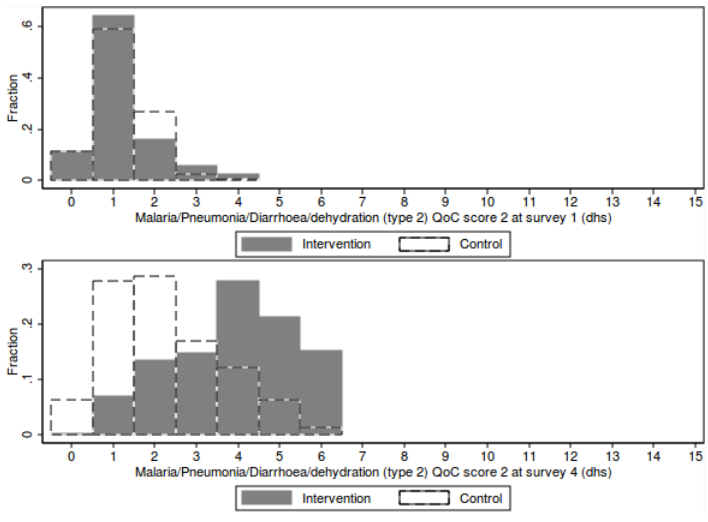


Figure 5.2-8: Distribution of the combined process-of-care score created from an all-or-none combination of disease-specific item scores, comparing baseline and main endpoint scores in the intervention and control hospitals

5.3. Measuring agreement between basic and modified scores through cluster level variability

The 7-point modified score mirrored the response of the basic 15-point score to hypothesised changes in process of care in the study hospitals across group and time. Nevertheless it was likely that some loss of variability of the score – and consequently ability to discriminate between clinically important differences in process of care – which was not apparent when comparing the distributions, could have occurred as a consequence of collapsing and grouping of indicators. In fact a number of studies have highlighted such losses occurring as a consequence of not only collapsing an ordered categorical measure into fewer categories as has been done between the 16- and 7-point scores [Srinivasan & Basu 1989] but also due to categorizing a continuous measure [Bennette & Vickers 2012, Naggara *et al.* 2011, Altman 2005] or using a categorical outcome to measure a continuous latent construct [Rhemtulla *et al.* 2012, Naggara *et al.* 2011].

In recognising loss of variability as a potential weakness of the described approach to score design, the amount of agreement between the 15-point basic score and the 7-point modified score was quantified by estimating the variability in the basic score that could be explained by the modified score alone at hospital and clinician level. Variations of two estimation models of the family of generalised linear models were fitted reflecting different assumptions about the level of measurement of the score: a linear model treating the score as a continuous measure, and an ordered categorical model assuming an ordered categorical measure. First a model of the basic score, Y_i, with the modified score, X_i, as the only covariate was fitted:

$$E[g(Y_{ijk})] = \alpha + \beta X_{ijk} + u_j + u_{jk}$$

u_j represents clinician-level random variation which was assumed to be normally distributed with a mean of zero and variance represented by $\sigma_{u_j}^2$, conditional on hospital-level variation represented by u_{jk} and also assumed to be normally distributed with a mean of zero and variance represented by $\sigma_{u_{jk}}^2$. The link function, g, was the identity link which assumed normally distributed scores for the linear regression models, and probit link for the ordered logistic regression models. A 'null' model of the basic score alone was then fitted:

$$E[g(Y_{ijk(null)})] = \alpha_{(null)} + u_{j(null)} + u_{jk(null)}$$

The increase in hospital and clinician-level variances in the null model compared to the calibration model represented the amount of variability in the basic score explained by the modified score alone, which was then calculated as:

$$\rho_{hosp} = \frac{u_{jk(null)} - u_{jk}}{u_{jk(null)}}$$

for variability at the hospital level and:

$$\rho_{clin} = \frac{u_{j(null)} - u_{j}}{u_{j(null)}}$$

for variability at the clinician level. The 95% confidence intervals for these estimates were obtained by bootstrapping: this involved repeated re-sampling with replacement from the estimation sample followed by re-estimation; the 95% confidence intervals was the range of values which included 95% of the repeated sample estimates.

Table 5.3-1: Cluster-level agreement between the basic and modified scores across the three diseases

Method	Percentage of variability in the basic score explained by the modified score (bootstrap 95% confidence intervals)		
	Malaria n=6,356	**Pneumonia** n=4,128	**Diarrhoea/ dehydration** n=1,910
Hierarchical linear regression to estimate hospital level variances	93.8% (92.4 – 95.3)	91.6% (88.1 – 95.1)	72.3% (66.1 – 78.7)
Hierarchical linear regression to estimate clinician level variances	95.3% (94.5 – 96.1)	90.0% (88.3 – 91.8)	51.8% (48.2 – 53.4)
Hierarchical ordered logistic regression to estimate hospital level variances	65.2% (56.2 – 74.3)	76.8% (64.3 – 89.2)	39.5% (18.1 – 60.9)
Hierarchical ordered logistic regression to estimate clinician level variances	73.8% (68.0 – 79.5)	73.6% (66.2 – 81.0)	14.5% (5.6 – 34.5)

There was some loss of information as a consequence of the score transformation. Assuming a normal distribution, the loss of variability was less than 10% at both hospital and clinician level for malaria and pneumonia, but comparatively larger for diarrhoea/dehydration, with close to 50% of clinician level variability lost in the transformation. The greater loss of variability in the diarrhoea score was most probably attributable to the fact that almost twice as many pairs of items in the disease-specific basic score were very strongly correlated (Table 5.2-6). When the score was treated as

an ordered categorical outcome, the amounts of variability explained in both levels across all three diseases were comparatively lower, ranging between 39.5% and 65.2% at hospital level and 14.5% to 73.8% at clinician level. Nevertheless in all cases, except the ordinal assumption on the diarrhoea score, the amount of explained variability was still 'large' according to Cohen's criteria (in the latter it was 'medium') [Cohen 1992]. It was thus concluded that despite the inevitable loss of variability upon collapsing the basic score items into fewer groups, there was still a good level of agreement between the basic and modified scores.

5.4. Summary

A score to measure process of care has been described in this Chapter, starting with the identification of items that represent key process-of-care recommendations from clinical practice guidelines. Binary scores of compliance with recommendations have been defined for each of these items, and a basic score has been constructed by summing up the binary scores. The score possesses good face-validity as a measure of adherence to good care practices and is potentially useful in its own right. However a major shortcoming is its inability to cater for direct comparison of the process of care across multiple diseases. Thus a second approach to scoring, a modification of the basic score, has been described, which has produced a score with this feature. The modification has come at the cost of some loss of variability in the score, a known consequence of collapsing continuous or categorical measures into fewer categories. Some validation of the proposed measure has also been undertaken in this Chapter. Content validity has been ensured through the selection from clinical practice guidelines of items describing how care ought to be provided. Face-validity has been explored by graphically examining the sensitivity of the score to changes in quality of care previously reported in the Kenyan district hospitals study. A shift has been observed in the distribution of the score between intervention and control hospitals at baseline and at the main study endpoint which is consistent with an improvement in quality of care. Further validation is presented in the next Chapter.

Chapter 6 – Further Validation

6.1. Introduction

In this Chapter three new aspects of the validity of the measure are examined in addition to those highlighted in Chapter 5, namely construct validity, criterion-related validity and external validity. Construct validity is tested by examining the extent to which items of the measure aggregate into domains which are consistent with generic phases of the clinical process. Criterion-related validity is assessed in terms of the association of the measure with an outcome of care. External validity is investigated by exploring how the measure works in situations different from those in which it has been designed and tested.

6.2. Construct validity: do the modified score items aggregate into the proposed domains?

Items of the modified and combined scores have been grouped into three domains, namely: assessment, which constitutes the indicators of documentation of the primary and secondary signs of illness; diagnosis, which is a single-item domain of an indicator of illness severity classification; and a treatment domain of indicators of correct choice of medical intervention (drug, intravenous fluids or oral fluids) and their correct use in terms of dose, frequency and duration of administration. This conceptual grouping of items into domains allows for measurement at the distinct stages of the process of care represented by the proposed domains, a key objective of this work. It is also consistent with the sequence of clinical process and is therefore arguably a sensible approach to measuring and reporting quality of care. What is still unclear at this stage is whether there is any statistical evidence that supports these conceptual groupings.

To address this question structural equation modelling (SEM) was used to explore the extent to which these conceptual groupings were consistent with patterns of correlations between observed items in the data across each of the three diseases of interest. Some statistical principles underlying CFA/SEM and generalised linear modelling (GLM, 'regression') have been described in sections 4.2 and 4.5 respectively.

6.2.1. Specification of a structural equation model of the modified score

The model specification was based on the expected association between the observed items and the latent factors that represent the conceptualised groupings. The first assumption was that the primary and secondary assessment indicators were the observable partial manifestations of a latent factor named 'assessment'. For this reason they were expected to be highly correlated with each other. Similarly, the indicators of correct drug choice and use were thought to be manifestations of a latent factor named 'treatment' and expected to be highly correlated with each other. The degree to which 'assessment' and 'treatment' manifested themselves through the items in those domains was measured through their factor *loadings*. These were symbolised by λ subscripted by an initial identifying the item. A 'diagnosis' domain was not included as a factor in this model because there was only one indicator item for this domain and a factor would not be identifiable from it for reasons explained in detail in section 4.2.3.

It was also assumed that the score items were partial manifestations of the latent factors, and there remained some quantities of the latent factors unmeasured by them. These *residuals*, symbolised by ε subscripted by an initial identifying the factor, were assumed to be normally distributed with a mean of zero and some variance σ^2_{ε}. The third assumption was that the items were completely unassociated with each other except through the latent factors. For example, any association between the indicator of assessment for primary signs of illness and that of correct drug use was through the association between the assessment and treatment factors by the correlation of the variances of their residuals. These three assumptions underlie the confirmatory factor analysis, described in section 4.2.4, which is considered the measurement component of the model.

The structural part of the model was the component to adjust for the independent associations between indicator items and exogenous variables; the latter were variables assumed to exert their effect on the indicators solely through the latent factors. Time/survey and intervention/group were included as exogenous variables since their effects on quality of care, therefore on 'assessment' and 'treatment' have been demonstrated previously [Ayieko *et al.* 2011]. These effects, analogous to regression coefficients, were represented by β subscripted with identifiers of the variables and the latent factors they affected.

A path diagram (Figure 6.2-1) presents these assumptions about the relationships between items, factors and residuals according to current convention which follows McArdle's reticular action model [McArdle & McDonald 1984]. Indicator items and exogenous variables are represented by rectangles, factors by ovals and residual variances by circles. A unidirectional arrow represents a causal relationship; the arrow points from the assumed causative factor, item or residual towards the effected one. A bi-directional arrow represents a correlation between factors, items or residuals.

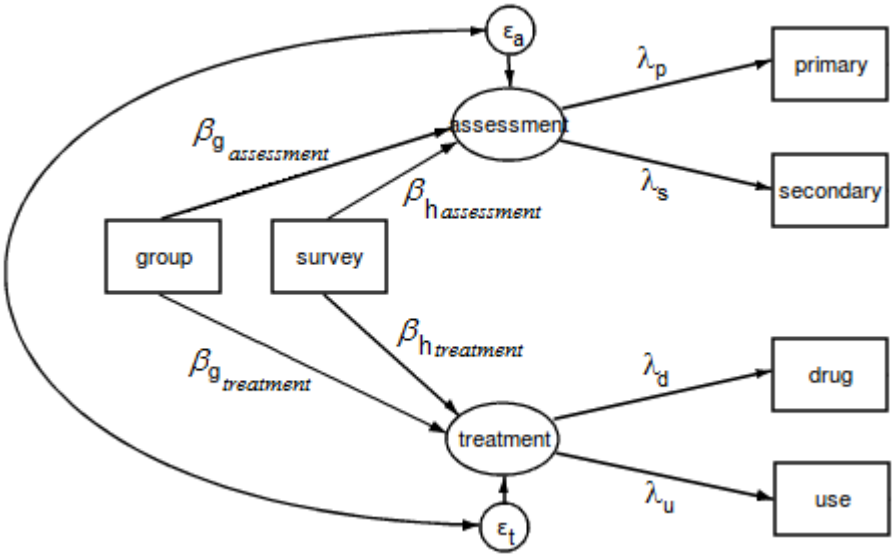


Figure 6.2-1: Path diagram of the structural equation model of the modified and combined scores

6.2.2. Estimation of parameters

The measurement part of the model quantifies the latent factors and expresses the probabilities of the observed indicator items in terms of their loadings on the latent factors. It is given by:

$$\Pr(primary = 1 \mid assessment) = \Phi\{\alpha_p + \lambda_p[assessment]\}$$

$$\Pr(secondary = 1 \mid assessment) = \Phi\{\alpha_s + \lambda_s[assessment]\}$$

$$\Pr(drug = 1 \mid treatment) = \Phi\{\alpha_d + \lambda_d[treatment]\}$$

$$\Pr(use = 1 \mid treatment) = \Phi\{\alpha_u + \lambda_u[treatment]\}$$

where $\Phi\{.\}$ is the cumulative normal distribution [StataCorp 2013]. The structural part of the model shows the effect of exogenous variables on the latent factors. It is:

$$assessment = \alpha_{assessment} + [group]\beta_{g\ assessment} + [survey]\beta_{h\ assessment} + \varepsilon_a$$

$$treatment = \alpha_{treatment} + [group]\beta_{g\ treatment} + [survey]\beta_{h\ treatment} + \varepsilon_t$$

Model parameters were estimated using M*plus* software version 5.1 [Muthén & Muthén]. An iterative robust weighted least-squares procedure – which is suitable for categorical items such as the process indicators in this model – was used for parameter estimation. This procedure assumes that residuals are normally distributed with very few extreme values. Estimation begins by fitting estimates of the measurement part of the model using weighted least squares, which involves identifying values of λ and the latent factors that solve the simultaneous equations expressing the probabilities of the observed indicator values for iteratively set values of the latent factors. The λ of the first item on each factor is fixed at 1 with a standard error of 0 to scale all other item loadings. This is followed by calculation of regression coefficients in the measurement part of the model and the standardised adjusted residuals. Robust weights are then estimated as a function of these residuals and applied to re-weight the estimates and the cycle repeated until convergence is achieved.

Although these data were clustered within hospital, this non-hierarchical model was preferred for two reasons. First, it was the more parsimonious option, since the alternative approach of fitting a model of 11 parameters using only 8 hospitals – effectively 8 observations – would have led to imprecise estimates. A minimum of 10 observations per parameter has been recommended to ensure precision of estimates [Schreiber *et al.* 2006]. Secondly, in these data the number of admissions for each disease varied across the 8 hospitals, and the hospital-level ICC was 0.17 [Opondo *et al.* 2011]. Simulation studies have shown that in such circumstances (i.e. unequal cluster sizes, few clusters, or very small or very large ICC) the assumption of independent and identically distributed variables in a hierarchical model is likely to be violated leading to unreliable between-group estimates [Hox & Maas 2001].

Listwise deletion was the preferred method of handling missing data in all analyses undertaken in this work. It is valid for data missing completely at random (MCAR). Data collection procedures followed in the district hospitals study made this the most likely mechanism for missingness. This is because all missing information in case record forms were coded ‘E’ (‘empty’) in the data collection tool and recoded to ‘0’ in

the indicators, and these implied poor quality of care. True missingness, which was most likely MCAR, occurred in the form of indicators for which either the actual observation made by the clinician or the explicit coding for the absence of such information was not done at data collection or lost during data entry.

6.2.3. Results of the structural equation model of the modified score

Table 6.2-1, Table 6.2-2 and Table 6.2-3 present the parameter estimates from the structural equation models of the modified score for malaria, pneumonia and diarrhoea/dehydration respectively. Each disease-specific model was limited to cases which had been diagnosed with that disease alone. This was done to exclude any potential effect of multiple diagnoses, since a number of the signs and symptoms of illness (e.g. level of consciousness, inability to drink/breastfeed, indrawing) were relevant to two or three diseases. For this reason only 2,930 cases of malaria, 1,409 of pneumonia and 529 of diarrhoea/dehydration were used to estimate these parameters and assess the model's goodness of fit.

The suitability of the proposed conceptual groupings was assessed through factor loadings – the parameters represented by the λ's – and model fit indices. All factor loadings across all three diseases were significantly greater than zero, except the secondary assessment indicator for the diarrhoea/dehydration model (which was the smallest model by sample size). Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were large and Root Mean Square Error of Approximation (RMSEA) small, indicative of good to marginal fit according to the suggested interpretation of fit indices in Table 4.2-2.

Modifications to the proposed factor model in an attempt to improve fit, including a one-factor model which assumed that all indicators loaded onto a single factor, and for diarrhoea/dehydration, a model replacing the secondary assessment indicator with the diagnosis indicator, resulted in poorer fitting models (Appendix A.11). For this reason the proposed aggregation of items into three domains was adopted as the most preferred conceptual grouping of indicators within the score since it was the one best supported by the patterns observed in these data and most likely to be intuitive to the intended audiences.

Table 6.2-1: Parameter estimates of the two-factor structural equation model of the modified score for malaria quality of care

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.534)	_	_
λ_s	1.958 (0.976)	0.250	< 0.001
λ_d	1.000 (0.706)	_	_
λ_u	0.886 (0.644)	0.055	< 0.001
$\beta_{g\ assessment}$	0.216 (0.196)	0.039	< 0.001
$\beta_{h\ assessment}$	0.183 (0.394)	0.025	< 0.001
$\beta_{g\ treatment}$	0.770 (0.466)	0.046	< 0.001
$\beta_{h\ treatment}$	0.423 (0.605)	0.023	< 0.001
var(ε_a)	0.246	0.039	< 0.001
var(ε_t)	0.310	0.043	< 0.001
cov(ε_a,ε_t)	0.229 (0.829)	0.031	< 0.001

n = 2,930; fit indices: χ^2_5 = 88.254, p-value < 0.001; CFI = 0.948; TLI = 0.874; RMSEA = 0.075

Table 6.2-2: Parameter estimates of the two-factor structural equation model of the modified score for pneumonia quality of care

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.650)	_	_
λ_s	1.614 (0.918)	0.254	< 0.001
λ_d	1.000 (1.050)	_	_
λ_u	0.516 (0.580)	0.064	< 0.001
$\beta_{g\ assessment}$	0.518 (0.358)	0.087	< 0.001
$\beta_{h\ assessment}$	0.378 (0.587)	0.057	< 0.001
$\beta_{g\ treatment}$	0.630 (0.273)	0.082	< 0.001
$\beta_{h\ treatment}$	0.304 (0.296)	0.043	< 0.001
var(ε_a)	0.286	0.093	0.002
var(ε_t)	1.125	0.160	< 0.001
cov(ε_a,ε_t)	0.287 (0.506)	0.053	< 0.001

n = 1,409; fit indices: χ^2_5 = 11.151, p-value = 0.0485; CFI = 0.992; TLI = 0.980; RMSEA = 0.030

Table 6.2-3: Parameter estimates of the two-factor structural equation model of the modified score for diarrhoea/dehydration quality of care

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.360)	_	_
λ_s	2.091 (0.709)	2.199	0.341
λ_d	1.000 (0.795)	_	_
λ_u	0.847 (0.685)	0.154	< 0.001
$\beta_{g\ assessment}$	0.260 (0.354)	0.278	0.349
$\beta_{h\ assessment}$	0.105 (0.333)	0.112	0.351
$\beta_{g\ treatment}$	0.501 (0.297)	0.105	< 0.001
$\beta_{h\ treatment}$	0.174 (0.241)	0.045	< 0.001
var(ε_a)	0.095	0.102	0.349
var(ε_t)	0.585	0.124	< 0.001
cov(ε_a,ε_t)	0.105 (0.444)	0.112	0.349

n = 529; fit indices: χ^2_5 = 13.093, p-value = 0.0108; CFI = 0.941; TLI = 0.822; RMSEA = 0.066

The two-factor model of the combined score further supported the proposed structure of the quality of care measure (Table 6.2-4). The fit indices were indicative of good fit. Additionally, factor loadings were all significantly greater than zero, with standardised loadings between 0.59 and 0.80. There was also a high correlation of 0.66 between the 'assessment' and 'treatment', conditional on group and survey (this was between 0.44 and 0.83 for the disease-specific scores), confirming that although the domains were distinct dimensions of the process of care they were still closely related as is expected of the dimensions of a construct.

Table 6.2-4: Parameter estimates of the two-factor structural equation model of the combined score for malaria, pneumonia and diarrhoea/dehydration

Parameter	**Estimate (standardised)**	**Standard error**	**p-value**
λ_p	1.000 (0.617)	_	_
λ_s	1.347 (0.801)	0.091	< 0.001
λ_d	1.000 (0.771)	_	_
λ_u	0.738 (0.591)	0.034	< 0.001
$\beta_{g\ assessment}$	0.369 (0.283)	0.031	< 0.001
$\beta_{h\ assessment}$	0.224 (0.401)	0.016	< 0.001
$\beta_{g\ treatment}$	0.517 (0.305)	0.029	< 0.001
$\beta_{h\ treatment}$	0.304 (0.419)	0.015	< 0.001
var(ε_a)	0.317	0.031	< 0.001
var(ε_t)	0.516	0.033	< 0.001
cov(ε_a,ε_t)	0.266 (0.657)	0.020	< 0.001

n = 8,476; fit indices: χ^2_5 = 187.008, p-value < 0.001; CFI = 0.980; TLI = 0.948; RMSEA = 0.039

6.3. Criterion-related validity: is the score associated with mortality?

Outcomes of care are not only intrinsic to the meaning of quality of care [Hammermeister *et al.* 1995] but are also the ultimate validation of quality of care [Donabedian 1966]. Mortality (or, more specifically, survival) is arguably the most relevant and objective outcome of care. As such, the demonstration of a favourable association between processes of care and mortality is a key step in validating this quality of care measure. Although indicators contributing to the proposed process of care measure should, at least in theory, be associated with outcomes such as death/survival, individual indicators can lack sufficient variability and power to show such associations or intervention effects except in very large samples and/or for very strong associations.

A case in point is a study exploring whether process indicators of care for elderly patients with pneumonia were associated with lower 30-day mortality which showed

improved survival with only 2 of the 4 processes examined [Meehan *et al.* 1997]. A similar finding in the Kenyan district hospitals study was that 6 of the 18 indicators used to compare performance between the intervention and control hospitals at the main study endpoint showed weak or no evidence of intervention effect [Ayieko *et al.* 2011], and only 1 of 3 key indicators of the process of care for malaria, pneumonia and diarrhoea/dehydration were associated with mortality, as shown in Table 6.3-1.

Table 6.3-1: Estimates of association between mortality and process of care based on a hierarchical logistic regression model of death on key process indicators adjusting for age, sex, group and survey.

Indicator	n	Adjusted odds ratio for death	95% confidence intervals	p-value
Quinine loading dose in severe malaria	1,792	1.32*	0.82 – 2.13	0.258
Correct dose of gentamicin in very severe pneumonia	434	0.70	0.38 – 1.27	0.239
Correct dose of IV fluids in severe dehydration	674	0.57	0.34 – 0.94	0.027

*Although not statistically significant, an increased odds of death associated with quinine loading dose could be due to poor management of harmful side effects of quinine, e.g. hypoglycaemia

Specifically, there was no evidence that the use of quinine loading dose in severe malaria and the correct dose of gentamicin in very severe pneumonia were associated with lower odds of death even though they are recommended best-practice for what ought to be good care, whereas there was strong evidence that correct use of IV fluids in dehydration was associated with a 43% reduction in odds of death.

To test the comparative efficiency of the score in detecting changes in the process of care, a hierarchical logistic regression model of the odds of death across the range of the combined score was fitted. The analysis was limited to data from children with at least one of the three diseases of interest to this work. The model adjusted for variables representing study design factors – group (intervention) and time (survey) and their interaction – and age and sex which were thought of *a priori* as possible confounders in the hypothesised association between the selected process and outcome measures. There was also adjustment for the nature, number and severity of disease all of which were known to increase the odds of death.

Proportions and associations of outcome and exposure variables with the combined score are presented in Table 6.3-2. There was evidence of association between the combined score and almost all the other exposure variables to be included in the model. There was also evidence of association between the all exposure variables and

mortality, as shown in Table 6.3-3. A number of variables had missing values. Potential causes and treatment of missing values is briefly discussed in section 6.2.2.

Table 6.3-2: Proportions (%) and associations of outcome and exposure variables with the combined score

Combined score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**n**	**p-value**
Outcome									
Alive	4.66	27.09	18.66	13.23	18.56	12.23	5.56	8,210	0.051
Dead	4.78	23.23	17.30	17.30	17.46	14.00	5.93		
Age									
< 12 months	5.05	28.40	18.17	14.19	17.64	11.57	4.97		
12 – 23 months	5.26	27.93	18.20	12.77	19.24	11.60	5.00		
24 to 35 months	3.78	22.34	20.45	13.23	18.38	14.43	7.39	8,476	<0.001
36 to 47 months	2.73	22.12	20.76	12.88	20.45	13.94	7.12		
48 to 59 months	4.73	26.72	18.68	13.53	18.43	12.32	5.59		
Sex									
Male	4.16	21.42	18.45	14.52	20.43	14.55	6.48	7,040	0.355
Female	3.75	22.80	18.89	15.01	20.58	12.92	6.04		
Number of diseases diagnosed									
1	3.27	20.87	20.62	12.76	18.30	15.90	8.28		
2	6.85	34.55	14.26	15.02	19.25	7.93	2.14	8,476	<0.001
3	5.29	35.29	33.24	10.29	12.35	3.24	0.29		
Severity									
Lowest	5.12	25.71	19.30	13.91	21.02	11.70	3.23		
Intermediate	1.83	9.56	17.69	21.20	28.52	13.52	7.68	5,812	<0.001
Highest	0.65	2.66	6.73	18.29	30.25	28.19	13.22		
Identity of disease									
Diarrhoea/dehydration	0.38	10.96	30.06	16.82	21.17	8.70	11.91		
Malaria	6.06	27.85	20.08	10.39	15.83	13.54	6.25	8,476	<0.001
Pneumonia	4.06	27.70	15.93	16.00	20.47	11.65	4.18		
Group									
Control	6.65	33.85	25.30	14.57	12.49	5.87	1.26	8,476	<0.001
Intervention	3.17	20.92	13.29	12.69	23.26	17.57	9.11		
Survey									
Baseline	11.33	61.56	21.12	4.25	1.74	0.00	0.00		
1st follow-up	1.96	13.31	16.07	17.50	26.78	17.92	6.47	8,476	<0.001
2nd follow-up	2.08	13.16	15.97	17.22	26.38	18.16	7.02		
End-point	3.06	16.69	20.60	15.85	20.65	14.40	8.75		
Overall	**4.73**	**26.72**	**18.68**	**13.53**	**18.43**	**12.32**	**5.59**		

Table 6.3-3: Proportions (%) and associations of exposure variables with mortality

Variable	Proportion of children who died (%)	Crude odds ratio	n	p-value
Combined score				
0	7.57	1.00		
1	6.41	0.84		
2	6.89	0.90		
3	9.45	1.27	8,210	0.051
4	6.99	0.92		
5	8.37	1.12		
6	7.84	1.04		
Age				
< 12 months	9.99	1.00		
12 – 23 months	5.65	0.54		
24 to 35 months	4.14	0.39	8,210	< 0.001
36 to 47 months	4.69	0.44		
48 to 59 months	5.56	0.53		
Sex				
Male	7.06	1.00	6,814	0.050
Female	8.32	3.83		
Number of diseases diagnosed				
1	7.51	1.00		
2	6.85	0.91	8,210	0.023
3	10.94	1.51		
Severity				
Lowest	5.34	1.00		
Intermediate	8.85	1.72	5,626	< 0.001
Highest	9.70	1.91		
Identity of disease				
Diarrhoea/dehydration	8.01	1.00		
Malaria	5.66	0.69	8,210	< 0.001
Pneumonia	7.39	1.12		
Group				
Control	6.51	1.00	8,210	0.006
Intervention	8.10	1.27		
Survey				
Baseline	5.82	1.00		
1st follow-up	7.27	1.27	8,210	0.007
2nd follow-up	8.37	1.48		
End-point	8.12	1.43		
Overall	**7.39**		**8,210**	

Although the three diseases considered in this analysis are often associated with illness of relatively short duration, it was not unusual to observe admission episodes lasting more than a few days. Adjustment was made of the duration of admission, but since admissions lasting more than a month were relatively rare – 2.56% of all cases – they were all collapsed into a single group. Excluding admission episodes lasting more than 7 days did not change the conclusions from the model (Appendix A.12). The presence and nature of clinical signs and symptoms and treatments were excluded from the model since they were either components of either the characteristics of the illnesses diagnosed – and already adjusted for – or the score itself. The model also adjusted for the clustered nature of observations by admitting clinician and hospital. Quadratic terms of statistically significant continuous predictor variables were fitted and retained in the model if they were significant and did not worsen model fit.

If Y_{ijk} is the outcome of patient i who was attended to by clinician j in hospital k, x_1 is the effect of the score or any indicator on mortality, and x_l are other covariates and quadratic terms associated with the outcome to be adjusted for, the model is:

$$\mathrm{E}[\mathrm{logit}(Y_{ijk})] = \alpha + \beta_1 x_{1ijk} + \Sigma_l \beta_l x_{lijk} + u_j + u_{jk}$$

with normally-distributed clinician-level random-effects u_j assumed to have a mean of zero and variance of $\sigma_{u_j}^2$ conditional on the normally-distributed hospital-level random effects $u_{jk,}$ whose variance is $\sigma_{u_{jk}}^2$. These assumptions are examined graphically in Appendix A.13.

Estimates from the model are presented in Table 6.3-4. They show strong evidence of a 14% reduction in adjusted odds of death per unit increase in the score. There were other notable findings as well. The odds of death decreased by 31% with each extra disease diagnosed. Although this may seem counterintuitive, it was probably a reflection of the guideline recommendation that clinicians should identify and treat all illnesses consistent with clinical signs and symptoms. It may also mean that children diagnosed with more than one disease were more likely to be admitted to hospital regardless of the severity of their illness.

Lastly, there was no difference in the adjusted odds of death comparing intervention and control groups across time. This was most probably due to the intervention exerting its

effect through improved process of care measured by the score which, once adjusted for, likely muted any direct group and time effects on mortality.

Table 6.3-4: Adjusted effect of quality of care measured using the combined score on death

	Adjusted odds ratio for death	95% confidence interval	p-value
Combined score	0.86	0.77 – 0.95	0.005
Age (years)	0.63	0.55 – 0.72	<0.001
Sex (female vs. male)	1.13	0.91 – 1.40	0.283
Number of diseases diagnosed	0.69	0.54 – 0.87	0.002
Severity (3=highest, 2=intermediate, 1=lowest)	1.49	1.27 – 1.75	<0.001
Identity of disease			
Diarrhoea/dehydration	1.00		
Malaria	0.83	0.54 – 1.29	< 0.001
Pneumonia	1.57	1.02 – 2.42	
Duration of admission (days)	0.95	0.92 – 0.98	< 0.001
Group			
Control	1.00		
Intervention	1.20	0.41 – 3.50	0.733
Survey			
Baseline	1.00		
1st follow-up	0.95	0.44 – 2.04	0.942
2nd follow-up	1.12	0.51 – 2.44	
End-point	1.06	0.50 – 2.24	
Group-survey interaction			
Intervention x 1st follow-up	1.50	0.55 – 4.12	
Intervention x 2nd follow-up	1.14	0.42 – 3.12	0.741
Intervention x End-point	1.45	0.55 – 3.85	
Random effects	**Variance**		
Clinician (i=391)	0.19	0.07 – 0.52	
Hospital (j=8)	0.17	0.05 – 0.59	

n = 4,732, died = 386 | Goodness of fit test: Hosmer-Lemeshow $\chi^2_8 = 13.12$, p = 0.1077, ROC AUC = 0.7535

6.4. External validity: can the score be systematically replicated in routine data?

The primary data used in the design of the score arose from a cluster randomised trial designed to mirror routine care as much as possible, but some characteristics of the trial were not typical of a routine care setting. Most notably, the use of the paediatric admission record (PAR) form – a central component of the intervention – probably increased the extent, content and detail of documentation of the process of clinical care. This could have made it comparatively easier to obtain case record data and calculate a process of care score. The provision of guidelines and work-aids along with supervision

and feedback could have also created an enabling environment for the completion of the processes represented by the indicators contributing to the score. Absolute scores at clinician and hospital level, and variations between and within these levels were likely affected by these factors, possibly influencing the observed characteristics of the score.

Systematic replication was therefore necessary to explore whether: (1) it was possible to calculate the score using data collected in non-intervention and routine quality assessment work, and (2) the score's domains and association with mortality observed in the trial data were generalisable to a non-trial setting. The pneumonia trial observation data and the Ministry of Health survey data, introduced and described in detail in section 2.3, are different from the Kenyan district hospitals study in a number of ways that make them suitable for external validation.

First, the pneumonia study focused on treatment of children with acute illnesses not randomised to either arm of a trial; there was therefore potentially no group (intervention) effect on the quality of care for these children. Secondly there was no survey effect in the pneumonia trial observational data because they were collected continuously over an 18-month period. This is in contrast with the staggered design of the district hospitals study where although data spanned the entire study period including the inter-survey periods, data collection followed 'waves' of supervision, instances of regular feedback and other substantive support activities designed to change the process of care.

Thirdly, although outcome assessment was undertaken at individual level in all three studies, the intervention was delivered at individual level in the pneumonia trial unlike the district hospitals study where the hospital (cluster) was the level of intervention. Any between-hospital differences in care in the observation study were therefore more likely to be due to hospital-level random variation, unlike the district hospitals study where there was likely an additional hospital-level effect due to the intervention. Lastly, observations were not linked to an identifier for the admitting clinician.

The Ministry of Health survey data were collected in a single cross-sectional study as part of a routine quality assessment exercise. Like the pneumonia trial observation dataset this survey involved neither group allocations nor multiple surveys. The hospital was the unit of interest in reporting of processes and outcomes, and the number of observations per hospital – between 16 and 55 – was comparatively smaller than those

in the pneumonia trial observation and district hospitals study. The data extraction and score calculation steps described in Chapter 5 were repeated on each of these two datasets. Next, a structural equation model to examine whether items contributing to the combined score aggregated in the proposed conceptual groupings were fitted on the replication data according to the path diagram in Figure 6.2-1. The results are presented in Table 6.4-1 and Table 6.4-2.

Table 6.4-1: Parameter estimates of the two-factor CFA model of the combined score for malaria, pneumonia and diarrhoea/dehydration in the Ministry of Health survey dataset

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.249)	_	_
λ_s	4.010 (0.998)	2.794	0.151
λ_d	1.000 (1.469)	_	_
λ_u	0.115 (0.169)	0.096	0.229
var(ε_a)	0.062	0.058	0.287
var(ε_t)	2.157	1.658	0.193
cov(ε_a,ε_t)	0.145 (0.398)	0.102	0.152

n = 802; fit indices: χ^2_5 = 182.375, p-value < 0.001; CFI = 1.000; TLI = 1.026; RMSEA < 0.001

Table 6.4-2: Parameter estimates of the two-factor CFA model of the combined score for malaria, pneumonia and diarrhoea/dehydration in the pneumonia trial observation dataset

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.331)	_	_
λ_s	1.606 (0.531)	0.334	<0.001
λ_d	1.000 (0.886)	_	_
λ_u	0.434 (0.384)	0.064	<0.001
var(ε_a)	0.109	0.037	0.003
var(ε_t)	0.784	0.117	<0.001
cov(ε_a,ε_t)	0.170 (0.581)	0.034	<0.001

n = 7,479; fit indices: χ^2_5 = 632.226, p-value < 0.001; CFI = 0.994; TLI = 0.968; RMSEA = 0.023

Validation on the Ministry of Health survey data showed a very good fit on CFI, TLI and RMSEA fit indices. However, none of the item loadings was significantly greater than zero. This may mean that variations in the items in these data is poorly linked to the underlying factors (domains), but could also reflect confounding by unidentified exogenous variables or item response patterns unique to the data. Validation on the pneumonia trial observation data also showed good fit on the same three fit indices and, unlike the former, significant factor loadings on all items. Correlations between the latent factors were estimated as 40% in the Ministry of Health survey data and 58% in the pneumonia trial observation data. These findings suggested that the observed

relationships between items and factors/domains were similar to those observed in the Kenyan district hospitals trial.

The final step was to investigate the score's association with mortality. The results are shown in Table 6.4-3. Unlike in the Kenyan district hospitals data, there was no evidence of a crude association of the combined score with mortality in each of the validation datasets.

Table 6.4-3: Proportions (%) and associations of exposure variables with mortality

Variable	Proportion of children who died (%)	Crude odds ratio	Trend odds ratio	n	p-value (trend p-value)
Combined score in the pneumonia observation data					
0	6.25	1.00			
1	5.32	0.84			
2	5.79	0.92			
3	6.88	1.11	0.95	7,058	0.062 (0.184)
4	4.36	0.68			
5	5.64	0.90			
6	3.85	0.60			
Combined score in the Ministry of Health survey data					
0	0.00	_			
1	1.89	1.00			
2	6.54	3.64			
3	5.15	2.82	0.96	783	0.149 (0.732)
4	9.28	5.32			
5	2.67	1.42			
6	5.49	1.81			

Nevertheless adjusted estimates association were obtained from logistic regression models analogous to those in section 6.3 appropriately modified to accommodate the differences between the datasets. For example, neither model included a group (intervention) or time (survey) variable, nor was there any adjustment for clustering within clinician. Both models were first attempted with adjustment for clustering within hospitals. Treating hospitals as random effects – an exact replication of the model of the district hospitals study data – and as fixed effects resulted in poorly fitting models even though the estimated association of the score with mortality remained consistent with previous findings (Appendix A.14). The best fitting models of the validation data did not include any adjustment for clustering. Estimates from these models are presented in Table 6.4-4 and Table 6.4-5. They are consistent with previous findings about the adjusted associations between each of the variables and death. The estimated reduction

in the odds of death per unit increase in the score was 22% in the pneumonia trial observation and 39% in the Ministry of Health survey, and the confidence intervals around these estimates overlapped with those from the district hospitals study. The greater uncertainty about the association in the Ministry of Health survey was likely attributable to the comparatively smaller size of the study.

Table 6.4-4: Adjusted effect of quality of care measured using the combined score on death estimated by a logistic regression model of the pneumonia trial observation data

	Adjusted odds ratio for death	95% confidence interval	p-value
Combined score	0.78	0.70 – 0.88	< 0.001
Age (years)	0.92	0.85 – 0.98	0.014
Sex (female vs. male)	1.05	0.84 – 1.32	0.653
Number of diseases diagnosed	0.83	0.58 – 1.18	0.302
Severity (3=highest, 2=intermediate, 1=lowest)	2.64	2.11 – 3.30	< 0.001
Identity of disease			
Diarrhoea/dehydration	1.00		
Malaria	0.39	0.23 – 0.66	< 0.001
Pneumonia	1.09	0.75 – 1.58	
Duration of admission (days)	1.00	0.99 – 1.01	0.785

n = 5,924, died = 324 | Goodness of fit test: Hosmer-Lemeshow $\chi^2_8 = 12.70$, p = 0.1226; ROC AUC = 0.6704

Table 6.4-5: Adjusted effect of quality of care measured using the combined score on death estimated by a logistic regression model of the Ministry of Health survey data

	Adjusted odds ratio for death	95% confidence interval	p-value
Combined score	0.61	0.42 – 0.88	0.008
Age (years)	0.64	0.43 – 0.96	0.033
Sex (female vs. male)	1.03	0.49 – 2.18	0.933
Number of diseases diagnosed	(omitted – varies very little with the outcome)		
Severity (3=highest, 2=intermediate, 1=lowest)	5.34	2.53 – 13.26	< 0.001
Identity of disease			
Diarrhoea/dehydration	1.00		
Malaria	0.18	0.03 – 1.11	0.162
Pneumonia	0.29	0.07 – 1.17	
Duration of admission (days)	0.94	0.86 – 1.04	0.210

n = 540, died = 32 | Goodness of fit test: Hosmer-Lemeshow $\chi^2_8 = 6.52$, p = 0.5890 | ROC AUC = 0.7689

More formally, similarity across estimates was tested by calculating a chi-squared statistic for heterogeneity based on how different each estimate was from the average relative to the variance of the estimate, as described in section 4.5.3 [Higgins *et al.* 2002, Thompson & Sharp 1999, Thompson 1994]. The statistic is a sum of the inverse variance-weighted differences between each study's estimate of effect and the overall pooled estimate of effect. For these three studies Q was estimated as 3.87 with 2 degrees

of freedom corresponding to a p-value of 0.144, meaning that there was no evidence of heterogeneity. The proportion of variability across studies attributable to heterogeneity, I^2, was 48.4%. This was a 'moderate' amount of variability which did not weaken the suggestion of a common effect especially in light of the consistency in magnitude and direction of association of the score with mortality [Higgins *et al.* 2003].

It was further assumed that the true effect of process of care – measured by the score – on mortality varied randomly across the studies. This was a fair assumption considering the possibility of confounding in the association between care and mortality caused by factors unidentifiable from the hospital record of the admission episode, such as socio-economic status of the patient or promptness of with which care was given, among others. Nevertheless this assumption did not change any conclusions about heterogeneity, estimates of effect, or strength of evidence for (or against) the hypotheses examined.

Figure 6.4-1 is a graphical comparison of the individual study effects along with the pooled effect which was estimated as a 20.1% reduction in the odds of death per unit increase in the score (95% confidence intervals 29.1% reduction to 9.9% reduction, p-value < 0.001). It shows very strong evidence of an association between the score and mortality which is consistent and of similar magnitude across the three datasets.

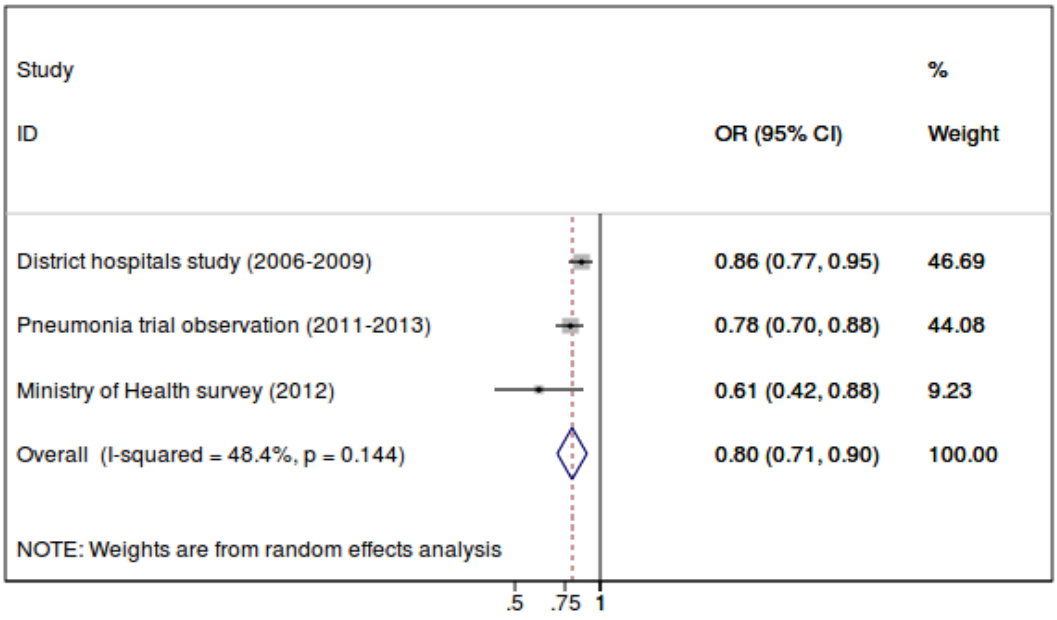


Figure 6.4-1: Study-specific and pooled estimates of the strength of association between the process of care score and mortality

6.5. Summary

There is evidence that items making up the score naturally aggregate into domains that are not only consistent with the clinical process encompassing assessment, diagnosis and treatment, but that are also measures of a common underlying construct. There is also strong evidence that the score is associated with mortality. Higher scores, representing better care, are associated with lower odds of mortality, and this suggests that the score as a measure of process of care is linked to an objective outcome, as expected. All these findings, first observed in data arising from a trial, have been replicated on external data from a prospective observation and those from a retrospective cross-sectional survey, albeit with lower-strength evidence of item groupings in one dataset. Nevertheless when put together with the results of initial validation in Chapter 5, there is sufficient evidence to conclude that the proposed score is a good measure of the process of care. Chapter 7 presents some practical applications of the score.

Chapter 7 – Applications of the Proposed Measure as a Process-based Outcome

7.1. Introduction

This Chapter demonstrates how the score could work in real life. The combined score is presented in the context of two common uses of quality measures. The first is an example of routine quality of care assessment and reporting similar to what would normally be undertaken in quality assurance and improvement programmes. In this example, alternative contemporary approaches to reporting summary measures are briefly discussed with the aim of selecting the most suitable one for reporting this score. Use of the score is demonstrated with data from the Ministry of Health survey which was the type of quality assessment exercise envisaged in the design of the proposed score. The second situation is an application of the score as a trial outcome measure for an intervention to improve quality of care. This is demonstrated through estimation of the effect of the intervention in the Kenyan district hospitals study using score as an outcome measure. An approach to the analysis is suggested, along with an appropriate sample size calculation for the study.

7.2. Example 1: reporting and monitoring quality of care

Once standards of process of care are in place, and an objective way of measuring whether care is provided according to these standards has been developed, then the next logical step is to work out an effective strategy for communicating the metric to its target audience. Three approaches to reporting metrics of quality of care for quality assurance and improvement and comparing hospitals are commonly encountered in literature, and these are cumulative sum (CUSUM) charts league tables and funnel plots.

CUSUM charts are plots of the cumulative number of successes or failures against the total number of patients, procedures or events [Williams *et al.* 1992]. They have been used to monitor clinical outcomes in routine clinical data such as low APGAR scores [Sibanda & Sibanda 2007], binary surgical outcomes [Steiner *et. al* 2001], assessment of organ transplant failure rates [Biswas & Kalbfleisch 2008], and even competency in conducting medical procedures [Sivaprakasam & Purva 2010, Sims *et al.* 2013]. They are designed for use with binary outcomes or measures that can be dichotomised, and for this reason are not applicable to the score.

League tables are listings, tabulations or graphical presentations of the institutions against their value of a measure of an attribute of interest. They are popular because they are easy to interpret and understand. They are often used when a standard against which performance on the measure is to be compared is not set [Noyez 2009]. They assume that the institutions being compared are inherently similar, and any variations between them are due to true differences in the attribute being measured. A league table comparing the quality of care in the 22 hospitals in the Ministry of Health survey of 2012 as measured by the combined score is presented in Figure 7.2-1.

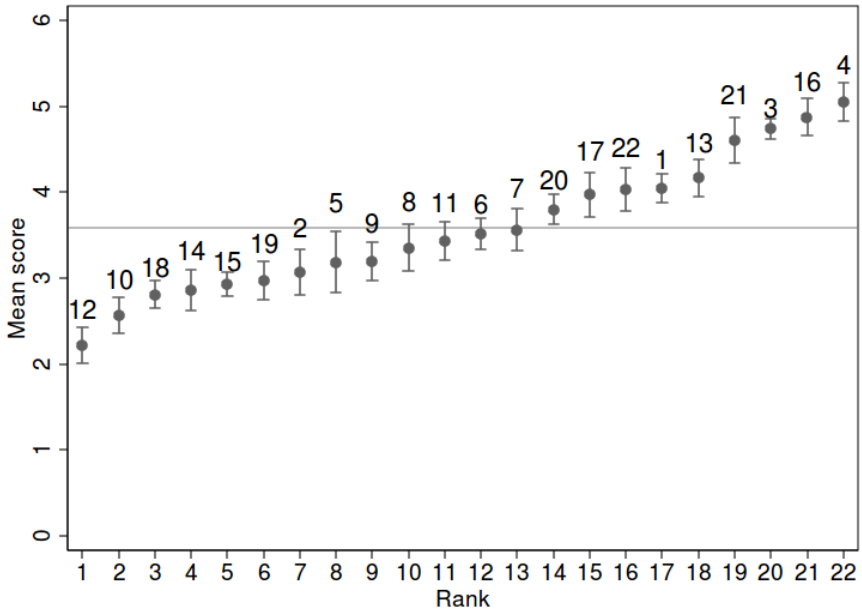


Figure 7.2-1: League table of the 22 hospitals in the Ministry of Health survey of 2012
The dots and capped vertical lines represent the hospital mean combined scores and their 95% confidence intervals respectively (numbers are hospital IDs), and the horizontal line is the overall mean for all hospitals.

League tables are suitable for identification of institutions in need of improvement. For example, hospitals 2, 5, 9, 10, 12, 14, 15, 18 and 19 in Figure 7.2-1 exhibit poor quality of care compared to their peers because they have mean scores below the overall mean and their 95% confidence intervals do not include the overall mean. However, the ranks implied by league tables are themselves sensitive to sampling variation just like the mean scores. This has been a source of great controversy and scepticism [Pandey *et al.* 2007, Jacobs *et al.* 2005, Gibberd *et al.* 2004, Botha *et al.* 2001] not least because the variability in ranking is not properly accounted for [Goldstein & Speigelhalter 1996].

Funnel plots, unlike CUSUM plots, can be applied to binary, categorical and continuous outcomes, and unlike league tables, they handle various sources of variability satisfactorily and do not explicitly rank institutions [Benneyan *et al.* 2003, Woodall & Montgomery 1999, Wheeler & Chambers 1992]. The principles underlying funnel plots are derived from statistical process control of product variation in manufacturing. Two sources of variation in a measurement are identified. One is chance which manifests itself through sampling variation. It is referred to as common cause variation and is expected in all measurement. The other – arguably the most important in institutional comparison but which league tables do not explicitly distinguish – is systematic differences, referred to in statistical process control terminology as special cause variation. Special cause variation can be attributed to an assignable cause such as a true improvement or deterioration, and if an assignable cause is not patent then special cause variation warrants further investigation.

Funnel plots have a series of points representing the units being compared, with a line indicating their mean on one axis. These are bound by contour lines representing control limits beyond which deviations in the metric are attributed to special cause variation. Precision is measured by the denominators of the metric and are plotted on the second axis [Spiegelhalter 2005, Spiegelhalter 2002]. For example, the measure of precision of the hospital mean quality of care scores is the number of case records sampled to estimate the mean score. Typically the control limits are set at 2 and 3 standard deviations above and below the overall quantity means. They are analogous to the 95% and 99.5% confidence intervals in the league tables, but are calculated differently: whereas the confidence intervals in the league tables are estimated around each hospital's mean score, those in the funnel plots are based on the overall mean of all the hospitals. The assumption behind this approach is that all hospitals are performing at the same level, and the funnel plot simply aims to establish a range of values within which their performance could still be characterised as 'average' within the limits of chance. It is a formal test of the null hypothesis that each of the hospital mean scores is not different from the overall mean score. If the mean of n scores in hospital i is $\bar{x}_i$ and the mean score of all hospitals is μ with standard deviation σ, then:

$$\frac{\bar{x}_i - \mu}{\sigma / \sqrt{n}}$$

$\frac{\bar{x}_i - \mu}{\sigma/\sqrt{n}}$ is assumed to be normally distributed. The control limits are a function of *n* and are set at $\mu \pm z_\alpha \frac{\sigma}{\sqrt{n}}$ where z_α is the cumulative normal distribution at level α. Figure 7.2-2 is a funnel plot of same 22 hospitals in Figure 7.2-1.

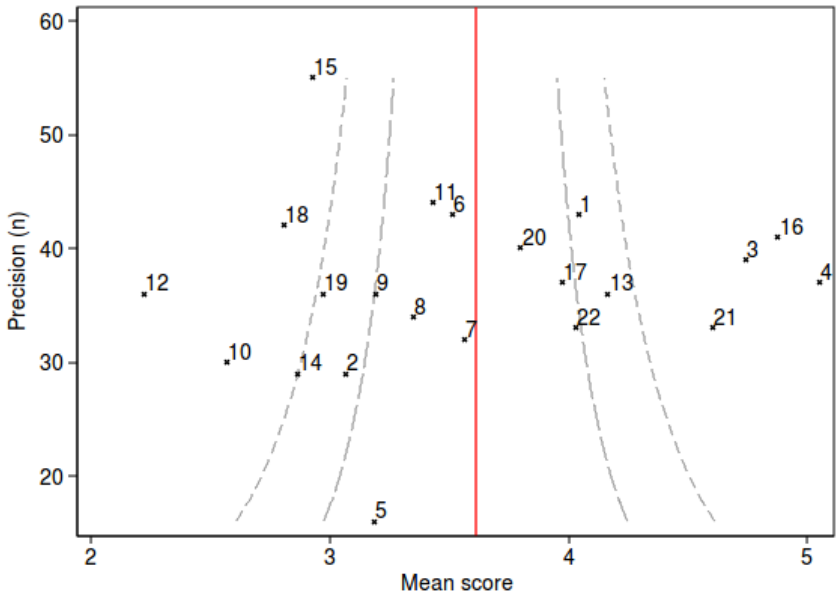


Figure 7.2-2: Funnel plot of the 22 hospitals in the Ministry of Health survey of 2012
Dots represent the hospital mean scores (numbers are the hospital IDs), the vertical line is the overall mean for all hospitals, and the dotted contour lines represent the range of values 2 (inner) and 3 (outer) standard deviations around the mean

A larger number of hospital mean scores are compatible with the overall mean compared to the league table. The funnel plot suggests that 7 hospitals have mean scores which are less than the lower 95% confidence limit of overall mean of about 3.6, another 6 greater than the upper 95% confidence limit, and 9 are within the 95% confidence limits of the overall mean. This is in contrast to the league table where only hospitals 6, 7, 8 and 11 are not different from the overall mean while all other hospitals have mean scores higher or lower than the overall mean.

Although the score is in fact an ordinal outcome, the assumption that its mean is asymptotically normally distributed underlies the construction of the control limits. This assumption may be problematic for means close to the lower and upper bounds of the range of the score since it may lead to control limits falling outside these bounds. To avoid this anomaly the items of the score are assumed to be a series of independent but not necessarily identically distributed Bernoulli experiments. The sum of successes of *n*

binary process of care indicators each with a distinct probability of success, p_{ij} in each hospital j has a Poisson-binomial distribution. The corresponding hospital mean scores, μ_j, are given by:

$$\mu_j = \sum_{i=1}^{n} p_{ij}$$

and are equivalent to the arithmetic hospital means of a normally distributed score. Hospital standard deviations, σ_j, are given by:

$$\sigma_j = \sqrt{\sum_{i=1}^{n} p_{ij}(1 - p_{ij})}$$

The standard deviation is equal to that of a normally distributed score at its median; it reaches a maximum when each of the Bernoulli experiments – the indicator items – has a probability of 50%, but tends towards zero for values greater or less than the median Figure 7.2-3, which is a funnel plot of the 22 hospitals in Figure 7.2-2, shows the narrowed control limits resulting from this assumption about the score distributions.

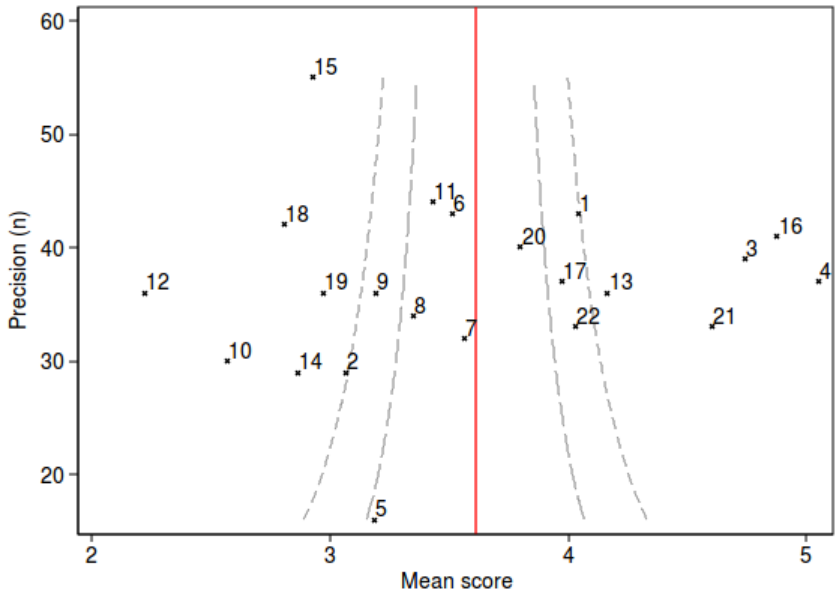


Figure 7.2-3: Funnel plot of the 22 hospitals in the Ministry of Health survey of 2012 with Poisson-binomial based 95% and 99.5% confidence bounds ('control limits')

A grading system can then be applied to report these mean scores (Table 7.2-1). Grading provides standardised descriptors of each hospital's performance relative to the group: hospitals with higher scores will on average have better grades, and vice-versa,

even when there's a change in the mean score of all hospitals. The grades may be reported alongside the absolute scores to communicate both absolute and relative changes in quality of care.

Table 7.2-1: Suggested 5-grade system for interpreting the funnel plots

Grade	Range of mean scores	Hospitals in this grade
A	Equal to or greater than the upper 99.5% confidence limit	3, 4, 13, 16, 21
B	Equal to or greater than the upper 95% confidence limit but less than the upper 99.5% limit	1, 17, 22
C	Equal to or greater than the lower 95% confidence limit but less than the upper 95% limit	5, 6, 7, 8, 11, 20
D	Equal to or greater than the lower 99.5% confidence limit but less than the lower 95% limit	9
E	Less than lower 99.5% confidence limit	2, 10, 12, 14, 15, 18, 19

Grades B and D refer to the region between the 95% and 99.5% confidence limits, which is small compared to the rest of the funnel-plot, and likely to include few hospitals especially when there is little variability between them. For example, although there is a wide variability of hospital mean scores in Figure 7.2-3, only 4 out of 22 hospitals fall in this region. A 3-grade system which considers 'average' performance to be within either 95% or 99.5% confidence limits – thus excluding this narrow range of performance – may be preferred (Table 7.2-2).

Table 7.2-2: Suggested 3-grade system with 'average' performance set at mean scores within the overall 95% confidence limits

Grade	Range of mean scores	Hospitals in this grade
A ('above average')	Equal to or greater than the upper 95% confidence limit	1, 3, 4, 13, 16, 17, 21, 22
B ('average')	Equal to or greater than the lower 95% confidence limit but less than the upper 95% limit	5, 6, 7, 8, 11, 20
C ('below average')	Less than the lower 95% confidence limit	2, 9, 10, 12, 14, 15, 18, 19

Other than comparing hospitals, funnel plots can be used to compare clinicians within a hospital, quality of care across a variety of diseases using disease-specific scores, or even showing how quality compares to a pre-set level of performance. Further examples of their uses are presented in Appendix A.15.

7.3. Example 2: estimating the effect of a quality improvement intervention

The proposed score can be used as a trial outcome measure in studies estimating the effect of quality improvement interventions. Group mean scores may be compared using parametric and non-parametric tests or regression models when adjustment for covariates is required. For example, in the Kenyan district hospitals study where a

quality improvement intervention was delivered to hospitals over an 18-month period, the effectiveness of the intervention could be assessed as follows.

The trial outcome measure, Y_i, is the score of process of care for patient i. The patient's group allocation is x_1 such that x_1=0 for a patient in the control group and x_1=1 for a patient in the intervention group. The surveys are represented by x_2 (x_2=0 at baseline, x_2=1 at first follow-up, x_2=2 at second follow-up and x_2=3 at the primary end-point), and x_3 is the interaction of group and survey. A generalised linear regression model for Y is:

$$g(Y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

where g is the link function for the outcome. The intervention effect on the score at baseline is the difference between the expressions of the outcome when x_1=1 and x_1=0 respectively, with x_2 held at 0:

$$\{\alpha + \beta_1 + \beta_3 x_{3i} + e_i\} - \{\alpha + \beta_3 x_{3i} + e_i\}$$

which is β_1. The score at any time T is predicted by:

$$g(Y_i) = \alpha + \beta_1 x_{1i} + \beta_2 T + \beta_3 T\, x_{1i} + e_i$$

which when rearranged to:

$$g(Y_i) = \alpha + x_{1i}(\beta_1 + \beta_3 T) + \beta_2 T + e_i$$

shows that the intervention effect on the score is the slope $\beta_1 + \beta_3 T$. The baseline-adjusted intervention effect is the difference between the intervention effect at baseline and that at any other time point:

$$\{\beta_1 + \beta_3 T\} - \{\beta_1\} = \beta_3 T$$

Therefore the intervention effect at time T is the coefficient of the interaction term for the intervention and time.

Table 7.3-1 shows the intervention effect on the score estimated by two generalised linear models both with adjustment for clustering within hospital: a linear model which assumes that the score is continuous and normally distributed as suggested by the histograms in Figure 5.2-7 and Figure 5.2-8, and an ordered logistic regression model which acknowledges the ordered categorical nature of the score and assumes

proportionality of odds between levels of the score. The model results were very similar. There was no evidence of a difference in scores between the groups at baseline. Scores in the control group at the first and second follow-up surveys and at the main end-point were 1.57, 1.31 and 1.09 units respectively greater than baseline scores, with a trend of decline over time (χ^2 test for trend p-value <0.001). In the ordered logistic model the odds ratio for a unit increase in control group scores followed the same trend: the odds of a unit increase in the score at the three time points were 9.32, 6.67 and 4.58 times those at baseline. Scores in the intervention group were progressively higher in successive surveys according to the linear model; the score was on average 0.96, 1.33 and 1.55 units higher at the second, third and fourth surveys respectively than baseline. Similarly in the ordered logistic model the odds ratio for a unit increase in the score were 3.90, 6.07 and 8.85 times higher than baseline in each of the three successive surveys. The largest effect was observed in the study's main end-point, with strong evidence of a linear trend (χ^2 test for trend p-value <0.001).

Table 7.3-1: Effect of the intervention on the process of care score in the Kenyan district hospitals study

	Linear model (n=8,453)			Ordered logistic model (n=8,476)		
Effect	**coef.**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Group						
Control	1.00			1.00		
Intervention	0.10	-0.26 – 0.47	0.591	0.90	0.50 – 1.63	0.724
Survey						
Baseline	1.00			1.00		
1st follow-up	1.57	1.46 – 1.68	<0.001	9.32	5.77 – 15.05	<0.001
2nd follow-up	1.31	1.20 – 1.43		6.67	3.15 – 14.14	
End-point	1.09	0.98 – 1.20		4.58	1.32 – 15.83	
Group-survey interaction						
Intervention x 1st follow-up	0.96	0.80 – 1.11		3.90	2.11 – 7.22	
Intervention x 2nd follow-up	1.33	1.17 – 1.48	<0.001	6.07	2.80 – 13.12	<0.001
Intervention x End-point	1.55	1.40 – 1.69		8.85	2.78 – 28.18	

An important consideration for the use of the score as a trial outcome is a sample size calculation. Unlike indicator outcomes where sample size calculations and effect sizes are based on an estimated or desired proportion of successes, sample size calculations when the score is the main outcome can be set as the change in the mean number of process of care tasks undertaken. Trials of this nature will almost invariably be of a cluster design due to the effect of higher-level clustering variables, such as the clinician or hospital, on the outcome. Assuming equal-sized clusters with *m* observations in each

cluster, the minimum required number of clusters per arm, n, for a two-arm trial to detect a difference in mean scores, d, between the arms is:

$$n = \frac{2(z_{\frac{\alpha}{2}} + z_{\beta})^2 (\sigma^2)\{1 + (m-1)\rho\}}{d^2 m}$$

where α and β are the required level of significance and 1-power respectively, $z_{\alpha/2}$ and z_{β} are the cumulative normal distribution corresponding to $\alpha/2$ and β respectively, and σ^2 is the overall variance [Hayes & Bennett 1999]. The intraclass correlation coefficient ρ is estimated from the between (σ^2_b) and within (σ^2_w) cluster variances of previous similar studies as $\sigma^2_b/\{\sigma^2_b + \sigma^2_w\}$. Poisson-binomial variances described in section 7.2 may be used if success probabilities of each of the indicator items contributing to the score are known. If the clusters are of different sizes, as is wont to be in many studies, then the sample size is adjusted by a coefficient of variation, c, which is the ratio of the standard deviation of cluster sizes to the mean of cluster size, ~~m~~ [Eldridge *et al.* 2006, Kang *et al.* 2003]. The required number of clusters per arm of uneven sizes with a known coefficient of variation is:

$$n = \frac{2(z_{\frac{\alpha}{2}} + z_{\beta})^2 (\sigma^2)\{1 + (\bar{m}[1 + c^2] - 1)\rho\}}{d^2 \bar{m}}$$

One likely difficulty with this adjustment is that very few studies report intraclass correlation coefficients and coefficients of variation of their main measures of interest [Rutterford *et al.* 2011]. It is also not unusual to find a difference between values used in sample size calculations prior to the study and those calculated with actual data from the study [Eldridge *et al.* 2006]. Figure 7.3-1 is a simulation of sample sizes (number of clusters per arm, n) across a range of between-arm mean score differences, d, assuming clusters of different mean sizes of 20 and 200, coefficients of variation of 0, 0.5 and 1, and ρ of 0.2 (low clustering) and 0.8 (high clustering). In each simulation α is 0.05, β is 90%, and σ^2 is assumed to be 1.5 which is the Poisson-binomial variance for six binary indicators each with a 50% probability. It shows that larger samples are required for smaller between group differences in the score (smaller d), higher clustering (larger ρ), smaller average cluster sizes (smaller m) and higher variability between cluster sizes (larger c) when all else are held constant. For example, a sample of 8 clusters per arm is required to demonstrate a difference of 1 between the groups if clusters have an average

size of 200, coefficient of variation of 0.5 and intraclass correlation of 0.2, but this shoots up to 32 for an intraclass correlation of 0.8, and to 50 with a coefficient of variation of 1 and an intraclass correlation of 0.8. For any combination of c, m and ρ, the n quadruples for any 50% reduction in d.

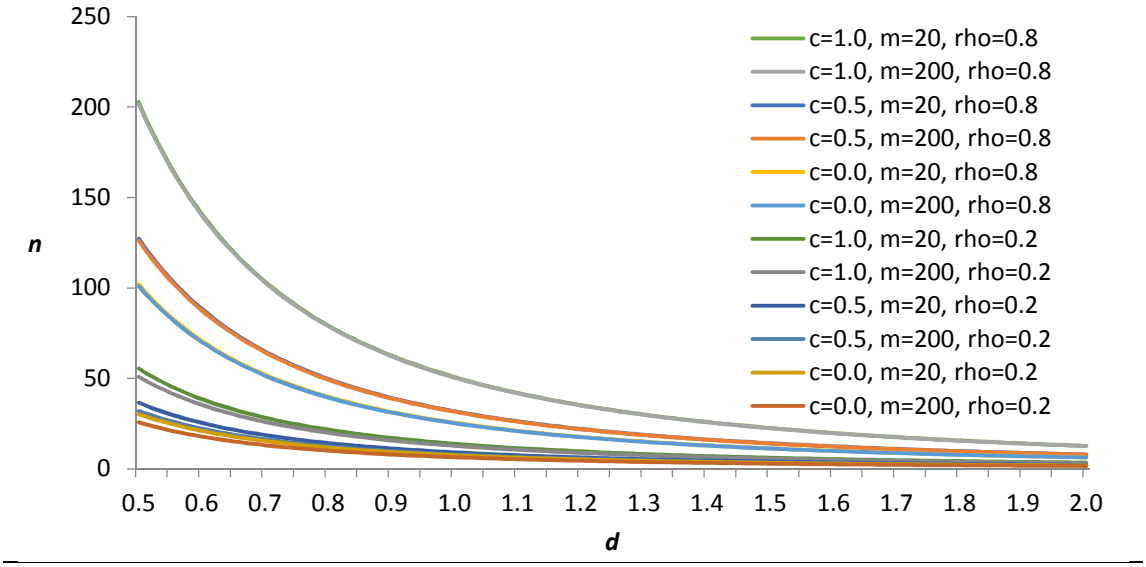


Figure 7.3-1: Sample size calculations for a range of between-group score differences, average cluster sizes, coefficients of variations and intraclass correlation coefficients
For ease of presentation the legends are listed in the same vertical order of the lines they represent on the graph

7.4. Summary

A variety of tools for summarising and presenting the score are available. Funnel plots are the preferred method of presenting hospital, clinician or other cluster-level mean scores because they appropriately deal with the various sources of variability in measurement. A qualitative interpretation of cluster mean scores below, around and above the overall mean provides a promising way of grading scores, an additional approach to reporting them.

The utility of the score as a trial outcome has also been demonstrated in two regression models of data from the Kenyan district hospitals trial. The models reflect different assumptions about the distribution of the score. They have been compared to explore the sensitivity of conclusions on quality of care to non-normality through the modelling of the score as a discrete, ordered categorical outcome. Both models have led to conclusions which are consistent with each other, and with traditional outcomes such as single indicators which have previously been used to analyse these data.

Chapter 8 – Summary and Conclusion

The aim of this work was to construct and validate a measure of the quality of care for children with acute illnesses admitted to hospitals in a low income country. A review of the literature highlighted a lack of quality of care measures relevant for routine use in this setting. This can be partly attributed to a lack of well-established reference standards and guidelines on which to base quality measurements and, even when guidelines have been available, a lack of effective dissemination and use in setting minimum standards of care [English *et al.* 2004, Nolan *et al.* 2001]. There are however several examples of studies illustrating the development of measures of quality of care for a variety of illnesses, mostly in high income countries. These examples provided guidance on the development of the measure of quality of care. The process began with a selection of a perspective of measurement (Chapter 1), followed by data selection (Chapter 2) and an exploration of approaches to analysis (Chapter 3 and Chapter 4), reliability and validity testing of the measure (Chapter 5), and culminated in exploration of it links to related measures and a brief presentation of examples of its application (Chapter 6 and Chapter 7). This was thought to be an objective and transparent approach to setting up and validating the measure.

The Donabedian framework which considers the overall quality of health care in terms of resources that support health service delivery ('structure'), the appropriateness of what is actually done by the providers of care ('process') and the results of care ('outcomes') was adopted in conceptualising quality of care and selecting a suitable perspective for its measurement [Donabedian 1988]. While each of these aspects is important in its own right, there is evidence that failure to translate evidence of good care into practice – weaknesses in the process of care according to Donabedian's framework – is a limiting factor to the achievement of good outcomes in many low income settings, including Kenya [Peabody *et al.* 2006]. For this reason this thesis adopted process of care as the perspective of measurement, with initial focus on children admitted with malaria, pneumonia and diarrhoea since these diseases are responsible for the majority of under-5 deaths in this setting [Liu *et al.* 2012, Black *et al.* 2010]. Measuring quality of care in terms of process by contrasting what was documented to have been undertaken in the process of care – or the lack of documentation of process, an indicator of poor practice – with guideline recommendations is admittedly a narrow perspective for measuring what is obviously a

multi-faceted concept; it is nevertheless a very important one considering the central role of clinical processes in providing the means by which health inputs are converted to desirable outcomes. Ideally, this approach should be applicable to quality of care measurement in other diseases as well as settings beyond what has been presented in this thesis, and also allow for direct comparison of quality of care across diseases.

Data for the initial development of the measure came from the case records of 12,036 children admitted with acute illnesses to 8 Kenyan district hospitals during a cluster-randomised trial [Ayieko *et al.* 2011, Gathara *et al.* 2011, English *et al.* 2009]. In Kenya, what constitutes good care is defined by clinical practice guidelines for paediatric care [MoH 2010, MoH 2007, MoH 2006]. Items representing specific recommendations on how to provide care were identified from these evidence-based guidelines. These items were then summarised into a basic score which was a count of the number of recommended tasks undertaken by the admitting clinician during each child's treatment. The use of score for quality of care measurement is not without precedence since a number of other measures of health related constructs, such as quality of life measures [Burckhardt & Anderson 2003, Hunt *et al.* 1985, Gilson *et al.* 1975] and prognostic scores in critical care [Knaus *et al.* 1985, Teasdale & Jennet 1974] take a similar approach to measurement.

This patient-level score focused on processes of care in the first 48 hours of an admission episode, a critical period for treatment of acute illnesses when there is sufficient opportunity to intervene and restore health [Couto *et al.* 2013, Adeboye *et al.* 2010, Campbell *et al.* 2004, Berkley *et al.* 2003, Sodemann *et al.* 1997, Commey *et al.* 1994]. Patient-level measurement provides the flexibility to adjust for characteristics such as age, sex, severity of illness, co-morbidity in statistical models, while allowing for aggregation at higher levels such as clinician, department and hospital level, and adjustment for factors at each of those levels that may affect quality of care.

Items making up the basic score were very specific to each of the three diseases. There was also redundancy between several items and lack of direct comparability of disease-specific scores resulting from differences in their scale ranges. Modifications to the basic score items produced generic items which eliminated redundancy by grouping items representing similar tasks. These generic items allowed the same scoring approach to be applied to all three diseases, and for this reason the modified score was similar in

scale and directly comparable across all three diseases. It also allowed the aggregation of scores for multiple treatment needs into a unified score – the combined score. This approach can be extended to other diseases by identifying equivalent domains of clinical process, since the principle underlying the three domains of clinical process is almost universal in medicine. The absence of a hypothetical fourth domain of items relating to diagnostic testing is a weakness worth noting. Perhaps this reflects on the low-resource nature of this setting where the use of such technologies is not widespread or emphasised as important for the delivery of care [Mabey *et al.* 2004]. Future work could explore adding this domain in settings where such elements of process are clearly part of the standards of admission care.

Structural equation modelling of the disease-specific modified scores and the combined score provided evidence that this conceptual aggregation of score items into domains was supported by observed patterns of correlations between items, and was therefore a valid representation of the construct being measured. It is possible that better fitting factor models could have been obtained through changes to the proposed groupings of items and relationships between domains. However, this was not the aim of the analysis, especially considering the known difficulty in fitting models with only two items per domain, and the sensitivity of model fit to unique patterns across datasets [Costello & Osborne 2005, Raubenheimer 2004]. Instead the focus was on establishing whether the proposed model was theoretically and practically sound while exhibiting a statistically acceptable fit to the data – these key aims of the analysis were met.

Association of the score with mortality was explored to test the theory that process of care should be associated with outcomes. Strong evidence that higher scores are associated with lower odds of mortality was found. This confirms that the score works well as a measure of quality of care, since better care – represented by higher process of care scores – should be associated with lower odds of mortality. It also strengthens the face-validity of the score, and will likely reassure policymakers and practitioners. However, as quality of care improves over time – resulting in less variability in scores – this measure is likely to start to fail to predict mortality. In the same way it may not be possible to link such scores with outcomes in high income settings.

The score was easily reproduced with case record data from a longitudinal observational study and a cross-sectional survey of hospitals which were broadly representative of the

varied characteristics of hospitals in a low income country. Successful replication of the score was important because it demonstrated the broad applicability of the score across a variety of circumstances. Key characteristics of the score were consistent with those observed during its development on data from the district hospitals study. Association of the score with mortality was similar in magnitude and direction to that observed in the trial data, sufficiently to assume a common effect across the three datasets examined, and possibly more generally. In addition, fit indices in structural equation model of the main data and both validation datasets confirmed that the proposed aggregation of items into domains was a good fit to the data. Although other domains may exist which could better explain item aggregation into groups that are different from those proposed in this thesis – as rigorous testing of the structure of many measures often reveals [Kupeli *et al.* 2013, van Prooijen & van der Kloot 2001] – all of the evidence established in this work supports the conclusion that the items in this score are good measures of the proposed domains. These include the well-fitting disease-specific and combined score models in two of the datasets, and a sound theoretical basis of this grouping of items.

A graphical and a semi-quantitative method for effective communication of the score to its target audiences were suggested. Funnel plots for graphing score summaries at clinician, department, hospital or group level (or any other desired level of clustering) allow for effective handling of the various sources of variability within and between groups [Benneyan *et al.* 2003, Woodall & Montgomery 1999]. They are good for showing such groups summaries and the differences between them, including differences that are unlikely to arise by chance alone. Grading of scores provides an additional layer of reporting which enables the descriptive presentation of group scores relative to other groups. Application of the score as a trial outcome measure was also demonstrated by estimating effect of the quality improvement intervention in the Kenyan district hospitals study. The observed effect of the intervention was consistent with the findings of previous analyses, showing the suitability of the score as a trial outcome and its efficiency in terms of the numbers required to test an intervention [Ayieko *et al.* 2011]. For this purpose a sample size calculation for clustered group comparison has been suggested which addresses some important considerations.

A number of statistical approaches and considerations underlie the score design process, including the identification of a suitable level of measurement, factor analysis to explore

the aggregation of items into domains, approaches to comparing agreement between measures and generalised linear modelling to estimate the association between the score and other variables. Some challenges were encountered in the course of applying these methods. In factor analysis, it is often assumed that the items thought to measure the underlying construct are continuous and normally distributed. This was not the case in this work. Items were binary, and a modification to the factor analysis involving the use of a polychoric correlation matrix was necessary for the estimation of factors. Difficulties were also encountered in interpreting factor analysis fit indices because recommendations in the literature vary widely. An approach examining the consistency of evidence for or against model fit from multiple indices and different datasets was adopted. It relied on qualitative interpretations of ranges of multiple fit indices rather than fixed thresholds which, according to literature, were least affected adversely by model peculiarities such as sample size and model complexity.

In measuring agreement between candidate scores it was necessary to consider the likely effect of clustering on scores. Specifically, scores of children attended to by the same clinician or in the same hospital were likely to be more similar than those not, thereby creating a false impression of higher agreement. To work around this problem an R^2 statistic was estimated at both clinician and hospital level to measure agreement between candidate scores in terms of the amount of variability in one score explained by the other. A theoretical distribution of this statistic was unknown. It was therefore not possible to derive estimates of uncertainty around these measures of agreement, and for this reason bootstrap 95% confidence intervals were calculated.

Estimates of R^2 were obtained by modelling the score as both a continuous and an ordinal measure. This involved the use of GLMs which assumed a Gaussian and an ordered logit distribution respectively of the dependent variable. These were plausible assumptions, since the score was observed to be normally distributed, but also known to be an ordered categorical measure. A Poisson-binomial distribution was an alternative plausible approach to modelling the score, since the score could have also been considered to be the number of successes in a series of independent binary items. Poisson-binomial means and standard errors were manually calculated when exploring and graphing variations in the score across hospitals and over time because none of the major statistical analysis software provided estimation routines using this approach. This is a gap for future software development.

In conclusion, this work shows how a score of process of care can be set up, used and presented in quality of care measurement. The initial focus has been on quality of care for three acute illnesses, and future work will be required to explore whether the methods that have been developed are suitable in other clinical settings, including non-paediatric care, non-medical care and non-communicable diseases. This work also demonstrates the link between process and an objective outcome of care, and shows the relevance of clinical processes in quality of care assessment.

Appendices

A.1. Search terms in the literature review

No.	Search strings	Number of hits
1	"design" OR "develop" OR "build" OR "theory" OR "construct" OR "create"	1,384,314
2	"Quality of Health Care"[Mesh]	4,230,882
3	"Process Assessment (Health Care)"[Mesh]	2,578
4	"score" OR "scale" OR "index" OR "measure" OR ("composite" AND "indicator")	1,218,431
5	"child, preschool"[Mesh] OR "Infant, Newborn"[Mesh]	1,025,427
6	#1 AND #2 AND #3 AND #4 AND #5	9
7	#1 AND #2 AND #3 AND #4	138
8	"Outcome Assessment (Health Care)"[Mesh]	565,146
9	"Inpatients"[Mesh]	10,569
10	#3 OR #8	565, 306
11	#1 AND #9 AND #10	277

A.2. Explanations of suggested steps for constructing a composite measure

Proposed step	**Explanation**
Perspective of quality measurement	Whether the measure is a structure, process or outcome measure or a combination
Data selection, sample size	Source data for the development of the measure including considerations for sampling and sample size calculation
Handling of missing data	Discussion on whether missing data is encountered, mechanism of missingness and solutions e.g. imputation, complete-case analysis
Multivariate analysis	Statistical methods to study the data structure and dimensions, to identify grouping scales or domains
Weighting, aggregation, generation of a summary measure	Identify weighting and aggregation procedures relevant to the data properties and clinical application
Assessment of robustness and sensitivity analysis	Exploring changes to the measure under different scenarios e.g. methods of aggregation, handling of missing data, summary generation
Reliability and validity	Using a validation sample or new data to test reliability and validity of the measure
Links to other indicators and measures	Comparing performance of new measure with existing ones

A.3. Summary of articles included in the review

Authors	Title	Description

Saloojee *et al.* 2009	Development of a measure of family-centred care for resource-poor South African settings: the experience of using a modified version of the MPOC-20	• objective: to adapt the Measure of Process of Care (MPOC) and develop a measure of family-centred care in a disadvantaged South African setting
		• setting: all public hospitals in two SA provinces with established rehabilitation services for children with cerebral palsy
		• population: caregivers of children aged 1-18 diagnosed with cerebral palsy living in poorly resourced areas in two provinces in South Africa
		• sampling: sample size based on Nunnally's (1978) suggested criterion for validating a scale (a minimum of 10 participants for each scale item). Convenience sample of 267 used.
		• statistical methods:
		  • exploratory factor analysis
		  • reliability: internal consistency (Cronbach's alpha >0.7 based on Nunnally's recommendation, very high = item redundancy Streiner & Norman 2003) and test-retest reliability (re-interviewing respondents by same interviewer then calculate ICC, 10% sample required)
		  • validity: convergent and discriminant validity using multi-trait scaling (Stewart *et al.* 1988); Kaiser-Meyer-Olkin measure of item sampling adequacy and content validity (Kaiser 1974), Bartlett's test of sphericity to ascertain whether factor analysis was appropriate (Sitzia 1999); concurrent validity by correlating scale scores with other variables it was hypothesised to be related to using Pearson correlation coefficients
		  • PCA to examine dimensions of the MPOC-SA
		• findings: modified MPOC did not work in new setting although underlying constructs were similar; a further revised 8-item version with 2 factors appears useful
		• key refs: King et al. 1995, 1996; King et al. 2004b (original MPOC); Rosenbaum et al. 1990; Ketelaar et al. 1998; Streiner & Norman 2003 (psychometric properties). Also lookup the Gross Motor Function Classification System (GMFCS) (Palisano et al. 1997)

Klassen *et al.* 2009	Evaluating family centred service in paediatric oncology with the measure of process of care (MPOC-20)	• objective: to evaluate the psychometric properties of the 20-item MPOC in parents of children undergoing treatment for cancer at 5 paediatric oncology centres in Canada • setting: 5 Canadian paediatric oncology centres • population: parents of children with cancer • sampling: convenient sample of 513 parents whose

Authors	Title	Description

		children were on active treatment at least 2 months post-diagnosis not on palliative care. 412 of 501 who agreed completed questionnaire (80% response rate)
		• statistical methods:
		  • exploratory factor analysis to determine how to group the questions (ML-FA without rotation, scree test to determine number of factors)
		  • score generated by summation
		  • reliability: internal consistency – Cronbach's alpha (random error in score due to intercorrelations among items); item scale correlations (Nunnally & Bernstein 1994)
		  • validity: within-scale construct validity – item-own scale correlations corrected for overall exceeding item-other scale correlations by at least 2 SDs; scale inter-correlation r=0.3-0.7 Cano *et al.* 2004
		  • ability of scale to differentiate between groups known to differ (King *et al.* 2004b)
		• findings: EFA yielded 2 factors; little missing data; scale scores spanned entire scale range; no notable floor or ceiling effects nor skewing; mean scale score nearer to high end than midpoint; scales exceeded criteria for internal consistency reliability; intercorrelations between scales showed related constructs; Spearman's rho correlations between the scales ranged 0.72-0.85 suggesting one concept was being measured
		• key refs: King *et al.* 1996; King *et al.* 2004b; Costello & Osborne 2005 (EFA); Scientific Advisory Committee of the Medical Outcomes Trust 2002; Streiner & Norman 2003; Cano *et al.* 2004 (psychometric tests and criteria – e.g. criterion for scale-level data quality is less than 5% missing, mean is near midpoint, floor according to Cano *et al.* 2004)

Siebes *et al.*
2008
	A family-specific use of the Measure of Process of Care for Service Providers (MPOC-SP)	• objective: to examine the validity and utility of the MPOC-SP as a family specific measure
		• setting: paediatric rehab in the Netherlands
		• population: professionals providing rehabilitation and educational services to children with cerebral palsy
		• sample size: 5-10 subjects per variable/item
		  • statistical methods:
		  • to explore appropriateness of parametric statistics: skewness and kurtosis
		  • reliability:
		  • validity: construct validity – Pearson's product-moment correlation coefficient between scales and comparable multidimensional structure on FA (general vs. family-specific MPOC-SP); ability of providers to discriminate between scales explored

Authors	Title	Description
		by absolute difference scores • findings: construct validity satisfactory (providers can distinguish family-specific and general ratings), Pearson's correlations not significant. • key refs: King *et al.* 2004b; King *et al.* 1996; Nunnally & Wilson 1975; Streiner & Norman OUP 1995

Wierenga *et al.* 2011	Quality indicators for in-hospital pharmaceutical care of Dutch elderly patients: development and validation of an ACOVE-based quality indicator set	• objective: develop and validate a set of explicitly-phrased QIs to measure efficiently the quality of pharmaceutical care of elderly hospitalized patients in the Netherlands by selecting and adapting ACOVE criteria
		• setting: a tertiary university hospital in the Netherlands
		• population: elderly hospitalized patients
		• sampling: NR
		• methods:
		  • expert opinion used to select QIs using a Delphi technique
		  • selected QIs tested against records of 10 preselected patients experiencing long-term hospital stay
		  • inter-rater reliability: using 3 raters' assessments of 10 randomly-selected patients using Fleiss κ values if all 3 raters or ICC if any two raters
		  • face and content validity – expert opinion
		  • final QIs categorized in 4 'domains' based on important aspects of pharmaceutical care described by Higashi et al.

Chen *et al.* 2011	Diabetes Empowerment Process Scale: development and psychometric testing of the Chinese version	• objective: to develop and test the psychometric properties of the Chinese version of the Diabetes Empowerment Process Scale • setting: China • population: outpatients living with diabetes • sample size: 211 outpatients; calculation not reported • statistical methods: stages – item generation and content validity testing (through concept analysis of literature and expert opinion respectively), item analysis (sampling – 10 participants per item, item analysis to check correlation of each item with total score) , validity testing (EFA, CFA, concurrent validity using similar tools), reliability testing (Cronbach's alpha for each scale – internal consistency; re-administration of tool to new sample then calculating ICC – test-retest reliability) • findings: satisfactory reliability and validity; no 'floor' or 'ceiling' effects; scale and subscales correlated with global self-care behaviours

Authors	Title	Description
		• key refs: Bear 1990 (Theoretical basis of measurement); Bentler & Chou 1986, MacCallum *et al.* 1996 (sample size); DeVellis 1991, Ferketich 1991 (item selection)

Bamm *et al.*
2010
	Validation of the measure of process of care for adults: a measure of client-centred care	• objective: to assess psychometric properties of MPOC-A
		• setting: regional orthopaedic service of a university-affiliated hospital in Canada
		• population: patients and their families who had joint replacement surgery in Jan-Aug 2007
		• sample size: convenient sample of all patients treated at the time; exclusion was lack of fluency in English.
		• statistical methods (expectations):
		  • internal consistency – Cronbach's alpha > 0.8
		  • reliability: inter-rater reliability correlation – ICC of 0.5-0.7 between patients' and families' scores on MPOC-A; high test-retest reliability– ICC 0.7-0.9 (both by comparing time 1 vs. time 2 scores)
		  • validity: cross-sectional – MPOC-A vs. CSQ (client satisfaction questionnaire) scores; convergent construct.
		• findings: high internal consistency, 'moderate to good' correlation between scales, inter-rater agreement; high test-retest reliability

Suhonen *et al.*
2010
	Individualized care scale – nurse version: a Finnish validation study	• objective: develop the ICS-nurse and ensure its validity and reliability
		• design: methodological design
		• setting: inpatient wards in one university, two regional hospitals, two psychiatric hospitals and four health centres in Finland
		• population: Finnish nursing staff
		• sample size: not reported
		• statistical methods:
		  • sum scores by averaging
		  • internal consistency reliability: Cronbach's alpha and item analysis including inter-tem, average inter-item and item-to-total correlations
		  • validity: content validity – producing the measure from a concept analysis including expert analysis; construct validity – PCA with Kaiser-Meyer-Olkin measure of sampling adequacy for FA; Spearman's rho correlations between subscales to check associations, SEM?
		• findings: good evidence of content validity, few missing data, satisfactory reliability, some ICCs high; PCA supported 3 component structure

Authors	Title	Description
		• key refs: Suhonen *et al.* 2000, Suhonen *et al.* 2005

Mael *et al.* 2010	Development of a model and measure of process-oriented quality of care for substance abuse treatment	• objective: to develop and validate a model of QoC for substance abuse treatment
		• setting: the US
		• population: substance abuse treatment staff
		• sample size: 'representative' sample of 51 substance abuse treatment agencies purposively selected
		• methods:
		  • development of the model and QoC scale using a critical incident technique (CIT) – staff describe incidents they have been involved in that affected organizational outcomes; these are sorted and categorized iteratively into a performance model (consensus based Delphi-like process)
		  • QoC scale based on 15 dimensions created above made up of 100 Likert-type items (5-10 per scale)
		  • Validation – internal reliability, confirmatory factor analysis, inter-rater reliability

Chevat *et al.* 2009	Development and psychometric validation of a self-administered questionnaire assessing the acceptance of influenza vaccination: the Vaccinees' Perception of Injection (VAPI) questionnaire	• objective: to develop and validate a self-administered questionnaire for use in clinical trials to assess subjects' perception and acceptance of influenza vaccination and injection site reactions
		• design: NR
		• setting: US, Germany, Switzerland; validation in France, Belgium, Germany, Italy and UK
		• population: adult subjects receiving IM influenza vaccination
		• sample size: NR
		• statistical methods:
		  • PCA with Varimax rotation and Kaiser-Guttman criterion to retain factors
		  • Final VAPI structure confirmed using Multitrait analysis based on item-scale Spearman correlations; item discriminant validity... floor and ceiling effects; scale-scale correlation using Spearman coefficients
		  • Clinical validity – comparing questionnaire ISR score vs. description in CRFs using Mann-Whitney-Wilcoxon and Kruskal-Wallis tests.
		  • Internal consistency reliability – Cronbach's alpha > 0.7
		• findings: PCA and MA resulted in deletion of 23 items; final PCA with 521 subjects; all items except one met convergent validity criterion for 4 dimensions; Cronbach's alpha > 0.7 indicating good internal

Authors	Title	Description
		consistency (but low for one dimension in one setting; clinical validity - scores higher on average for those with more severe reactions in CRF • key refs: Campbell *et al.* 1959

Najjar-Pellet *et al.* 2008	Quality assessment in intensive care units: proposal for a scoring system in terms of structure and process	• objective: to develop a score for assessing quality of ICU care in terms of structure and process based on literature review, expert opinion, field tests and analysis
		• design: feasibility observational study
		• setting: a French regional clinical research project
		• population: ICUs with focus on members of the ICU teams
		• sample size: 40; no report of how this was arrived at
		• methods:
		  • bibliographic review (including brainstorming to define 'quality') > expert consideration (Delphi type) > field test (in 2 ICUs) > descriptive analysis ()frequency tables, correlations etc) > final consensus to determine which variables to retain
		  • grouping into 'dimensions' of variables thought to be discriminating, uncorrelated, non-redundant and representative of QoC
		  • summarised by simple addition (total score)
		  • levels of performance measured by percentage achievement of maxima
		  • relationships between scores tested using Pearson's correlation matrix
		  • findings: adherence to methodology, use of Delphi technique, direct observations, elimination of subjectivity, coverage of components of quality in intensive care, independence of dimensions, feedback of participating ICUs, and equivalence to some available standards all support this score as a useful one

Llewellyn *et al.* 2007	An index of orthodontic treatment complexity	• objective: to develop an index for measuring treatment complexity with input factors related to complexity and output being a score reflecting the degree of treatment complexity • design: NR • setting: 2 dental hospitals in the UK • population: (dental casts and records of) patients treated for dental problems between June 1996 and December 2003 • sample size: 120 – determination not stated. • methods (steps): 1. scoring of the study casts using unweighted peer

Authors	Title	Description
		assessment rating (PAR) 2. grading perceived treatment complexity on a 6-point scale 3. selection of up to 3 occlusion factors contributing to complexity from a pre-set list 4. multiple regression to assess relationship between complexity grade and occlusion factors – partial regression coefficients used to derive weightings for occlusion factors 5. orthodontic treatment complexity = occlusion component score x weighting 6. summing up of component score 7. Spearman ranked correlation coefficients to study relationship between calculated and perceived complexity • key refs: Brook and Shaw 1989; Richmond et al., 1992; DeGuzman et al., 1995; Hamdan and Rock, 1999; Daniels and Richmond, 2000

Siebes *et al.*
2006
	Family-centred services in The Netherlands: validating a self-report measure for paediatric service providers	• objective: to validate the Dutch translation of the Canadian Measure of Processes of Care for Service Providers questionnaire MPOC-SP
		• setting: paediatric rehab in the Netherlands
		• population: service providers representing 7 children's rehab centres and affiliated schools
		• sample size: 163 service providers based on 5-10 subjects per item
		• statistical methods:
		  • face-validity: service providers to add or omit what they found most or least important
		  • test-retest reliability: rerunning questionnaire to a subset of initial sample
		  • skweness/kurtosis to check suitability of parametric statistics
		  • validity: construct validity – Spearman's rank correlations to confirm original factor structure and examine correlations between item and scale scores;
		  • content and face validity: qualitative and quantitative data from additional 26-item questionnaire given to subsample
		• findings: good evidence of content and construct validity, internal consistency and moderate-to-good reliability
		• key refs: King *et al.* 2004b; King *et al.* 1996; Nunnally & Wilson 1975; Streiner & Norman OUP 1995

Authors	Title	Description

Sixma *et al.*
2000
	Quality of care from the perspective of elderly people: the QUOTE-elderly instrument	• objective: to develop an instrument to measure quality of care for elderly people from their own expectations and experiences
		• setting: the Netherlands
		• population: non-institutionalized elderly people of 65 years or older who have used general practice services in last 2 months
		• sample size: 13 for FGD; 961 for questionnaire
		• statistical methods:
		  • 59 items answered on 4-point Likert scale
		  • Likert responses transformed to standardised z-scores ranging 0-10
		  • Response categories dichotomised
		  • Data analysis: non-response, skewness, correlations
		  • EFA with Varimax rotation and Kaiser normalization, CFA, reliability analysis
		  • Reliability and stability checked using sub-sample

Ashton *et al.*
1999
	An empirical assessment of the validity of explicit and implicit process-of-care criteria for quality assessment	• setting: the US
		• population: men treated between 1987 and 1989 at southern VA hospitals
		• sample size: 12 hospitals for the case-control study and 159 hospitals for the readmission study
		• methods:
		  • QoC assessed using chart reviews by an administrative and a quality reviewer
		  • Quality reviewer used a large set of disease specific explicit process-of-care criteria marked as met or not met
		  • Inter-rater reliability assessed be comparing ratings of two assessors
		  • Adherence score = number of met criteria divided by number of applicable criteria, expressed as a percentage
		  • Categories = ‘domains’
		  • Differential weightings assigned by panel of experts
		  • Validation: convergence with measures of same construct, divergence with different construct – Spearman rank correlations; association of process with outcome
		  • findings: weighting did not make a difference in score’s outcome predictive ability (predictive validity) but samples

Authors	Title	Description
		too small to show important differences; apparent dose-response relationship – worse outcomes in lower percentiles of score

Symmons *et al.* 1995	Development and preliminary assessment of a simple measure of overall status in rheumatoid arthritis (OSRA) for routine clinical use	• objective: to develop and assess a simple measure of overall status for rheumatoid arthritis
		• setting: the UK
		• population: patients receiving care of rheumatoid arthritis in routine clinical settings
		• sample size: NR
		• methods:
		  • score has 4 components: demographic info, disease activity, damage score and treatment category
		  • items in the activity and damage scores selected using a heuristic approach
		  • overall status is indicated by a sequence of categorical scores for each components e.g. M50B.8.2.C is a 50y old man in second decade of disease with active disease but little cumulative damage on second-line therapy.
		  • validation: content and construct validity – Spearman rank correlation between different scales to demonstrate some independence; criterion validity – longitudinal studies proposed; discriminant validity – comparison of OSRA to actual outcomes
		• key refs: Apgar 1953

A.4. Assessment score vs. expanded criteria set for measuring adherence to malaria guidelines for the first 20 patients in the Kenyan district hospitals study data

Patient ID	No. of assessment tasks performed (assessment score)	Illness severity recorded (1) and correct (2)	No. of correct treatment tasks (correct drug, dose, route, frequency, duration)	Total number of correctly completed tasks (expanded criteria set)
1	6	1	2	9
2	1	0	2	3
3	6	0	0	6
4	6	1	3	10
5	5	0	3	8
6	6	1	1	8
7	6	2	1	9
8	6	0	2	8
9	6	2	3	11
10	6	1	0	7
11	2	1	3	6
12	6	1	1	8
13	4	2	3	9
14	6	2	2	10
15	6	0	2	8
16	6	1	1	8
17	6	1	3	10
18	6	1	3	10
19	2	1	1	4
20	5	2	2	9

A.5. The paediatric data abstraction form

Hosp ID.		HW code		Questionnaire No.		IP NO.		QID	
Name (First, Last)					Age		yrs		mths
Adm Date	/ / 20		Wt (kg)		Ht (cm)		WHZ		
Sex	M / F / E		Temp (^{0}C)		Vaccines	*PCV10 X...../E*	*M'sles* Y / N / E		*DTP/ Penta X...../E*
Is there a hospital folder with an IP number?									Y / N

History					Examination							
Length of illness	days				Airway	Clear	Stridor	Needs active support to open				E
Fever	Y		E		Breathing	Respiratory Rate		/min / E				
Cough	Y	N	E			Oxygen saturation		_ _ % SpO2				
						Central Cyanosis		Y	N	E		
Cough > 3 weeks	Y	N	E			Indrawing		Y	N	E		
Difficulty breathing	Y	N	E			Grunting		Y	N	E		
						Acidotic breathing		Y	N	E		
						Wheeze		Y	N	E		
						Crackles		Y	N	E		
					Circulation	Pulse	Weak	Norm	E	/min		
						Cap Refill	X	<2	2-3	>3	E	
						Skin temp	Not cold	hand	fore-arm	elbow	E	
						Pallor / Anaemia		0	+	+++	E	
Diarrhoea	Y		E		Dehydration	Sunken eyes		Y	N	E		
Diarrhoea > 14d	Y	N	E			Skin pinch (sec)		0	1	2	E	
Diarrhoea bloody	Y	N	E		Disability	AVPU	A	V	P	U	E	
Convulsions	Y	N	E			Can drink / breastfeed?		Y	N	E		
If yes, no of fits						Bulging fontanelle		Y	N	E		
Difficulty feeding	Y	N	E			Visible severe wasting		Y	N	E		
						Oedema	none	foot	knee	face	E	

Admission Diagnoses			
Malaria	□ Severe □ Non-sev □ No classif'n	Anaemia	□ Sev □ Non-sev
Pneumonia	□ V. Sev □ Sev □ Non-sev	Meningitis	

Diarrhoea	□ Non-bloody □ Bloody		
Dehydration	□ Shock □ Sev □ Some □ No classif'n		
HIV / AIDS	□ previous diagnosis □ Clinical suspicion		
Malnutrition	□ Kwash □ Marasm □ M. Kwash		
Other 1			

Investigations ordered			
	Ordered?	Results documented same day	Result (give units)
Malaria Slide	Y / N	Y / N / E	Pos / Neg
Hb / HCT / PCV	Y / N	Y / N / E	
HIV test	Y / N	Y / N / E	Pos / Neg
Glucose	Y / N	Y / N / E	
Lumbar Puncture	Y / N	Y / N / E (microscopy)	Pos / Neg
HIV test (inpatient)	Y / N	Y / N / E	Pos / Neg
Other tests			

Oxygen and Blood Transfusion – *Record data only about the <u>immediate admission</u> events*				
	Ordered?	Describe how prescribed (flow rate, device)		
Oxygen	Y / N	□ No detail	Flow rate =	Ncath / NP / mask / Oth
Transfusion	Y / N	□ No detail	Volume of Blood (mls) =	Duration prescribed (hrs)

Treatment – *Record only the initial treatment prescribed for the admission episode*						
	Was drug prescribed?	Drug prescription				
		Route	Dose	Units	Freq	Days
Antibiotics						
Penicillin	Yes / No	iv / im		mg / iu		
Gentamicin	Yes / No	iv / im		mg		
Amoxicillin	Yes / No	po		mg / mls / tabs		
ceftriaxone	Yes / No	iv / im		mg / mls / tabs		
chloramphenical	Yes / No	iv / im				
Metronidazole	Yes / No	iv / po		mg / mls / tabs		
Antimalarials						
Quinine (Load)	Yes / No	iv / im		mg	stat	
Quinine (Maint)	Yes / No	iv / im / po		mg		
Artemether - load	Yes / No	iv / im		mg	stat	
Artemether - maint	Yes / No	iv / im		mg		
Coartem	Yes / No	po		tabs		

Supportive Care						
Paracetamol	Yes / No	im / po		mg / mls / tabs		
Others						

Did the child have a prescription for fluids to treat dehydration.			Yes / No	
A. Was the child given fluids using iv route?		Yes / No	*If No, skip to part B, if yes continue*	
Fluid prescribed	Fluid prescription for Step 1 and 2 (up to 6 hours)			
	Step 1 / Step 2 plan used	Total Vol.	Time (hrs)	
Ring / NSal / HS Darr / Other	Yes / No			
B. Was the child given fluids using oral or ng route?		Yes / No	*If No, then skip this section, if yes continue.*	
Fluid prescribed	Fluid prescription for first 4 hours			
	Ng tube used	Total Volume	Time (hrs)	OR Volume with each stool
ORS / Ring / NSal / HSD / Other	Yes / No			Mls/Not indicated
Was the child prescribed any feeds?		Yes / No	*If No, then skip this section, if yes continue.*	
Feed prescribed	Feed prescription for first 24 hours			
	Route		Feed vol	Freq / 24hrs
F75 / F100 / Sp Milk / HPD / Other	ngt / po		mls / E	/ E

Fluid and Feed Monitoring	
Is there a feed/fluid monitoring chart for the first 24 hrs	Yes / No
Were feeds/fluids monitored as prescribed for the first 24 hrs	Yes / No
Ward rounds	
No of documented ward rounds	
No of documented major (consultant) ward rounds	

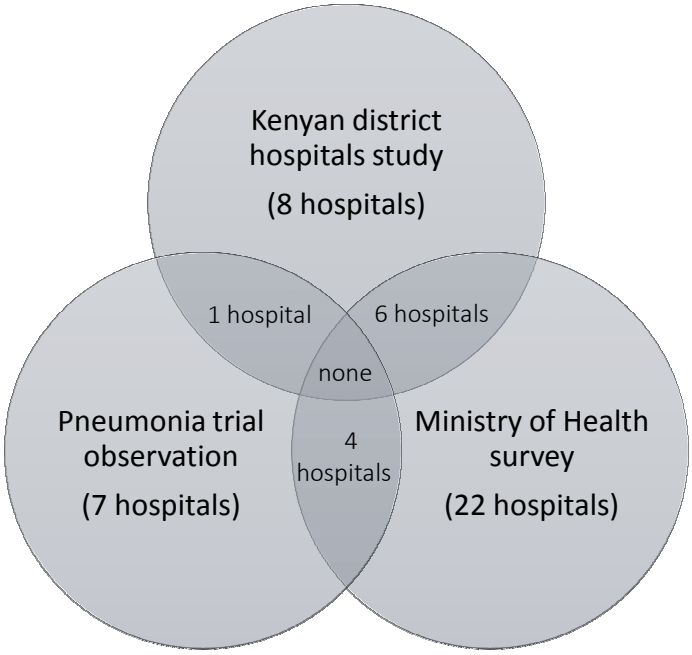
Vital signs chart				
Is there a vital signs chart?				Y ☐ N☐
What parameters are recorded?	Temperature	Respiratory rate	Pulse rate	Oxygen saturation
No of times documented in 48 hrs.				

Is there a discharge/Death summary in the case record? Y ☐ N ☐			
Discharge Date / /2001	Outcome	Alive / Dead / Refer'd / Absc'd	

Discharge Diagnoses: Select ONE *primary diagnosis (tick 1) and secondary diagnoses (tick 2)*						
Is there a clear primary diagnosis			Y ☐ N ☐			
Malaria	1	2	□ Severe □ Non-sev □ No classif'n	Meningitis	1	2
Pneumonia	1	2	□ V. Sev □ Sev □ Non-sev			
Diarrhoea	1	2	□ Non-bloody □ Bloody	*Other diagnosis (name)*	1	2
Dehydration	1	2	□ Sev □ Some □ No classif'n			

HIV / AIDS	1	2	□ previous □ clinical suspicion diagnosis		
Malnutrition	1	2	□ Kwash □ Marasm □ M. Kwash		
Anaemia	1	2	□ Sev □ Non-sev		
Last weight recorded				Date recorded	
Follow Up	Not arranged			Hospital	Disp / H. Centre
Did the child have a diagnosis of malaria or meningitis?					Yes / No
Were there convulsions				Yes / No	*If No, skip to part B, if yes continue*
	How many convulsions in 24 hrs			□ 1 -2 □ 3 -4 □>5	
	Type of convulsion experienced			□Generalized □ Focal □ Empty	
Was Lumber puncture done				Yes / No	*If No, then skip this section, if yes continue.*
	Bed side exam of CSF			□ dry tap □ under pressure □ turbid □ bloody □ clear □ not done	
		CSF investigations			
		Test done		Test results	
	CSF microscopy	Yes / No			
	CSF Biochemistry	Yes / No			
	CSF culture	Yes / No		□ no growth □ growth (list)……...	
What was the outcome?					
	□ Died □ Alive no sequelae □ Alive with sequelae □ Referred □ Absconded □ empty				
	If sequelae please give details				

A.6. Overlap between hospitals in the Kenyan district hospitals study, the pneumonia trial observation and the Ministry of Healthsurvey



A.7. Reliability and validity as conceptualised in this work

Reliability

The reliability of a measure refers to its consistency on repeated use. According to the classical test theory an observed measure is made up of two components, namely the 'true' level of the underlying trait being measures and some random error. Reliability is formally defined as 1 minus the ratio of error variance to observed-measure variance. Correlation coefficients are often used to estimate reliability of a measure; this coefficient quantifies the amount of variance attributable to differences in a measure on repeated use.

A measure's reliability can be tested in a number of ways. Internal consistency reliability, alternatively referred to as the coefficient of internal consistency, is the correlation between items that make up a measure. It is investigated by calculating the average correlation between pairs of items measuring the same domain of a construct (average inter-item correlation) or the average correlation between measurements from two halves of items contributing to a measure (split-half reliability). Internal consistency is related to the number of items making up the measure since it tends to decrease for measures made up of fewer items. This is because more items in a measure increase its ability to better characterise the underlying construct, and this is consistent with the domain sampling theory which attempts to explain the relationship between a construct, items and measure.

Other tests of reliability are: inter-item consistency which quantifies the degree to which items in a measure correlate to each other; test-retest reliability, also known as the coefficient of stability, which measures the correlation between two sets of measures on the same entity; alternate-forms reliability, also called the coefficient of equivalence or parallel-forms reliability, which is the correlation between two versions of measures of a construct when applied to the same entity; and inter-rater or inter-scorer reliability which measures the correlation between measurements by different users of the same measure on an entity. Differences in a construct across time, differences between items in a measure, and differences between users introduce variability to a measure. For this reason no measure is perfectly reliable, and the play of error in measurement must be recognized when interpreting a measure. Nevertheless reliability can be improved by minimizing these sources of error whenever possible.

Validity

The validity of a measure is how well it quantifies the construct it is intended to. A measure's validity is what determines the truth of inferences made on a construct based on the use of measure. Similar to reliability there are a number of ways in which the validity of a measure can be conceptualised but unlike reliability it is a rather subjective concept.

Construct validity, as the name implies, is the association between the measure and the construct it is intended to measure to the exclusion of other constructs. Face validity defines the credibility of a measure to its potential audience and users. Content validity refers to the suitability of the items making up the measure in sufficiently covering the construct. Construct validity, content validity and face validity of a measure can be assessed and improved by incorporating the opinions of experts and potential users of a measure in its design.

Convergent and discriminant validity define how similar a measure is to others it should theoretically be related to, and dissimilar to those pertaining to unrelated constructs, respectively. Closely aligned with this is criterion-related validity which is the correlation between a measure and other methods of quantifying a construct. The measure may be able to predict those other methods, in which case it possesses predictive criterion-related validity; alternatively it may only concur with related measurements made at the same time, and this is referred to as concurrent criterion-related validity. Since this form of validity is a quantitative concept suitable statistical techniques may be applied to investigate it.

Validity may also relate to the use of a measure to steer progress towards desired outcomes. This is called formative validity and it refers to the ability of a measure to provide the information necessary to improve the construct it measures. It can only be realised when the measure is put to use; for this reason it constitutes a form of post-hoc validation.

A.8. Levels of measurement

The units of all measures belong to fundamental levels of measurement, also referred to as measurement scales. These are, in increasing order of complexity: (1) the nominal level which represents measurements that take on the same numeric value if they share the same value of an attribute; in a sense it represents 'labels' rather than 'quantities' implied by measurement; (2) the ordinal level where measurements reflect an order on the defined attribute; (3) the interval level which is similar to the ordinal level but additionally differences between assigned numbers reflect equal differences of an attribute; (4) the ratio level where both differences and ratios of assigned numbers reflect differences and ratios of attributes; and (5) the absolute level in which all properties of the assigned numbers are analogous to properties of the attributes of interest.

In each of these levels (except the nominal level) there exist transformations which preserve relevant relationships of the measurement process. Linear transformations preserve relevant relationships on interval or higher levels; an example of a linear transformation is the conversion of temperature measurements from degrees Celsius to Fahrenheit by multiplying by $^{9}/_{5}$ and adding 32. Only monotone increasing transformations preserve relationships on the ordinal level [Stevens 1946, Killeen 1976]. A monotone increasing relationship is one in which whenever an attribute value of the construct increases, associated values of the item(s) used to measure them increase or stay the same (contrast this with *affine* relationships where the values of items measuring a construct would necessarily increase whenever attribute values of the construct do). Monotone increasing transformations leave a scale form-invariant: this means that the relative order of the elements of the scale is unaffected by a transformation. For example, the linear transformation of [1, 2, 3, 4] to [1, 3, 5, 7], and the non-linear transformation of the same to [4, 5, 13, 20], are monotone increasing transformations because they do not affect the order of relationships between the elements of this matrix.

Some authors have disagreed with this classification of levels of measurement. Velleman (1993) argued that it was too strict to apply to real-world data. Indeed a measure may lie squarely at any one of these levels, be a mixture of more than one level, lie somewhere between any two levels or even be too arbitrarily defined to

correspond to any of the levels originally proposed by Stevens. Other authors have gone further to suggest alternative, less restrictive classifications. For example Mosteller and Tukey (1977) listed levels of measurement in increasing order of complexity as follows: (1) names; (2) grades and ordered labels; (3) ranks starting at 1 (smallest) to largest; (4) counted fractions, including percentages, bounded by zero and one; (5) counts of non-negative integers; (6) amounts represented by non-negative real numbers; (7) balances which are unbounded positive or negative values.

Three distinct tiers of scale become apparent from these classifications. At the lowest tier is the simplest level of measurement in both classifications which can only assign named identities to attributes of a construct. Above this is an intermediate tier corresponding to measurement of discrete quantities of the construct of interest; at this tier are the ordinal and interval levels of measurement in Stevens' classification, and grades, ranks, counted fractions and counts in Mosteller and Tukey's classification. Many measures of complex constructs such as human intelligence and academic ability, and many academic grading systems often correspond to this tier of scale. At the highest tier are continuous quantities, represented by the ratio and absolute levels of measurement in Stevens' classification, and amounts and balances in Mosteller and Tukey's classification. Normality, approximate normality or normality upon transformation of a measure is only achievable with measures at the level of measurement at or above the intermediate tier of scale.

A.9. Drug dosage charts from the Basic Paediatric Protocols

Anti-malarial drug doses - ** *Please check the tablets*.

200 mg Quinine Sulphate = 200mg Quinine Hydrochloride or Dihydrochloride
200 mg Quinine Sulphate = 300mg Quinine Bisuphate

The table below assumes the use of a 200mg Quinine Sulphate tablet.

If the tablets are **300mg Quinine sulphate or dihydrochloride** then the table is **NOT** appropriate.

For **im Quinine** take 1ml of the 2mls in a 600mg Quinine suphate iv vial and add 5mls water for injection – this makes a 50mg/ml solution. Do not give more than 3mls per injection site. (See nursing chart for more detail)

Weight kg	Quinine loading, 15mg/kg iv infusion / im Once only	Quinine, maintenance, 10mg/kg iv infusion / im 12 hrly	Quinine, tabs, 10mg/kg 200mg QN sulphate** 8 hourly
3.0	45	30	1/4
4.0	60	40	1/4
5.0	75	50	1/4
6.0	90	60	1/2
7.0	105	70	1/2
8.0	120	80	1/2
9.0	135	90	1/2
10.0	150	100	3/4
11.0	165	110	3/4
12.0	180	120	3/4
13.0	195	130	3/4
14.0	210	140	3/4
15.0	225	150	1
16.0	240	160	1
17.0	255	170	1
18.0	270	180	1
19.0	285	190	1 1/4
20.0	300	200	1 1/4

AL = Artemether (20mg) – Lumefantrine (120mg) (with food): Give: Stat, after 8hrs, and then 12 hrly on Day 2 and Day 3		
Weight	**Age**	**Dose**
5 – 15 kg	3 months – 35 months	1 tablet
15 – 24 kg	3 years – 8 years	2 tablets
25 – 34 kg	9 years – 14 years	3 tablets

Oral antibiotic doses

Weight kg	Amoxicillin, oral, 25mg/kg/dose		Cloxacillin / Flucloxacillin 15mg/kg/dose		Chloramphenicol 25mg/kg/dose		Ciprofloxacin 15mg/kg/dose	Metronidazole 7.5mg/kg/dose
	mls susp 125mg/5ml	250mg caps	mls susp 125mg/5ml	250mg caps or tabs	mls susp 125mg/5ml	250mg caps	250mg tabs	200mg tabs
	12 hrly	12 hrly	8 hrly	8 hrly	6 hrly	6 hrly	12 hrly (for 3 days)	8 hrly
3.0	5	1/2*	2.5	1/4	4	n/a		
4.0	5	1/2*	2.5	1/4	4	n/a	1/4	
5.0	5	1/2*	5	1/4	6	n/a	1/4	1/4
6.0	5	1/2*	5	1/2	6	n/a	1/4	1/4
7.0	7.5	1/2*	5	1/2	8	n/a	1/2	1/2
8.0	7.5	1/2*	5	1/2	8	n/a	1/2	1/2
9.0	7.5	1	5	1/2	8	n/a	1/2	1/2
10.0	10	1	5	1	12	1	1/2	1/2
11.0	10	1	10	1	12	1	1	1/2
12.0	10	1	10	1	12	1	1	1/2
13.0	10	1	10	1	12	1	1	1/2
14.0	15	2	10	1	12	1	1	1
15.0	15	2	10	1	15	1	1	1
16.0	15	2	10	1	15	1	1	1
17.0	15	2	10	1	15	1	1	1
18.0	15	2	10	1	15	1	1	1
19.0	20	2	10	1	15	1	1	1
20.0	20	2	10	1		2	1	1

***Amoxicillin syrup should be used and capsules divided ONLY if syrup is not available**

Intravenous / intramuscular antibiotic doses – Ages 7 days and older.

Weight (kg)	Penicillin* (50,000iu/kg)	Ampicillin or Flucloxacillin (50mg/kg)	Chloramphenicol (25mg/kg)	Gentamicin (7.5mg/kg) im or iv over 3-5 mins	Ceftriaxone iv/im Max 50mg/kg 24hrly for neonates Meningitis / V Sev Sepsis, 50mg/kg BD	Metronidazole (7.5mg/kg)
	iv / im	iv / im	iv / im			iv
	6 hrly	8 hrly	6hrly - meningitis 8hrly – V.S. LRTI	24 hrly	50mg/kg	12 hrly < 1m, ≥ 1m 8 hrly
3.0	150,000	150	75	20	150	20
4.0	200,000	200	100	30	200	30
5.0	250,000	250	125	35	250	35
6.0	300,000	300	150	45	300	45
7.0	350,000	350	175	50	350	50
8.0	400,000	400	200	60	400	60
9.0	450,000	450	225	65	450	65
10.0	500,000	500	250	75	500	75
11.0	550,000	550	275	80	550	80
12.0	600,000	600	300	90	600	90
13.0	650,000	650	325	95	650	95
14.0	700,000	700	350	105	700	105
15.0	750,000	750	375	110	750	110
16.0	800,000	800	400	120	800	120
17.0	850,000	850	425	125	850	125
18.0	900,000	900	450	135	900	135
19.0	950,000	950	475	140	950	140
20.0	1,000,000	1000	500	150	1000	150

***NB. Double Penicillin doses if treating Meningitis and age > 1 month**

Urgent Fluid management – Child WITHOUT severe malnutrition.*

Weight kg	Shock, 20mls/kg Ringer's or Saline Immediately	Plan C – Step 1 **30mls/kg** Ringer's **Age <12m,** 1 hour **Age ≥1yr,** ½ hour	Plan C – Step 2 **70mls/kg** Ringer's or ng ORS **Age <12m,** over 5 hrs = drops/min**	**Volume**	**Age ≥ 1yr,** over 2½ hrs = drops/min**	Plan B - 75mls/kg Oral / ng ORS Over 4 hours
2.00	40	**50**	*10*	**150**	** ***Assumes 'adult' iv giving sets where 20drops=1ml***	**150**
2.50	50	**75**	*13*	**200**		**150**
3.00	60	**100**	*13*	**200**		**200**
4.00	80	**100**	*20*	**300**		**300**
5.00	100	**150**	*27*	**400**	*55*	**350**
6.00	120	**150**	*27*	**400**	*55*	**450**
7.00	140	**200**	*33*	**500**	*66*	**500**
8.00	160	**250**	*33*	**500**	*66*	**600**
9.00	180	**250**	*40*	**600**	*80*	**650**
10.00	200	**300**	*50*	**700**	*100*	**750**
11.00	220	**300**	*55*	**800**	*110*	**800**
12.00	240	**350**	*55*	**800**	*110*	**900**
13.00	260	**400**	*60*	**900**	*120*	**950**
14.00	280	**400**	*66*	**1000**	*135*	**1000**
15.00	300	**450**	*66*	**1000**	*135*	**1100**
16.00	320	**500**	*75*	**1100**	*150*	**1200**
17.00	340	**500**	*80*	**1200**	*160*	**1300**
18.00	360	**550**	*80*	**1200**	*160*	**1300**
19.00	380	**550**	*90*	**1300**	*180*	**1400**
20.00	400	**600**	*95*	**1400**	*190*	**1500**

**Consider Immediate blood transfusion if severe pallor or Hb <5g/dl on admission*

A.10. Tetrachoric correlation matrices of score items at baseline

Table A.10-1: Tetrachoric correlation matrix of assessment items in the basic malaria process-of-care score at baseline

	fever	convulsions	acidotic breathing	pallor	(in)ability to drink	level of consciousness	indrawing	malaria test
fever	1.00							
convulsions	0.38	1.00						
acidotic breathing	1.00	0.16	1.00					
pallor	0.15	0.14	1.00	1.00				
(in)ability to drink	0.08	0.13	-1.00	0.15	1.00			
level of consciousness	0.18	0.41	0.62	0.10	0.42	1.00		
indrawing	0.12	0.19	0.43	0.09	-0.07	0.25	1.00	
malaria test	0.12	0.03	1.00	0.05	0.15	0.15	0.01	1.00

Table A.10-2: Tetrachoric correlation matrix of assessment items in the basic pneumonia process-of-care score at baseline

	cough	difficult breathing	central cyanosis	(in)ability to drink	level of consciousness	grunting	indrawing	resp. rate
cough	1.00							
difficult breathing	0.06	1.00						
central cyanosis	0.02	0.03	1.00					
(in)ability to drink	-0.15	-0.08	0.29	1.00				
level of consciousness	-0.37	-0.22	0.05	0.37	1.00			
grunting	-0.17	0.22	0.32	-1.00	0.27	1.00		
indrawing	0.00	0.25	0.01	-0.09	0.33	0.29	1.00	
resp. rate	0.09	-0.31	-0.38	0.06	0.15	0.02	-0.23	1.00

Table A.10-3: Tetrachoric correlation matrix of assessment items in the basic diarrhoea/dehydration process-of-care score at baseline

	diarrhoea	vomiting	capillary refill	level of consciousness	(in)ability to drink	sunken eyes	skin pinch	pulse*
diarrhoea	1.00							
vomiting	0.12	1.00						
capillary refill	1.00	-1.00	1.00					
level of consciousness	1.00	0.08	-1.00	1.00				
(in)ability to drink	1.00	0.47	-1.00	0.58	1.00			
sunken eyes	1.00	-0.06	1.00	0.23	-1.00	1.00		
indrawing	1.00	0.07	-1.00	0.22	-1.00	0.54	1.00	
pulse*	–	–	–	–	–	–	–	1.00

*does not vary

Table A.10-4: Tetrachoric correlation matrix of treatment items in the basic process-of-care score at baseline

		Malaria					Pneumonia					Diarrhoea/Dehydration		
		drug	**route**	**dose**	**freq.**	**dur.**	**drug**	**route**	**dose**	**freq.**	**dur.**	**drug**	**dose**	**freq.**
Malaria	**drug**	1.00												
	route	1.00	1.00											
	dose	1.00	-1.00	1.00										
	freq.	1.00	1.00	0.92	1.00									
	dur.	1.00	1.00	1.00	1.00	1.00								
Pneumonia	**drug**						1.00							
	route						0.83	1.00						
	dose						0.19	1.00	1.00					
	freq.						0.82	1.00	0.96	1.00				
	dur.						0.83	0.99	0.96	0.99	1.00			
DnD	**drug**											1.00		
	dose											0.43	1.00	
	freq.											0.76	0.28	1.00

Table A.10-5: Tetrachoric correlation matrix of items in the modified process-of-care score at baseline

Complete assessment indicator has been excluded because it is perfectly correlation with the other two assessment items by design. 'pri.', 'sec.' and 'class.' are primary signs, secondary signs and illness severity classification indicator items respectively

		Malaria					Pneumonia					Diarrhoea/Dehydration				
		pri.	**sec.**	**class.**	**drug**	**use**	**pri.**	**sec.**	**class**	**drug**	**use**	**pri.**	**sec.***	**class**	**drug**	**use**
Malaria	**pri.**	1.00														
	sec.	0.31	1.00													
	class.	0.21	0.08	1.00												
	drug	0.26	0.11	1.00	1.00											
	use	1.00	0.25	-1.00	-1.00	1.0										
Pneumonia	**pri.**						1.00									
	sec.						1.00	1.00								
	class.						-0.20	1.00	1.00							
	drug						-0.27	-1.00	1.00	1.00						
	use						-0.06	-1.00	0.21	0.36	1.00					
DnD	**pri.**											1.00				
	sec.*											_	1.00			
	class.											0.11	_	1.00		
	drug											-0.10	_	0.09	1.00	
	use											0.20	_	-0.05	0.25	1.00

A.11. One-factor structural equation model of the modified and combined process of care scores

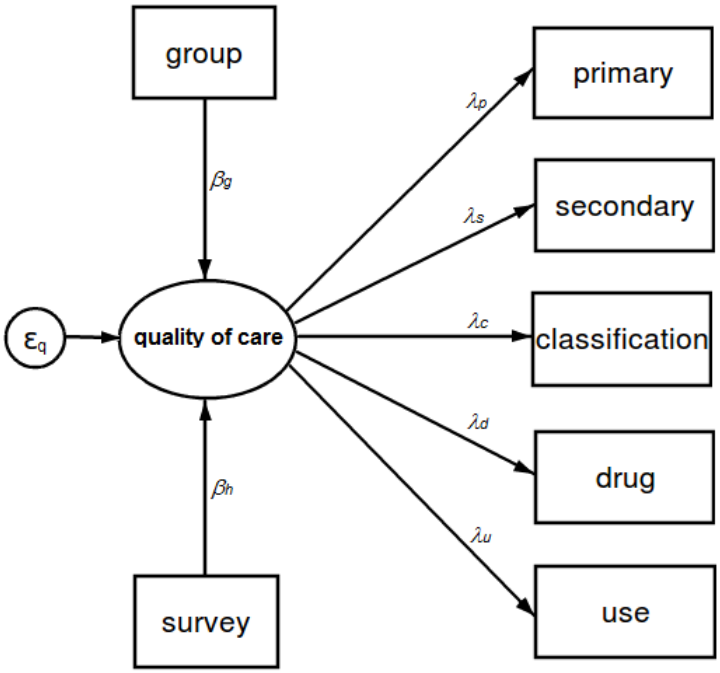


Figure A.11-1: Path diagram of the one-factor structural equation model of the modified and combined scores

Table A.11-1: Parameter estimates of a one-factor structural equation model of the modified score for malaria quality of care

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.555)	_	_
λ_s	1.422 (0.745)	0.119	< 0.001
λ_c	_	_	_
λ_d	1.520 (0.785)	0.127	< 0.001
λ_u	1.187 (0.643)	0.113	< 0.001
β_g	0.431 (0.365)	0.042	< 0.001
β_h	0.259 (0.520)	0.022	< 0.001
var(ε_q)	0.216	0.033	< 0.001

n = 2,930; fit indices: χ^2_7 = 163.888, p-value <0.001; CFI = 0.901; TLI = 0.831; RMSEA = 0.087

Table A.11-2: Parameter estimates of a one-factor structural equation model of the modified score for pneumonia quality of care

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.543)	_	_
λ_s	2.084 (0.962)	0.275	< 0.001
λ_c	2.195 (0.994)	0.293	< 0.001
λ_d	1.424 (0.731)	0.198	< 0.001
λ_u	0.751 (0.418)	0.116	< 0.001
β_g	0.376 (0.326)	0.058	< 0.001
β_h	0.281 (0.548)	0.037	< 0.001
var(ε_q)	0.203	0.052	< 0.001

n = 1,409; fit indices: χ^2_{11} = 129.204, p-value <0.001; CFI = 0.937; TLI = 0.908; RMSEA = 0.087

Table A.11-3: Parameter estimates of a one-factor structural equation model of the modified score for diarrhoea/dehydration quality of care

Parameter	Estimate (standardised)	Standard error	p-value
λ_p	1.000 (0.261)	_	_
λ_s	4.039 (0.961)	1.306	0.002
λ_c	3.059 (0.757)	0.904	0.001
λ_d	1.846 (0.474)	0.589	0.002
λ_u	1.455 (0.376)	0.500	0.004
β_g	0.152 (0.290)	0.053	0.004
β_h	0.060 (0.266)	0.021	0.005
var(ε_q)	0.056	0.033	0.093

n = 529; fit indices: χ^2_{16} = 383.026, p-value <0.001; CFI = 0.817; TLI = 0.707; RMSEA = 0.113

Item loadings on the single factor were significant and positive in all three diseases. Despite a 0.047 rise in the CFI of the malaria model indicative of a slightly better fit, all model fit indices indicate poorer fit compared to the two-factor model.

A.12. Hierarchical logistic regression model in section 6.3 on admission episodes lasting up to 7 days

Table A.12-1: Adjusted effect of quality of care measured using the combined score on death in admission episodes lasting up to 7 days

	Adjusted odds ratio for death	95% confidence interval	p-value
Combined score	0.87	0.78 – 0.98	0.026
Age (years)	0.56	0.48 – 0.65	<0.001
Sex (female vs. male)	1.19	0.92 – 1.52	0.170
Number of diseases diagnosed	0.79	0.60 – 1.04	0.097
Severity (3=highest, 2=intermediate, 1=lowest)	1.59	1.32 – 1.91	<0.001
Identity of disease			
Diarrhoea/dehydration	1.00		
Malaria	0.81	0.48 – 1.35	< 0.001
Pneumonia	1.52	0.91 – 2.55	
Duration of admission (days)	0.53	0.48 – 0.57	< 0.001
Group			
Control	1.00		
Intervention	1.72	0.46 – 6.40	0.416
Survey			
Baseline	1.00		
1st follow-up	1.30	0.48 – 3.56	0.610
2nd follow-up	1.64	0.59 – 4.52	
End-point	1.71	0.64 – 4.58	
Group-survey interaction			
Intervention x 1st follow-up	1.25	0.36 – 4.32	
Intervention x 2nd follow-up	0.83	0.24 – 2.84	0.697
Intervention x End-point	0.78	0.26 – 2.80	
Random effects	**Variance**		
Clinician (i=391)	0.21	0.07 – 0.69	
Hospital (j=8)	0.25	0.07 – 0.85	

n = 4,119, died = 332 | Goodness of fit test: Hosmer-Lemeshow $\chi^2_8 = 7.63$, p = 0.4706 | ROC AUC = 0.8106

A.13. Fit diagnostic test for the hierarchical logistic regression model in section 6.3

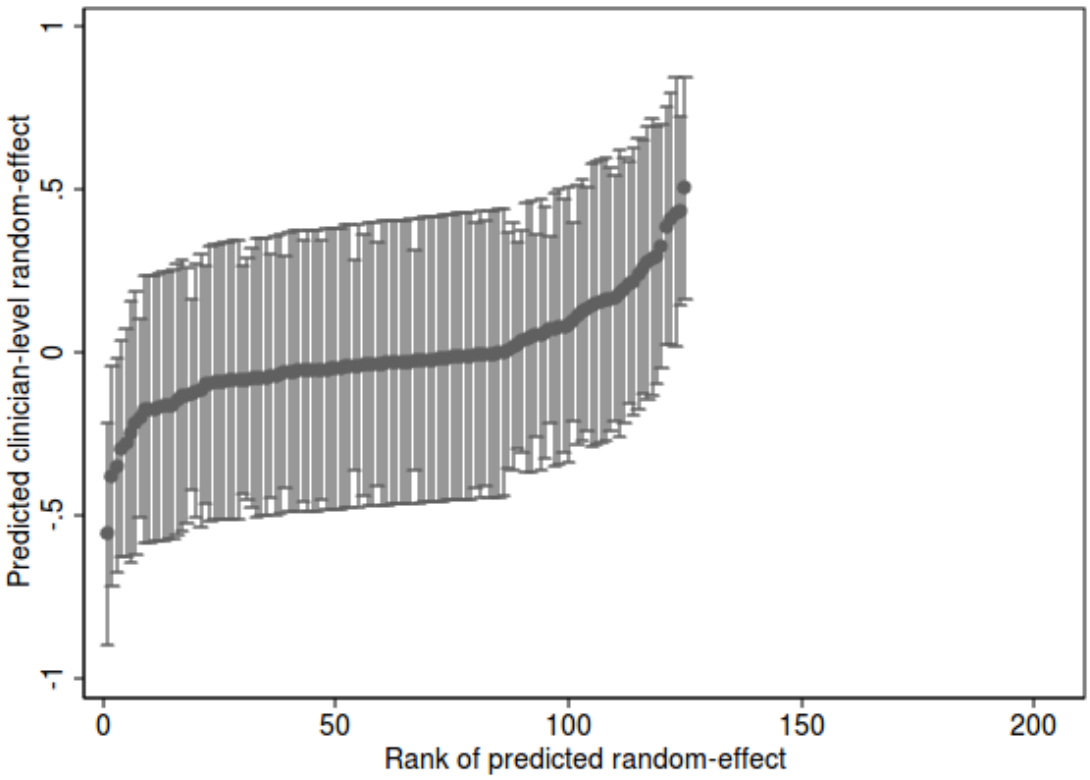


Figure A.13-1: Plot of predicted clinician-level random effects versus their rank for logistic regression model in section 6.3

The sigmoid shape of the scatter of random effects estimates implies that the assumption of normal distribution is not violated

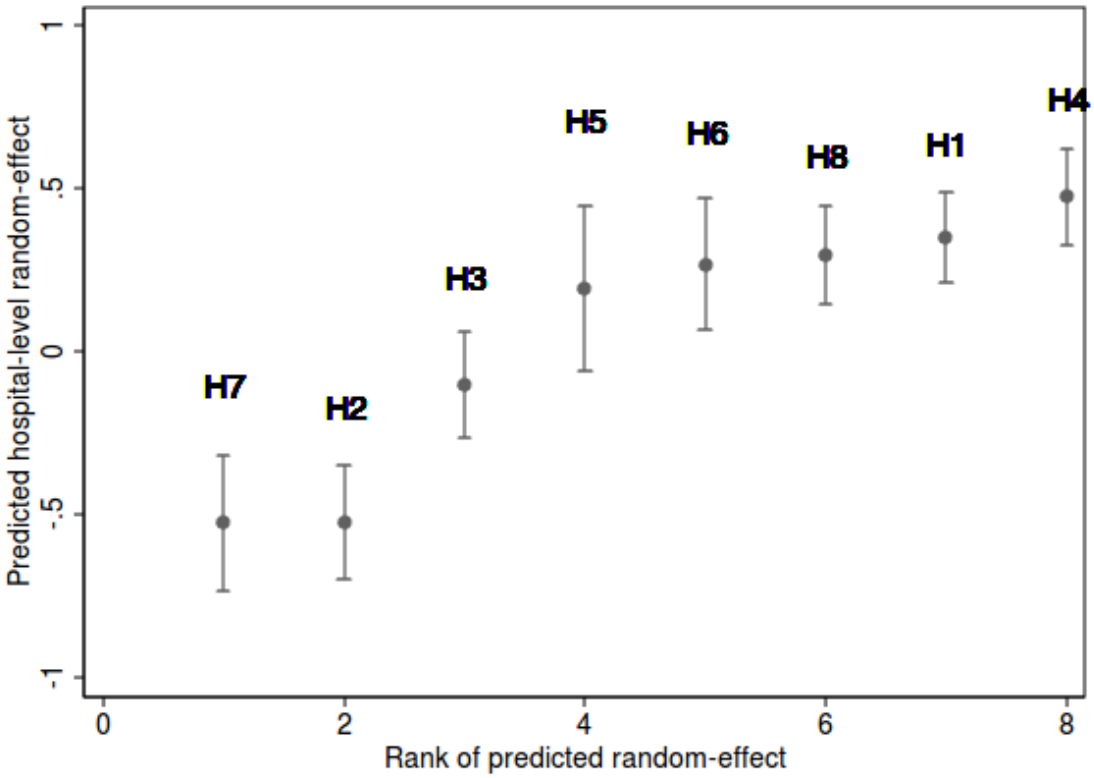


Figure A.13-2: Plot of predicted hospital-level random effects versus their rank for logistic regression model in section 6.3

The small number of hospitals makes it difficult to identify the same sigmoid shape of the scatter of random effects as seen in the clinician-level estimates. However there are no signs of extreme values

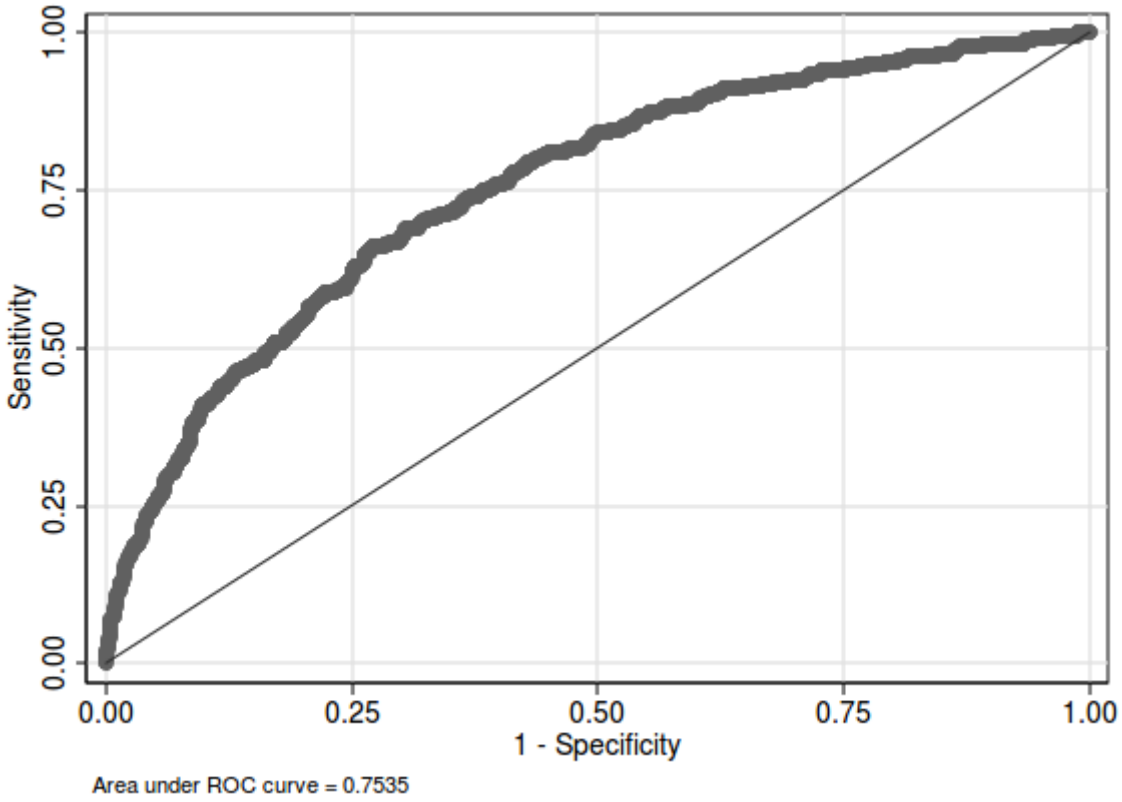


Figure A.13-3: Receiver operating characteristics (ROC) curve for logistic regression model in section 6.3

The area under the curve (AUC) of 0.7535 implies to an 'acceptable' model fit [Hosmer & Lemeshow 2000]

A.14. Hierarchical logistic regression models of the association between the 7-point combined score and mortality on the validation data

Table A.14-1: Adjusted effect of quality of care measured by the combined score on death in the pneumonia trial observation data

	Adjusted odds ratio for death	95% confidence interval	p-value
Combined score	0.72	0.64 – 0.81	< 0.001
Age (years)	0.95	0.89 – 1.02	0.167
Sex (female vs. male)	1.06	0.84 – 1.34	0.602
Number of diseases diagnosed	0.79	0.55 – 1.14	0.211
Severity (3=highest, 2=intermediate, 1=lowest)	2.53	2.01 – 3.19	< 0.001
Identity of disease			
Diarrhoea/dehydration	1.00		
Malaria	0.29	0.16 – 0.51	< 0.001
Pneumonia	0.86	0.59 – 1.29	
Duration of admission (days)	1.00	0.99 – 1.01	0.267
Random effects	**Variance**		
Hospital (j=7)	0.59	0.19 – 1.86	

n= 5,924 | Goodness of fit for logistic regression model: Hosmer-Lemeshow $\chi^2_8 = 85.45$, $p < 0.0001$ | ROC AUC = 0.7552

Table A.14-2: Adjusted effect of quality of care measured by the combined score on death in the Ministry of Health survey data

	Adjusted odds ratio for death	95% confidence interval	p-value
Combined score	0.55	0.36 – 0.84	0.006
Age (years)	0.63	0.40 – 0.98	0.041
Sex (female vs. male)	0.98	0.44 – 2.17	0.957
Number of diseases diagnosed	(omitted – varies very little with the outcome)		
Severity (3=highest, 2=intermediate, 1=lowest)	3.88	1.44 – 10.49	0.007
Identity of disease			
Diarrhoea/dehydration	1.00		
Malaria	0.11	0.01 – 0.94	0.118
Pneumonia	0.21	0.04 – 1.08	
Duration of admission (days)	0.95	0.87 – 1.05	0.338
Random effects	**Variance**		
Hospital (j=22)	1.12	0.21 – 5.89	

n= 592 | Goodness of fit for logistic regression model: Hosmer-Lemeshow $\chi^2_8 = 17.84$, $p = 0.0224$ | ROC AUC = 0.8842

A.15. Further examples of application of funnel plots for comparing mean quality of care scores across clinicians, diseases and time

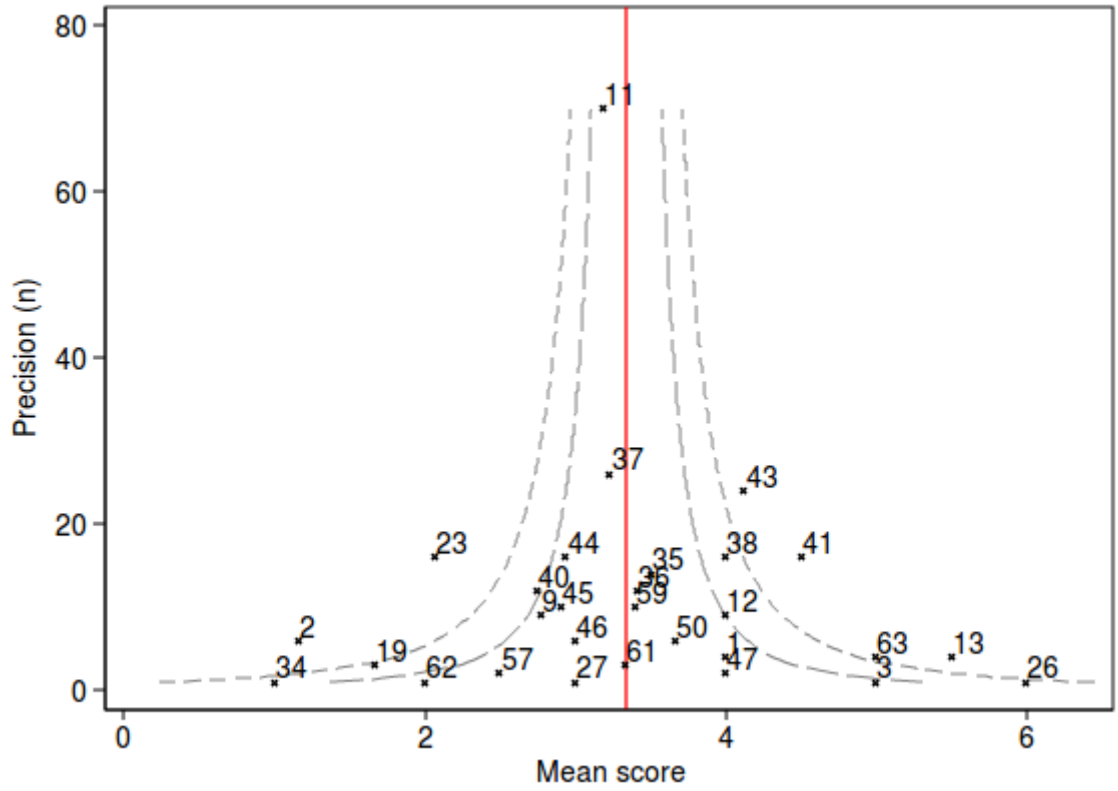
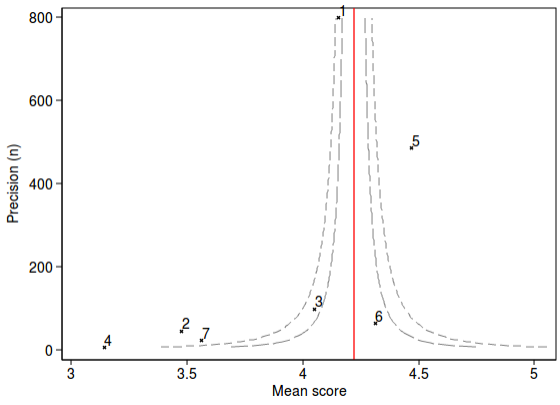


Figure A.15-1: Funnel plot comparing clinician performance (mean scores) in one hospital Numbered dots are clinician IDs. This plot illustrates a wide variability in clinician performance



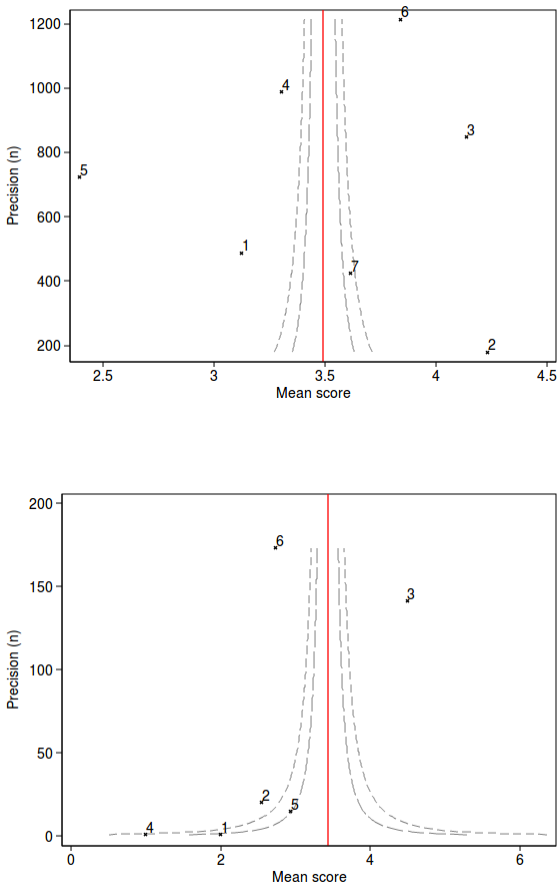


Figure A.15-2: Funnel plots showing disease-specific quality of care scores for malaria (top), pneumonia (middle) and diarrhoea/dehydration (bottom) in the seven Pneumonia observational study hospitals

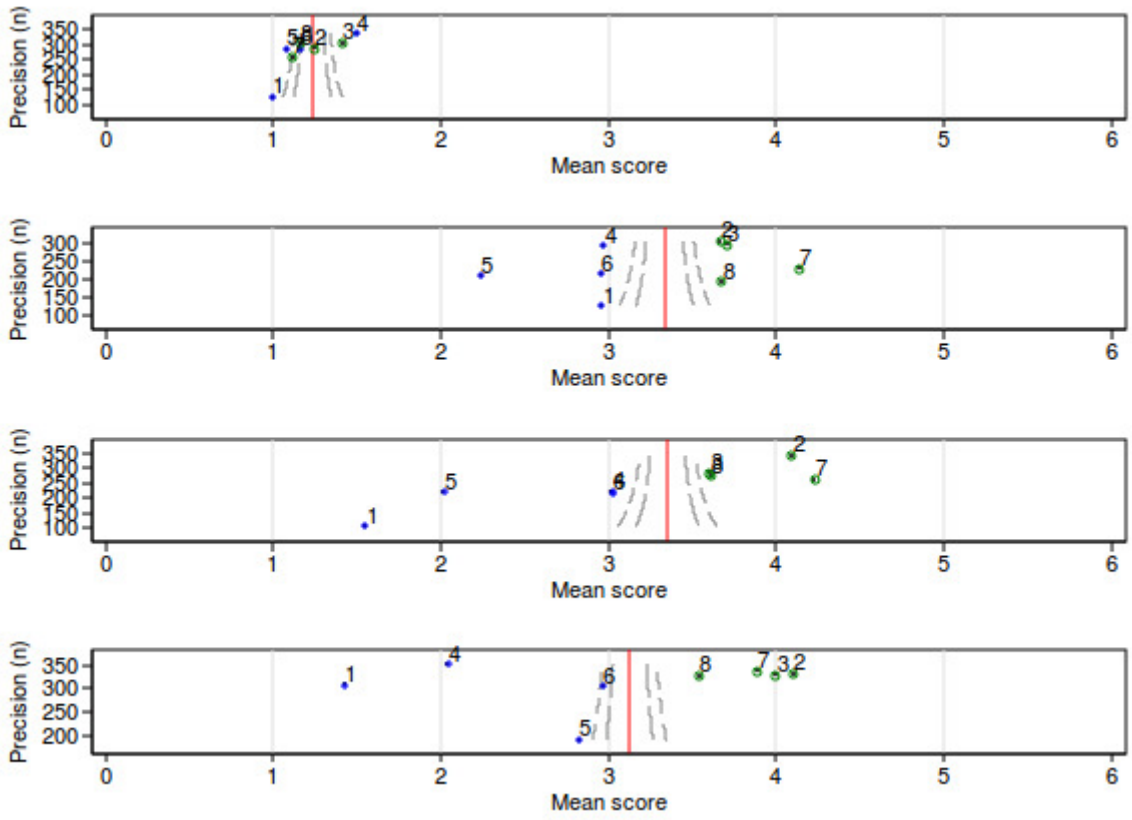


Figure A.15-3: Funnel plots charting changes in quality of care in the intervention (2, 3, 7 and 8) and control (1, 4, 5 and 6) hospitals in the Kenyan district hospitals study at baseline (top), first and second follow-up, and main end-point (bottom) surveys respectively

At baseline all hospitals had low mean scores. One control and one intervention hospital had means above the overall average, one intervention and two control hospitals had means below the average, and the remaining three hospitals were within the average. After baseline, all control hospitals had means below the overall average and all intervention hospitals had means above the average, and this pattern was maintained all through the study until the main end-point.

A.16. Stata commands for main analyses

Obtaining patient-level basic, modified and combined scores from case record data

```
1.    *Create indicator variables
2.    *Malaria
3.    local loop = 1
4.    while `loop' < 8 {
5.    foreach sign_symptom of varlist fever convulsions acidotic__breathing
      pallor can_drink avpu indrawing {
6.    gen mal_assess_`loop' = 0
7.    replace mal_assess_`loop' = 1 if `sign_symptom' != "E" & `sign_symptom'
      != ""
8.    label var mal_assess_`loop' "`sign_symptom'"
9.    local loop = `loop' + 1
10.   }
11.   }
12.   gen mal_assess_8 = 0
13.   replace mal_assess_8 = 1 if mal1_order == "Y"
14.   label var mal_assess_8 "BS for MPS"
15.   *Pneumonia
16.   local loop = 1
17.   while `loop' < 9 {
18.   foreach sign_symptom of varlist cough diff_breath c_cyanosis can_drink
      avpu grunting indrawing resp_rate {
19.   gen pneum_assess_`loop' = 0
20.   replace pneum_assess_`loop' = 1 if `sign_symptom' != "E" &
      `sign_symptom' != ""
21.   label var pneum_assess_`loop' "`sign_symptom'"
22.   local loop = `loop' + 1
23.   }
24.   }
25.   *Diarhoea/dehydration
26.   local loop = 1
27.   while `loop' < 9 {
28.   foreach sign_symptom of varlist diarrhoea vomits cap_refill avpu
      can_drink sunk_eyes skin_pinch pulse{ // recode if cold hands and
      restlessness/irritability known
29.   gen dnd_assess_`loop' = 0
30.   replace dnd_assess_`loop' = 1 if `sign_symptom' != "E" & `sign_symptom'
      != ""
31.   label var dnd_assess_`loop' "`sign_symptom'"
32.   local loop = `loop' + 1
33.   }
34.   }
```

```
35.  Assessment score type 1
36.  gen mal_assess_sum1 =
     mal_assess_1+mal_assess_2+mal_assess_3+mal_assess_4+mal_assess_5+mal_ass
     ess_6+mal_assess_7+mal_assess_8
37.  gen pneum_assess_sum1 =
     pneum_assess_1+pneum_assess_2+pneum_assess_3+pneum_assess_4+pneum_assess
     _5+pneum_assess_6+pneum_assess_7+pneum_assess_8
38.  gen dnd_assess_sum1 =
     dnd_assess_1+dnd_assess_2+dnd_assess_3+dnd_assess_4+dnd_assess_5+dnd_ass
     ess_6+dnd_assess_7+dnd_assess_8
39.  Assessment score type 2
40.  *Malaria
41.  gen mal_assess_1_primary = fever != "E" & fever != ""
42.  label var mal_assess_1_primary "fever"
43.  replace dx1_malaria = "Non-severe" if strmatch(dx1_malaria, "*Non*sev*")
44.  replace dx1_malaria = "No classification" if strmatch(dx1_malaria,
     "*No*clas*")
45.  gen mal_assess_1_secondary =
     ((mal_assess_2+mal_assess_3+mal_assess_5+mal_assess_6>0) |
     (mal_assess_4==1 & mal_assess_7==1) & dx1_malaria=="Severe") |
     ((mal_assess_2+mal_assess_3+(mal_assess_5 | mal_assess_6)==3) |
     (mal_assess_4+mal_assess_7==2))
46.  label var mal_assess_1_secondary "convulsion or acidosis or inability to
     drink or AVPU <A or pallor with resp. distress"
47.  gen mal_assess_1_complete = mal_assess_sum==8
48.  label var mal_assess_1_complete "complete assessment for malaria"
49.  *Pneumonia
50.  gen pneum_assess_1_primary = (cough != "E" & cough != "") | (diff_breath
     != "E" & diff_breath != "")
51.  label var pneum_assess_1_primary "cough or difficult breathing"
52.  replace dx1_pneum = "Very severe" if strmatch(dx1_pneum, "*V. Sev*")
53.  replace dx1_pneum = "Non-severe" if strmatch(dx1_pneum, "*Non*sev*")
54.  replace dx1_pneum = "No classification" if strmatch(dx1_pneum,
     "*No*clas*")
55.  destring resp_rate, force replace
56.  gen pneum_assess_1_secondary =
     ((pneum_assess_3+pneum_assess_4+pneum_assess_5+pneum_assess_6>0 &
     dx1_pneum=="Very severe") | (pneum_assess_3+(pneum_assess_4 |
     pneum_assess_5)+pneum_assess_6+pneum_assess_7==4 & dx1_pneum=="Severe")
     | (pneum_assess_3+(pneum_assess_4 |
     pneum_assess_5)+pneum_assess_6+pneum_assess_7==4 & resp_rate != . &
     (age_years != . | age_mths != . | age_days != .) & dx1_pneum=="Non-
     severe"))
57.  label var pneum_assess_1_secondary "at least one of cyanosis, inability
     to drink, AVPU, grunting or acidotic breathing for very severe
     pneumonia, OR all of the above plus indrawing for severe pneumonia, OR
     all of the above cumulatively plus respiratory rate for non-severe
     pneumonia"
58.  gen pneum_assess_1_complete = pneum_assess_sum == 8
59.  label var pneum_assess_1_complete "complete assessment for pneumonia"
```

```
60.  *Diarrhoea/dehydration

61.  gen dnd_assess_1_primary = dnd_assess_1 | dnd_assess_2

62.  label var dnd_assess_1_primary "diarrhoea or vomiting"

63.  replace dx1_dehydrat = "Shock" if dx1_other_1 == "Shock" | dx1_other_2
     == "Shock"

64.  replace dx1_dehydrat = "Severe" if dx1_dehydrat == "Sev"

65.  replace dx1_dehydrat = "No classification" if strmatch(dx1_dehydrat,
     "*No*clas*")

66.  gen dnd_assess_1_secondary = ((dnd_assess_3 | dnd_assess_4 |
     dnd_assess_5) & dnd_assess_8 & dx1_dehydrat == "Shock") |
     (dnd_assess_3+(dnd_assess_4 |
     dnd_assess_5)+dnd_assess_6+dnd_assess_7+dnd_assess_8==5 & (dx1_dehydrat
     == "Severe" | dx1_dehydrat == "Some" | dx1_dehydrat == "No
     classification"))

67.  label var dnd_assess_1_secondary "capillary refill or AVPU, plus weak
     pulse for shock OR these plus sunken eyes and skin pinch for all other
     categories"

68.  gen dnd_assess_1_complete = dnd_assess_sum == 8

69.  label var dnd_assess_1_complete "complete assessment for
     diarrhoea/dehydration"

70.  Diagnosis score

71.  *Malaria

72.  gen mal_diag_id_guide = fever == "Y" & mal1_order == "Y"

73.  gen mal_diag_id_clin = dx1_malaria == "No classification" | dx1_malaria
     == "Non-severe" | dx1_malaria == "Severe"

74.  gen mal_diag_id_corr = mal_diag_id_clin == 1 & mal_diag_id_guide == 1

75.  gen mal_diag_sev_guide = fever == "Y" & (convulsions == "Y" |
     acidotic__breathing == "Y" | (pallor == "+++" & indrawing == "Y") |
     can_drink == "N" | avpu == "V" | avpu == "P" | avpu == "U")

76.  gen mal_diag_sev_clin = dx1_malaria == "Severe"

77.  gen mal_diag_nonsev_guide = fever == "Y" & convulsions == "N" &
     acidotic__breathing == "N" & (pallor == "0" | pallor == "+") &
     (can_drink == "Y" | avpu == "A")

78.  gen mal_diag_nonsev_clin = (dx1_malaria == "Non-severe" | dx1_malaria ==
     "Non-sev")

79.  gen mal_diag_noclass_clin = dx1_malaria == "No classif'n"

80.  gen mal_diag_class_corr = (mal_diag_sev_clin == 1 & mal_diag_sev_guide
     == 1) | (mal_diag_nonsev_clin == 1 & mal_diag_nonsev_guide == 1)

81.  gen mal_diag_any_class = mal_diag_sev_clin + mal_diag_nonsev_clin

82.  *Pneumonia

83.  gen pneum_diag_id_guide = cough == "Y" | diff_breath == "Y"

84.  gen pneum_diag_id_clin = dx1_pneum == "No classification" | dx1_pneum ==
     "Non-severe" | dx1_pneum == "Severe" | dx1_pneum == "Very severe"

85.  gen pneum_diag_id_corr = pneum_diag_id_clin == 1 & pneum_diag_id_guide
     == 1

86.  gen pneum_diag_very_sev_guide = (cough == "Y" | diff_breath == "Y") &
     (c_cyanosis == "Y" | can_drink == "N" | avpu == "V" | avpu == "P" | avpu
     == "U" | grunting == "Y" | acidotic__breathing == "Y")

87.  gen pneum_diag_very_sev_clin = dx1_pneum == "Very severe"

88.  gen pneum_diag_sev_guide = (cough == "Y" | diff_breath == "Y") &
     c_cyanosis == "N" & can_drink == "Y" & avpu == "A" & grunting == "N" &
```

```
      acidotic__breathing == "N" & indrawing == "Y"

89.   gen pneum_diag_sev_clin = dx1_pneum == "Severe"

90.   gen pneum_diag_nonsev_guide = (cough == "Y" | diff_breath == "Y") &
      c_cyanosis == "N" & can_drink == "Y" & avpu == "A" & grunting == "N" &
      acidotic__breathing == "N" & indrawing == "N" & ((age_years < 1 &
      age_mths > 1 & age_mths < 12 & resp_rate >= 50) | (age_years > 0 &
      resp_rate >= 40))

91.   gen pneum_diag_nonsev_clin = dx1_pneum == "Non-severe"

92.   gen pneum_diag_noclass_clin = dx1_pneum == "No classification"

93.   gen pneum_diag_class_corr = (pneum_diag_very_sev_clin == 1 &
      pneum_diag_very_sev_guide == 1) | (pneum_diag_sev_clin == 1 &
      pneum_diag_sev_guide == 1) | (pneum_diag_nonsev_clin == 1 &
      pneum_diag_nonsev_guide == 1)

94.   gen pneum_diag_any_class = pneum_diag_very_sev_clin +
      pneum_diag_sev_clin + pneum_diag_nonsev_clin


95.   *Diarrhoea/dehydration

96.   gen dnd_diag_id_guide = diarrhoea == "Y" | vomits == "Y"

97.   gen dnd_diag_id_clin = dx1_diarrhoea == "Bloody" | dx1_diarrhoea ==
      "bloody" | dx1_diarrhoea == "Non-bloody" | dx1_diarrhoea == "No
      classification"

98.   gen dnd_diag_id_corr = dnd_diag_id_clin == 1 & dnd_diag_id_guide == 1

99.   gen dnd_diag_shock_guide = pulse == "Weak" & (cap_refill == ">3" | avpu
      == "V" | avpu == "P" | avpu == "U")

100.  gen dnd_diag_shock_clin = dx1_dehydrat == "Shock"

101.  gen dnd_diag_sev_guide = (can_drink == "N" | avpu == "V" | avpu == "P" |
      avpu == "U") & sunk_eyes == "Y" & skin_pinch == "2" // recode if cold
      hands known

102.  gen dnd_diag_sev_clin = dx1_dehydrat == "Severe"

103.  gen dnd_diag_some_guide = can_drink == "Y" & sunk_eyes == "Y" &
      (skin_pinch == "1" | skin_pinch == "2") // recode if
      restlessness/irritability known

104.  gen dnd_diag_some_clin = dx1_dehydrat == "Some"

105.  gen dnd_diag_none_clin = dx1_dehydrat != "Shock" & dx1_dehydrat !=
      "Severe" & dx1_dehydrat != "Some"

106.  gen dnd_diag_none_guide = pulse == "Normal" & (cap_refill == "<2" |
      cap_refill == "X") & (avpu == "A" | can_drink == "Y") & sunk_eyes == "N"
      & skin_pinch == "0" // recode if cold hands and
      restlessness/irritability known

107.  gen dnd_diag_class_corr = (dnd_diag_shock_clin == 1 &
      dnd_diag_shock_guide == 1) | (dnd_diag_sev_clin == 1 &
      dnd_diag_sev_guide == 1) | (dnd_diag_some_clin == 1 &
      dnd_diag_some_guide == 1) | (dnd_diag_none_clin == 1 &
      dnd_diag_none_guide == 1)

108.  gen dnd_diag_any_class = dnd_diag_shock_clin + dnd_diag_sev_clin +
      dnd_diag_some_clin

109.  Treatment score

110.  *Malaria

111.  gen mal_trt_sev_drug = quinl1_pres == "Y" | quinm1_pres == "Y"

112.  gen mal_trt_sev_route = (quinl1_route == "im" | quinl1_route == "iv") &
      (quinm1_route == "im" | quinm1_route == "iv")

113.  gen mal_trt_sev_dose = (quinl1_dose <= 20*i_weight & quinl1_dose >=
      20*i_weight & quinl1_unit == "mg") & (quinm1_dose <= 10*i_weight &
```

```
      quinm1_dose >= 10*i_weight & quinm1_unit == "mg")
114.  gen mal_trt_sev_freq =  quinm1_freq == 2
115.  gen mal_trt_sev_dur = (quinl1_freq == "stat" | quinl1_freq == "STAT" |
      quinl1_freq == "1") & (quinm1_days != "E" | quinm1_days != "")
116.  gen mal_trt_nonsev_drug = (coart1_pres == "Y" | quinm1_pres == "Y") &
      quinl1_pres == "N"
117.  gen mal_trt_nonsev_route = coart1_route == "po" | quinm1_route == "po"
118.  gen mal_trt_nonsev_dose = (((coart1_dose == "1" & i_weight >= 5 &
      i_weight < 15) | (coart1_dose == "2" & i_weight >= 15 & i_weight < 25) |
      (coart1_dose == "3" & i_weight >= 25 & i_weight < 35) | (coart1_dose ==
      "4" & i_weight >= 35)) & coart1_unit == "tabs") | (quinm1_dose <=
      10*i_weight & quinm1_dose >= 10*i_weight & quinm1_unit == "mg")
119.  gen mal_trt_nonsev_freq = coart1_freq == "2" | quinm1_freq == 3
120.  gen mal_trt_nonsev_dur = (coart1_days != "E" | coart1_days != "") |
      (quinm1_days != "E" | quinm1_days != "")
121.  replace mal_trt_sev_route = 1 if mal_trt_nonsev_drug &
      mal_trt_nonsev_route
122.  replace mal_trt_nonsev_route = 1 if mal_trt_sev_drug & mal_trt_sev_route
123.  replace mal_trt_sev_dose = 1 if mal_trt_nonsev_drug &
      mal_trt_nonsev_dose
124.  replace mal_trt_nonsev_dose = 1 if mal_trt_sev_drug & mal_trt_sev_dose
125.  replace mal_trt_sev_freq = 1 if mal_trt_nonsev_drug &
      mal_trt_nonsev_freq
126.  replace mal_trt_nonsev_freq = 1 if mal_trt_sev_drug & mal_trt_sev_freq
127.  replace mal_trt_sev_dur = 1 if mal_trt_nonsev_drug & mal_trt_nonsev_dur
128.  replace mal_trt_nonsev_dur = 1 if mal_trt_sev_drug & mal_trt_sev_dur
129.  gen mal_trt_drug = (mal_diag_sev_clin==1 & mal_trt_sev_drug==1) |
      (mal_diag_nonsev_clin==1 & mal_trt_nonsev_drug==1)
130.  gen mal_trt_route = (mal_diag_sev_clin==1 & mal_trt_sev_route==1) |
      (mal_diag_nonsev_clin==1 & mal_trt_nonsev_route==1)
131.  gen mal_trt_dose = (mal_diag_sev_clin==1 & mal_trt_sev_dose==1) |
      (mal_diag_nonsev_clin==1 & mal_trt_nonsev_dose==1)
132.  gen mal_trt_freq = (mal_diag_sev_clin==1 & mal_trt_sev_freq==1) |
      (mal_diag_nonsev_clin==1 & mal_trt_nonsev_freq==1)
133.  gen mal_trt_dur = (mal_diag_sev_clin==1 & mal_trt_sev_dur==1) |
      (mal_diag_nonsev_clin==1 & mal_trt_nonsev_dur==1)
134.  gen mal_trt_use = (mal_trt_sev_route==1 & mal_trt_sev_dose==1 &
      mal_trt_sev_freq==1 & mal_trt_sev_dur==1) | (mal_trt_nonsev_route==1 &
      mal_trt_nonsev_dose==1 & mal_trt_nonsev_freq==1 & mal_trt_nonsev_dur==1)
135.  treatment score type 1 (sum of appropriate drug(s), dose, route,
      frequency and duration for severity classification)
136.  gen mal_trt_sum1 = 0
137.  replace mal_trt_sum1 = mal_trt_sev_drug + mal_trt_sev_route +
      mal_trt_sev_dose + mal_trt_sev_freq + mal_trt_sev_dur if
      mal_diag_sev_clin == 1// severe
138.  replace mal_trt_sum1 = mal_trt_nonsev_drug + mal_trt_nonsev_route +
      mal_trt_nonsev_dose + mal_trt_nonsev_freq + mal_trt_nonsev_dur if
      mal_diag_nonsev_clin == 1 // non-severe
139.  treatment score type 2 (appropriate drug + correct use)
140.  gen mal_trt_sum2 = mal_trt_drug + mal_trt_use
141.  *Pneumonia
```

```
142. gen pneum_trt_very_sev_drug = (oxy_order == "Y" &
     strmatch(clinician_oxy, "*linician*1*")) & (pen_pres == "Y") &
     (gent1_pres == "Y")

143. gen pneum_trt_very_sev_route = (pen1_route != "E" & pen1_route != "") &
     (gent1_route != "E" & gent1_route != "")

144. gen pneum_trt_very_sev_dose = (pen1_dose <= 50000*i_weight*1.2 &
     pen1_dose >= 50000*i_weight*0.8 & pen1_unit == "iu") & (gent1_dose <=
     7.5*i_weight*1.2 & gent1_dose >= 7.5*i_weight*0.8 & gent1_unit == "mg")

145. gen pneum_trt_very_sev_freq = (pen1_freq == 4) & (gent1_freq == 1) // &
     (oxy_prescr != "No detail" & oxy_prescr != "")

146. gen pneum_trt_very_sev_dur = (pen1_days != "E" & pen1_days != "") &
     (gent1_days != "E" & gent1_days != "")

147. gen pneum_trt_sev_drug = (pen_pres == "Y") & (gent1_pres != "Y")

148. gen pneum_trt_sev_route = pen1_route != "E" & pen1_route != ""

149. gen pneum_trt_sev_dose = pen1_dose <= 50000*i_weight*1.2 & pen1_dose >=
     50000*i_weight*0.8 & pen1_unit == "iu"

150. gen pneum_trt_sev_freq = pen1_freq == 4

151. gen pneum_trt_sev_dur = pen1_days != "E" & pen1_days != ""

152. gen pneum_trt_nonsev_drug = amox1_pres == "Y" | sept1_pres == "Y"

153. gen pneum_trt_nonsev_route = (amox1_pres == "Y" & amox1_route == "po") |
     (sept1_pres == "Y" & sept1_route == "po")

154. gen pneum_trt_nonsev_dose = (amox1_pres == "Y" & ((amox1_dose <=
     25*i_weight*1.2 & amox1_dose >= 25*i_weight*0.8 & amox1_unit == "mg") |
     (((amox1_dose == 5 & i_weight < 7) | (amox1_dose == 7.5 & (i_weight >= 7
     & i_weight < 10)) | (amox1_dose == 10 & (i_weight >= 10 & i_weight <
     14)) | (amox1_dose == 15 & (i_weight >= 14 & i_weight < 19)) |
     (amox1_dose == 20 & (i_weight >= 19 & i_weight < 21))) & amox1_unit ==
     "mls") | (((amox1_dose == 0.5 & i_weight < 9) | (amox1_dose == 1 &
     (i_weight >= 9 & i_weight < 14)) | (amox1_dose == 2 & (i_weight >= 14 &
     i_weight < 21))) & (amox1_unit == "tabs" | amox1_unit == "caps")))) |
     (sept1_pres == "Y" & ((sept1_dose <= 24*i_weight*1.2 & sept1_dose >=
     24*i_weight*0.8 & sept1_unit == "mg") | (((sept1_dose == 2.5 & (i_weight
     >= 2 & i_weight < 4)) | (sept1_dose == 5 & (i_weight >= 4 & i_weight <
     11)) | (sept1_dose == 7.5 & (i_weight >= 11 & i_weight < 16)) |
     (sept1_dose == 10 & (i_weight >= 16 & i_weight < 21))) & sept1_unit ==
     "mls") | (((sept1_dose == 0.25 & (i_weight >= 2 & i_weight < 4)) |
     (sept1_dose == 0.5 & (i_weight >= 4 & i_weight < 16)) | (sept1_dose == 1
     & (i_weight >= 16 & i_weight < 21))) & (sept1_unit == "tabs" |
     sept1_unit == "caps"))))

155. gen pneum_trt_nonsev_freq = (amox1_pres == "Y" & amox1_freq == 3) |
     (sept1_pres == "Y" & sept1_freq == 2) // recheck this

156. gen pneum_trt_nonsev_dur = (amox1_pres == "Y" & amox1_days != "E" &
     amox1_days != "") | (sept1_pres == "Y" & sept1_days != "E" & sept1_days
     != "")

157. replace pneum_trt_very_sev_route = 1 if (pneum_trt_sev_drug &
     pneum_trt_sev_route) | (pneum_trt_nonsev_drug & pneum_trt_nonsev_route)

158. replace pneum_trt_sev_route = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_nonsev_drug &
     pneum_trt_nonsev_route)

159. replace pneum_trt_nonsev_route = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_sev_drug & pneum_trt_sev_route)

160. replace pneum_trt_very_sev_dose = 1 if (pneum_trt_sev_drug &
     pneum_trt_sev_route) | (pneum_trt_nonsev_drug & pneum_trt_nonsev_dose)

161. replace pneum_trt_sev_dose = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_nonsev_drug &
     pneum_trt_nonsev_dose)
```

```
162. replace pneum_trt_nonsev_dose = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_sev_drug & pneum_trt_sev_dose)

163. replace pneum_trt_very_sev_freq = 1 if (pneum_trt_sev_drug &
     pneum_trt_sev_route) | (pneum_trt_nonsev_drug & pneum_trt_nonsev_freq)

164. replace pneum_trt_sev_freq = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_nonsev_drug &
     pneum_trt_nonsev_freq)

165. replace pneum_trt_nonsev_freq = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_sev_drug & pneum_trt_sev_freq)

166. replace pneum_trt_very_sev_dur = 1 if (pneum_trt_sev_drug &
     pneum_trt_sev_route) | (pneum_trt_nonsev_drug & pneum_trt_nonsev_dur)

167. replace pneum_trt_sev_dur = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_nonsev_drug &
     pneum_trt_nonsev_dur)

168. replace pneum_trt_nonsev_dur = 1 if (pneum_trt_very_sev_drug &
     pneum_trt_very_sev_route) | (pneum_trt_sev_drug & pneum_trt_sev_dur)

169. gen pneum_trt_drug = (pneum_diag_very_sev_clin==1 &
     pneum_trt_very_sev_drug==1) | (pneum_diag_sev_clin==1 &
     pneum_trt_sev_drug==1) | (pneum_diag_nonsev_clin==1 &
     pneum_trt_nonsev_drug==1)

170. gen pneum_trt_route = (pneum_diag_very_sev_clin==1 &
     pneum_trt_very_sev_route==1) | (pneum_diag_sev_clin==1 &
     pneum_trt_sev_route==1) | (pneum_diag_nonsev_clin==1 &
     pneum_trt_nonsev_route==1)

171. gen pneum_trt_dose = (pneum_diag_very_sev_clin==1 &
     pneum_trt_very_sev_dose==1) | (pneum_diag_sev_clin==1 &
     pneum_trt_sev_dose==1) | (pneum_diag_nonsev_clin==1 &
     pneum_trt_nonsev_dose==1)

172. gen pneum_trt_freq = (pneum_diag_very_sev_clin==1 &
     pneum_trt_very_sev_freq==1) | (pneum_diag_sev_clin==1 &
     pneum_trt_sev_freq==1) | (pneum_diag_nonsev_clin==1 &
     pneum_trt_nonsev_freq==1)

173. gen pneum_trt_dur = (pneum_diag_very_sev_clin==1 &
     pneum_trt_very_sev_dur==1) | (pneum_diag_sev_clin==1 &
     pneum_trt_sev_dur==1) | (pneum_diag_nonsev_clin==1 &
     pneum_trt_nonsev_dur==1)

174. gen pneum_trt_use = (pneum_trt_very_sev_route==1 &
     pneum_trt_very_sev_dose==1 & pneum_trt_very_sev_freq==1 &
     pneum_trt_very_sev_dur==1) | (pneum_trt_sev_route==1 &
     pneum_trt_sev_dose==1 & pneum_trt_sev_freq==1 & pneum_trt_sev_dur==1) |
     (pneum_trt_nonsev_route==1 & pneum_trt_nonsev_dose==1 &
     pneum_trt_nonsev_freq==1 & pneum_trt_nonsev_dur==1)

175. treatment score type 1 (sum of appropriate drug(s), dose, route,
     frequency and duration for severity classification)

176. gen pneum_trt_sum1 = 0

177. replace pneum_trt_sum1 = pneum_trt_very_sev_drug +
     pneum_trt_very_sev_route + pneum_trt_very_sev_dose +
     pneum_trt_very_sev_freq + pneum_trt_very_sev_dur if
     pneum_diag_very_sev_clin == 1

178. replace pneum_trt_sum1 = pneum_trt_sev_drug + pneum_trt_sev_route +
     pneum_trt_sev_dose + pneum_trt_sev_freq + pneum_trt_sev_dur if
     pneum_diag_sev_clin == 1

179. replace pneum_trt_sum1 = pneum_trt_nonsev_drug + pneum_trt_nonsev_route
     + pneum_trt_nonsev_dose + pneum_trt_nonsev_freq + pneum_trt_nonsev_dur
     if pneum_diag_nonsev_clin == 1

180. treatment score type 2 (appropriate drug + correct use)
```

```
181.  gen pneum_trt_sum2 = pneum_trt_drug + pneum_trt_use

182.  *Diarrhoea/dehydration

183.  gen dnd_trt_shock_drug = fluid_pres1 == "nsal" | fluid_pres1 == "hs
      darr"

184.  gen dnd_trt_shock_dose = (total_vol1/(fluid_time1*4)) <= 20*i_weight*1.2
      & (total_vol1/(fluid_time1*4)) >= 20*i_weight*0.8

185.  gen dnd_trt_shock_freq = fluid_time1/4 >= 1

186.  gen dnd_trt_sev_drug = fluid_pres1 == "ring" | fluid_pres2 == "ORS"

187.  gen dnd_trt_sev_dose = (((total_vol1/fluid_time1) <= (((30*i_weight*1.2)
      + (70*i_weight*1.2))/3) & (total_vol1/fluid_time1) >=
      (((30*i_weight*0.8) + (70*i_weight*0.8))/3)) & age_years !=.) |
      (((total_vol1/fluid_time1) <= (((30*i_weight*1.2) +
      (70*i_weight*1.2))/6) & (total_vol1/fluid_time1) >= (((30*i_weight*0.8)
      + (70*i_weight*0.8))/6)) & age_years ==.) | (total_vol2/fluid_time2 <=
      (100*i_weight*1.2/6) & total_vol2/fluid_time2 >= (100*i_weight*0.8/6))

188.  gen dnd_trt_sev_freq = step1_2 == "Y"

189.  replace fluid_pres2 = "hs darr" if fluid_pres2 == "HS Darr" |
      fluid_pres2 == "HSD"

190.  replace fluid_pres2 = "resomal" if fluid_pres2 == "RESOMAL"

191.  replace fluid_pres2 = "nsal" if fluid_pres2 == "Nsal"

192.  gen dnd_trt_some_drug = fluid_pres2 == "ORS"

193.  gen dnd_trt_some_dose = (total_vol2/fluid_time2 <= (75*i_weight*1.2/4) &
      total_vol2/fluid_time2 >= (75*i_weight*0.8/4))

194.  gen dnd_trt_some_freq = fluid_time2/24 >= 1

195.  gen dnd_trt_none_drug = fluid_pres2 == "ORS"

196.  gen dnd_trt_none_dose = vol_stool <= 10*i_weight*1.2 & vol_stool >=
      10*i_weight*0.8

197.  gen dnd_trt_none_freq = vol_stool !=.

198.  replace dnd_trt_shock_dose = 1 if (dnd_trt_sev_drug & dnd_trt_sev_dose)
      | (dnd_trt_some_drug & dnd_trt_some_dose) | (dnd_trt_none_drug &
      dnd_trt_none_dose)

199.  replace dnd_trt_sev_dose = 1 if (dnd_trt_shock_drug &
      dnd_trt_shock_dose) | (dnd_trt_some_drug & dnd_trt_some_dose) |
      (dnd_trt_none_drug & dnd_trt_none_dose)

200.  replace dnd_trt_some_dose = 1 if (dnd_trt_shock_drug &
      dnd_trt_shock_dose) | (dnd_trt_sev_drug & dnd_trt_sev_dose) |
      (dnd_trt_none_drug & dnd_trt_none_dose)

201.  replace dnd_trt_none_dose = 1 if (dnd_trt_shock_drug &
      dnd_trt_shock_dose)| (dnd_trt_sev_drug & dnd_trt_sev_dose) |
      (dnd_trt_some_drug & dnd_trt_some_dose)

202.  replace dnd_trt_shock_freq = 1 if (dnd_trt_sev_drug & dnd_trt_sev_freq)
      | (dnd_trt_some_drug & dnd_trt_some_freq) | (dnd_trt_none_drug &
      dnd_trt_none_freq)

203.  replace dnd_trt_sev_freq = 1 if (dnd_trt_shock_drug &
      dnd_trt_shock_freq) | (dnd_trt_some_drug & dnd_trt_some_freq) |
      (dnd_trt_none_drug & dnd_trt_none_freq)

204.  replace dnd_trt_some_freq = 1 if (dnd_trt_shock_drug &
      dnd_trt_shock_freq) | (dnd_trt_sev_drug & dnd_trt_sev_freq) |
      (dnd_trt_none_drug & dnd_trt_none_freq)

205.  replace dnd_trt_none_freq = 1 if (dnd_trt_shock_drug &
      dnd_trt_shock_freq)| (dnd_trt_sev_drug & dnd_trt_sev_freq) |
      (dnd_trt_some_drug & dnd_trt_some_freq)

206.  gen dnd_trt_drug = (dnd_diag_shock_clin==1 & dnd_trt_shock_drug==1) |
```

```
      (dnd_diag_sev_clin==1 & dnd_trt_sev_drug==1) | (dnd_diag_some_clin==1 &
      dnd_trt_some_drug==1) | (dnd_diag_none_clin==1 & dnd_trt_none_drug==1)

207.  gen dnd_trt_dose = (dnd_diag_shock_clin==1 & dnd_trt_shock_dose==1) |
      (dnd_diag_sev_clin==1 & dnd_trt_sev_dose==1) | (dnd_diag_some_clin==1 &
      dnd_trt_some_dose==1) | (dnd_diag_none_clin==1 & dnd_trt_none_dose==1)

208.  gen dnd_trt_freq = (dnd_diag_shock_clin==1 & dnd_trt_shock_freq==1) |
      (dnd_diag_sev_clin==1 & dnd_trt_sev_freq==1) | (dnd_diag_some_clin==1 &
      dnd_trt_some_freq==1) | (dnd_diag_none_clin==1 & dnd_trt_none_freq==1)

209.  gen dnd_trt_use = (dnd_trt_shock_dose==1 & dnd_trt_shock_freq==1) |
      (dnd_trt_sev_dose==1 & dnd_trt_sev_freq==1) | (dnd_trt_some_dose==1 &
      dnd_trt_some_freq==1) | (dnd_trt_none_dose==1 & dnd_trt_none_freq==1)

210.  *replace dnd_trt_none_dur = 1 if

211.  treatment score type 1 (sum of appropriate drug(s), dose, route,
      frequency and duration for severity classification)

212.  gen dnd_trt_sum1 = 0

213.  replace dnd_trt_sum1 = dnd_trt_shock_drug + dnd_trt_shock_dose +
      dnd_trt_shock_freq if dnd_diag_shock_clin == 1

214.  replace dnd_trt_sum1 = dnd_trt_sev_drug + dnd_trt_sev_dose +
      dnd_trt_sev_freq if dnd_diag_sev_clin == 1

215.  replace dnd_trt_sum1 = dnd_trt_some_drug + dnd_trt_some_dose +
      dnd_trt_some_freq if dnd_diag_some_clin == 1

216.  replace dnd_trt_sum1 = dnd_trt_none_drug + dnd_trt_none_dose +
      dnd_trt_none_freq if dnd_diag_none_clin == 1

217.  treatment score type 2 (appropriate drug + correct use)

218.  gen dnd_trt_sum2 = dnd_trt_drug + dnd_trt_use

219.  foreach score_part in assess_1_primary assess_1_secondary
      assess_1_complete diag_any_class diag_class_corr trt_drug trt_use {

220.  gen mpd1_`score_part' = 0

221.  replace mpd1_`score_part' = mal_`score_part' if mal_diag_id_clin == 1 &
      pneum_diag_id_clin == 0 & dnd_diag_id_clin == 0

222.  replace mpd1_`score_part' = pneum_`score_part' if mal_diag_id_clin == 0
      & pneum_diag_id_clin == 1 & dnd_diag_id_clin == 0

223.  replace mpd1_`score_part' = dnd_`score_part' if mal_diag_id_clin == 0 &
      pneum_diag_id_clin == 0 & dnd_diag_id_clin == 1

224.  replace mpd1_`score_part' = (mal_`score_part' + pneum_`score_part')/2 if
      mal_diag_id_clin == 1 & pneum_diag_id_clin == 1 & dnd_diag_id_clin == 0

225.  replace mpd1_`score_part' = (pneum_`score_part' + dnd_`score_part')/2 if
      mal_diag_id_clin == 0 & pneum_diag_id_clin == 1 & dnd_diag_id_clin == 1

226.  replace mpd1_`score_part' = (mal_`score_part' + dnd_`score_part')/2 if
      mal_diag_id_clin == 1 & pneum_diag_id_clin == 0 & dnd_diag_id_clin == 1

227.  replace mpd1_`score_part' = (mal_`score_part' + pneum_`score_part' +
      dnd_`score_part')/3 if mal_diag_id_clin == 1 & pneum_diag_id_clin == 1 &
      dnd_diag_id_clin == 1

228.  }

229.  *mpd - combined scores types 1 and 2

230.  foreach score_part in assess_1_primary assess_1_secondary
      assess_1_complete diag_any_class diag_class_corr trt_drug trt_use {

231.  gen mpd2_`score_part' = (mal_`score_part' == 1 & mal_diag_id_clin == 1 &
      pneum_diag_id_clin == 0 & dnd_diag_id_clin == 0) | (pneum_`score_part'
      == 1 & mal_diag_id_clin == 0 & pneum_diag_id_clin == 1 &
      dnd_diag_id_clin == 0) | (dnd_`score_part' == 1 & mal_diag_id_clin == 0
      & pneum_diag_id_clin == 0 & dnd_diag_id_clin == 1) | (mal_`score_part'
      == 1 & pneum_`score_part' == 1 & mal_diag_id_clin == 1 &
```

```
      pneum_diag_id_clin == 1 & dnd_diag_id_clin == 0) | (pneum_`score_part'
      == 1 & dnd_`score_part' == 1 & pneum_diag_id_clin == 1 &
      dnd_diag_id_clin == 1 & mal_diag_id_clin == 0) | (mal_`score_part' == 1
      & dnd_`score_part' == 1 & mal_diag_id_clin == 1 & dnd_diag_id_clin == 1
      & pneum_diag_id_clin == 0) | (mal_`score_part' == 1 & pneum_`score_part'
      == 1 & dnd_`score_part' == 1 & mal_diag_id_clin == 1 &
      pneum_diag_id_clin == 1 & dnd_diag_id_clin == 1)
232.  }
233.  gen disease_count = mal_diag_id_clin + pneum_diag_id_clin +
      dnd_diag_id_clin
234.  foreach mpd_score in 1 2 {
235.  gen mpd`mpd_score'_diag_id_clin = disease_count > 0
236.  gen mpd`mpd_score'_trt_sum2 = mpd`mpd_score'_trt_drug +
      mpd`mpd_score'_trt_use
237.  }
238.  foreach disease in mal pneum dnd mpd1 mpd2 {
239.  gen `disease'_assess_sum2 = `disease'_assess_1_primary +
      `disease'_assess_1_secondary + `disease'_assess_1_complete
240.  gen `disease'_assess_sum3 = `disease'_assess_1_primary +
      `disease'_assess_1_secondary
241.  if "`disease'" != "mpd1" & "`disease'" != "mpd2" {
242.  gen `disease'_sum1 = `disease'_assess_sum1 + `disease'_diag_any_class +
      `disease'_trt_sum1
243.  }
244.  gen `disease'_sum2 = `disease'_assess_sum2 + `disease'_diag_any_class +
      `disease'_trt_sum2
245.  gen `disease'_sum3 = `disease'_assess_sum3 + `disease'_diag_any_class +
      `disease'_trt_sum2
246.  gen `disease'_sum4 = `disease'_assess_sum3 + `disease'_diag_any_class +
      `disease'_diag_class_corr + `disease'_trt_sum2
247.  }
```

Estimating clinician- and hospital level agreement between the basic and modified scores

```
1.    foreach disease in mal pneum dnd {
2.          use datasets/scores_dhs, clear
3.          cap drop hospital
4.          encode clinician1, gen(clinician)
5.          encode hosp_id, gen(hospital)
6.          bootstrap "do bootstrap_rho.do `disease'" ((hosp_var_null[1,1] -
            hosp_var_alt[1,1]) / hosp_var_null[1,1]), reps(100)
            saving(bootstrap_rho_hosp_`disease') replace
7.          bootstrap "do bootstrap_rho.do `disease'" ((clin_var_null[1,1] -
            clin_var_alt[1,1]) / clin_var_null[1,1]), reps(100)
            saving(bootstrap_rho_clin_`disease') replace
8.    }
9.    foreach disease in mal pneum dnd {
10.         use datasets/scores_dhs, clear
```

```
11.         cap drop hospital
12.         keep if `disease'_diag_id_clin == 1 & survey_no < 5
13.         encode hosp_id, gen(hospital)
14.         encode clinician1, gen(clinician)
15.         bootstrap "do bootstrap_pred_prob.do `disease'" (r(mean)),
            reps(100) saving(bootstrap_pred_prob_`disease') replace
16.   }
17.   foreach disease in mal pneum dnd {
18.         use datasets/scores_dhs, clear
19.         keep if `disease'_diag_id_clin == 1 & survey_no < 5
20.         bootstrap "do bootstrap_pred_prob_gologit.do `disease'" (1 -
            (ss_error[1,1]/ss_total[1,1])), reps(100)
            saving(bootstrap_pred_prob_gologit_`disease') replace
21.   }
22.   foreach disease in mal pneum dnd {
23.         use datasets/scores_dhs, clear
24.         cap drop hospital
25.         encode clinician1, gen(clinician)
26.         encode hosp_id, gen(hospital)
27.         bootstrap "do bootstrap_rho_gllamm.do `disease'"
            ((var_null_g[2,2]^2 - var_alt_g[2,2]^2) / var_null_g[2,2]^2),
            reps(10) saving(bootstrap_rho_gllamm_hosp_`disease') replace
28.         bootstrap "do bootstrap_rho_gllamm.do `disease'"
            ((var_null_g[1,1]^2 - var_alt_g[1,1]^2) / var_null_g[1,1]^2),
            reps(10) saving(bootstrap_rho_gllamm_clin_`disease') replace
29.   }
30.   (bootstrap_pred_prob_gologit.do)
31.   foreach var in s_pred agree2 sq_err sq_tot {
32.         cap drop `var'
33.         gen `var' = .
34.   }
35.   foreach var in s0 s1 s2 s3 s4 s5 s6 s_max {
36.         cap drop `var'
37.   }
38.   qui gologit2 `1'_sum2 `1'_sum1
39.   predict s0 s1 s2 s3 s4 s5 s6
40.   egen s_max = rowmax(s0-s6)
41.   local loop = 0
42.         while `loop' < 7 {
43.               foreach s_prob of varlist s0 - s6 {
44.                         replace s_pred = `loop' if `s_prob' == s_max
45.                         local loop = `loop' + 1
46.               }
47.         }
48.   replace agree2 = (`1'_sum2 == s_pred)
```

```
49.   replace sq_err = (s_pred - `1'_sum2)^2
50.   qui mean `1'_sum2
51.   matrix means = e(b)
52.   replace sq_tot = (`1'_sum2 - (means[1,1]))^2
53.   qui tabstat sq_tot, stats(sum) save
54.   qui matrix ss_total = r(StatTotal)
55.   qui tabstat sq_err, stats(sum) save
56.   qui matrix ss_error = r(StatTotal)
```

bootstrap_pred_prob.do *for stimating clinician- and hospital level agreement between the basic and modified scores*

```
1.    foreach var in `1'_sum2_pred `1'_sum2_pred_round agree1 {
2.          cap drop `var'
3.          gen `var' = .
4.    }
5.    cap drop u
6.    gen u = runiform()
7.    qui xtmixed `1'_sum2 `1'_sum1 if u > 0.5 || hospital: || clinician:
8.    matrix coefficients = e(b)
9.    replace `1'_sum2_pred = ((coefficients[1,1]) * `1'_sum1) +
      (coefficients[1,2]) if u < = 0.5
10.   replace `1'_sum2_pred_round = round(`1'_sum2_pred, 1)
11.   replace agree1 = (`1'_sum2 == `1'_sum2_pred_round)
12.   replace agree1 = . if u > 0.5
13.   summ agree1
```

bootstrap_rho.do *for stimating clinician- and hospital level agreement between the basic and modified scores*

```
1.    qui xtmixed `1'_sum1 || hospital: || clinician: if `1'_diag_id_clin ==
      1 & survey_no < 5, var
2.    estat recovariance, level(hospital)
3.    matrix hosp_var_null = r(cov)
4.    estat recovariance, level(clinician)
5.    matrix clin_var_null = r(cov)
6.    qui xtmixed `1'_sum1 `1'_sum2 || hospital: || clinician: if
      `1'_diag_id_clin == 1 & survey_no < 5, var
7.    estat recovariance, level(hospital)
8.    matrix hosp_var_alt = r(cov)
9.    estat recovariance, level(clinician)
10.   matrix clin_var_alt = r(cov)
```

`bootstrap_rho_gllamm.do` *for stimating clinician- and hospital level agreement between the basic and modified scores*

```
1.    gllamm `1'_sum1 if `1'_diag_id_clin == 1 & survey_no < 5, i(clinician
      hospital) family(binomial) link(ologit) adapt
2.    matrix var_null_g = e(chol)
3.    gllamm `1'_sum1 `1'_sum2 if `1'_diag_id_clin == 1 & survey_no < 5,
      i(clinician hospital) family(binomial) link(ologit) adapt
4.    matrix var_alt_g = e(chol)
```

Structural equation modelling with Mplus™ (input and output piped through Stata™)

```
1.    foreach disease in mal pneum dnd mpd2 {
2.          tempname cfa_mod2_`disease'_inp
3.          file open `cfa_mod2_`disease'_inp' using
            "cfa_mod2_`disease'.inp", write replace
4.          #delimit ;
5.               foreach line in
5.                        "TITLE:"
6.                        "Factor Analysis with Categorical Outcome
                          Variables;"
7.                        "DATA:"
8.                        "FILE IS cfa_mod_`disease'.raw;"
9.                        "VARIABLE:"
10.                       "NAMES ARE id hosp a_1 a_2 d_1 t_1 t_2 group
                          survey;"
11.                       "USEVARIABLES ARE a_1 d_1 t_1 t_2 group survey;"
12.                       "IDVARIABLE IS id;"
13.                       "AUXILIARY = hosp;"
14.                       "CATEGORICAL ARE a_1 d_1 t_1 t_2;"
15.                       "ANALYSIS:"
16.                       "MODEL:"
17.                       "assess BY a_1 d_1;"
18.                       "treat BY t_1 t_2;"
19.                       "assess ON group survey;"
20.                       "treat ON group survey;"
21.                       "assess WITH treat;"
22.                       "OUTPUT:" "standardized modindices tech1;"
23.                       "SAVEDATA:"
24.                       "file is cfa_mod2_`disease'_scores.dat;"
25.                       "save = fscores;"
26.              {
```

```
27.              ;
28.         #delimit cr
29.              file write `cfa_mod2_`disease'_inp' "`line'" _n
30.              }
31.         file close `cfa_mod2_`disease'_inp'
32.         shell wine mplus.exe cfa_mod2_`disease'.inp
33.         type cfa_mod2_`disease'.out
34.   }
```

Generating funnel plots for comparing hospitals

```
1.    foreach distribution in 1 2 {
2.          preserve
3.          keep if `disease'_diag_id_clin == 1 & disease_count > 0 &
            hospital > 0
4.          if `distribution' == 1 {   // normal distribution
5.               collapse (mean) mean_score=`disease'_sum2 (semean)
                 semean_score=`disease'_sum2 (count) n_score=`disease'_sum2,
                 by (hospital survey_no group)
6.               gen sd_score = semean_score * sqrt(n_score)
7.          }
8.          if `distribution' == 2 {   // Poisson-binomial distribution
9.               collapse (mean) p1=`disease'_assess_1_primary
                 p2=`disease'_assess_1_secondary
                 p3=`disease'_assess_1_complete p4=`disease'_diag_any_class
                 p5=`disease'_trt_drug p6=`disease'_trt_use (count)
                 n_score=`disease'_sum2, by (hospital survey_no group)
10.              gen mean_score = p1 + p2 + p3 + p4 + p5 + p6
11.              gen sd_score = sqrt((p1*(1-p1))+(p2*(1-p2))+(p3*(1-
                 p3))+(p4*(1-p4))+(p5*(1-p5))+(p6*(1-p6)))
12.         }
13.         cap funnelcompar mean_score n_score hospital sd_score, contours(5
            0.2) continuous markall vertical legend(off) ytitle("Precision
            (n)") xtitle("Mean score")
14.         restore
15.   }
```

Bibliography

Adeboye, MA., Ojuawo, A., Ernest, SK., Fadeyi, A. and Salisu, OT. (2010). Mortality pattern within twenty-four hours of emergency paediatric admission in a resource-poor nation health facility. West Afr J Med *29*, 249-52.

Ajayi, IO., Falade, CO., Bamgboye, EA. et. al. (2008). Assessment of a treatment guideline to improve home management of malaria in children in rural south-west Nigeria. Malar J *7*, 24.

Akaike, H. (1974). A new look at the statistical model identification. IEE Transactions on Automatic Control *19*, 716-23.

Allison, P. (2014). Measures of fit for logistic regression - paper presented at the SAS Global Forum, March 25, Washington, DC.

Altman, DG. (1991). Practical statistics for medical research. London. Chapman & Hall.

Altman, DG. (2005). Categorizing Continuous Variables. Encyclopedia of Biostatistics.

Ashton, CM., Kuykendall, DH., Johnson, ML. et al. (1999). An empirical assessment of the validity of explicit and implicit process-of-care criteria for quality assessment. Med Care *37*, 798-808.

Ayieko, P., Ntoburi, S., Wagai, J. et al. (2011). A multifaceted intervention to implement guidelines and improve admission paediatric care in Kenyan district hospitals: a cluster randomised trial. PLoS Med. *8*, e1001018.

Bamm, EL., Rosenbaum, P. and Stratford, P. (2010). Validation of the measure of processes of care for adults: a measure of client-centred care. Int J Qual Health Care *22*, 302-309.

Barlow, DH. and Hersen, M. (1984). Single case experimental designs. Strategies for studying behaviour change. New York: Pergamon Press.

Barnhart, HX., Haber, MJ. and Lin, Li. (2007). An overview on assessing agreement with continuous measurements. J Biopharm Stat *17*, 529-69.

Bartlett, MS. (1950). Tests of significance in factor analysis. Br J Clin Psychol *3*, 77-85.

Beck, DH., Taylor, BL., Millar, B. and Smith, G.B. (1997). Prediction of outcome from intensive care: a prospective cohort study comparing Acute Physiology and Chronic Health Evaluation II and III prognostic systems in a United Kingdom intensive care unit. Crit Care Med *25*, 9-15.

Benneyan, JC., Lloyd, RC., and Plsek, PE. (2003). Statistical process control as a tool for research and healthcare improvement. Qual Saf Health Care, *12*, 458-464.

Bentler, PM. (1990). Comparative fit indexes in structural models. Psychol Bull *107*, 238-46.

Berkley, JA., Maitland, K., Mwangi, I., Ngetsa, C., Mwarumba, S., Lowe, BS., Newton, CR., Marsh, K., Scott, JA. and English M. (2005). Use of clinical syndromes to target antibiotic prescribing in seriously ill children in malaria endemic area: observational study. BMJ *330*, 995.

Berkley, JA., Ros, A., Mwangi, I., Osier, FHA., Mohammed, M., Shebe, M *et al.* (2003) Prognostic indicators of early and late deaths in children admitted to district hospital in Kenya: a cohort study. BMJ *326*, 361-7.

Biswas, P. and Kalbfleisch, JD. (2008). A risk-adjusted CUSUM in continuous time based on the Cox model. Stat Med *27*, 3382-406

Bjerrum, L., Munck, A., Gahrn-Hansen, B. et al. (2010). Health Alliance for Prudent Prescribing, Yield and Use of Antimicrobial Drugs in the Treatment of Respiratory Tract Infections (HAPPY AUDIT). BMC Fam Pract *11*, 29.

Black, RE., Cousens, S., Johnson, HL. et. al. (2010). Global, regional, and national causes of child mortality in 2008: a systematic analysis. Lancet *375*, 1969-87.

Bland, JM. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet *1*, 307-10.

Bland, JM. and Altman, DG. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med *20*, 337-40.

Botha, JL., Silcocks, PB., Bright, N. and Redgrave, P. (2001). Breast and cervical cancer survival: making sense of league tables. Public Health *115,* 165-72.

Brewer, M. (2000). Research design and issues of validity. In Reis, H. and Judd, C. (eds.) Handbook of research methods in social and personality psychology. Cambridge: Cambridge University Press.

Burckhardt, CS. and Anderson, KL. (2003). The Quality of Life Scale (QOLS): reliability, validity, and utilization. Health Qual Life Outcomes *1*, 60.

Campbel, JD., Sow, SO., Levine, MM. and Kotlof, KL. (2004) The causes of hospital admision and death among children in Bamako, Mali. J Trop Ped *50*:158-63.

Chanthong, P., Abrishami, A., Wong, J. et al. (2009). Systematic review of questionnaires measuring patient satisfaction in ambulatory anesthesia. Anesthesiology *110*, 1061-1067.

Chen, M., Wang, R., Cheng, C. et al. (2011). Diabetes Empowerment Process Scale: development and psychometric testing of the Chinese version. J Adv Nurs *67*, 204-214.

Chetter, IC., Spark, JI., Dolan, P. et. al. (1997). Quality of life analysis in patients with lower limb ischaemia: suggestions for European standardisation. Eur J Vasc Endovasc Surg *13*, 597-604.

Chevat, C., Viala-Danten, M., Dias-Barbosa, C. et al. (2009). Development and psychometric validation of a self-administered questionnaire assessing the acceptance of influenza vaccination: the Vaccinees' Perception of Injection (VAPI) questionnaire. Health Qual Life Outcomes *7*, 21.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educ Psychol Meas *20*, 37-46

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull *70*, 213-220.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. Psychol Bull *112*, 155-159.

Colman, AM., Norris, CE. and Preston, CC. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. Psychol Rep *80*, 355-362.

Commey, JO., Rodrigues, OP., Akita, FA. and Newman, M. (1994). Bacterial meningitis in children in southern Ghana. East Afr Med J *71*, 113-7.

Cooper, RA., Getzen, TE., McKee, HJ. and Laud, P. (2002). Economic and demographic trends signal an impending physician shortage. Health Affairs *21*:140-54.

Corfield, AR., Graham, CA., Adams, JN. et al. (2004). Emergency department thrombolysis improves door to needle times. Emerg Med J *21*, 676-680.

Costello, AB. and Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical Assessment Research & Evaluation. Available online: http://pareonline.net/getvn.asp?v=10&n=7

Couto, TB., Farhat, SC., Reid, T., Schvartsman, C. (2013). Mortality in a paediatric secondary care hospital in post-conflict Liberia in 2009. Einstein *11*, 413-20.

Cozby, PC. (1993). Methods in Behavioral Research. Mountain View, California: Mayfield.

Cudeck, R. and Henly, SJ. (1991). Model selection in covariance structures analysis and the 'problem' of sample size: a clarification. Psychol Bull *109*, 512-519.

Cutler, DM., Meara, E. (2001). Changes in the age distribution of mortality over the 20th century. National Bureau of Economic Research. [cited 12 April 2012]. Available from http://www.nber.org/papers/w8556

De Maeseneer, JM. and De Sutter, A. (2004). Why research in family medicine? A superfluous question. Ann Fam Med *2 Suppl 2*, S17-22.

DeCoster, J. (1998). Overview of factor analysis. Retrieved February 22 2013 from http://www.stat-help.com/notes.html.

Doherty, RE., Keller, EG. (1949). Mathematics of modern engineering. John Wiley & Sons, Inc. Vol I, p 58-60 [cited 30 July 2014]. Available from http://www.archive.org/stream/mathematicsofmod029509mbp#page/n83/mode/1up/search/laplace.

Donabedian, A. (1966). Evaluating the quality of medical care. Milbank Memorial Fund Quarterly *44*, 166-206.

Donabedian, A. (1988). The quality of care. How can it be assessed?. JAMA *260*, 1743-1748.

Donabedian, A. (2005). Evaluating the quality of medical care. 1966. Milbank Q *83*, 691-729.

Doyle, C., Reed, J., Woodcock, T. et al. (2010). Understanding what matters to patients - identifying key patients' perceptions of quality. JRSM Short Rep *1*, 3.

Duke, T., Tamburlini, G. and Silimperi, D. (2003). Improving the quality of paediatric care in peripheral hospitals in developing countries. Arch Dis Child *88*, 563-5.

Eldridge, SM., Ashby, D. and Kerry, S. (2006). Sample size for cluster randomised trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol *35*, 1292-300

English, M., Berkley, J., Mwangi, I. et. al. (2003). Hypothetical performance of syndrome-based management of acute paediatric admissions of children aged more than 60 days in a Kenyan district hospital. Bull World Health Organ *81*, 166-73.

English, M., Berkley, J., Mwangi, I., Mohammed, S., Ahmed, M., Osier, F., Muturi, N., Ogutu, B., Marsh, K. and Newton, CR. (2003). Hypothetical performance of syndrome-based management of acute paediatric admissions of children aged more than 60 days in a Kenyan district hospital. Bull World Health Organ *81*, 166-73.

English, M., Esamai, F., Wasunna, A. et al. (2004). Assessment of inpatient paediatric care in first referral level hospitals in 13 districts in Kenya. Lancet *363*, 1948-1953.

English, M., Wamae, A., Nyamai, R. et al. (2011). Implementing locally appropriate guidelines and training to improve care of serious illness in Kenyan hospitals: a story of scaling-up (and down and left and right). Arch. Dis. Child. *96*, 285-290.

Evans, JD. (1996). Straightforward statistics for the behavioural sciences. Pacific Grove, CA: Brooks/Cole Publishing.

Fan, X. and Sivo, SA. (2007). Sensitivity of model fit indices to model misspecification and model types. Multivar Behav Res *42*, 509-529.

Fernandez, R. and Fernandez, G. (2009). Validating the Bland-Altman method of agreement. Retrieved April 2 2013 from http://www.wuss.org/proceedings09/09WUSSProceedings/papers/pos/POS-Fernandez.pdf.

Fibrinolytic Therapy Trialists' (FTT) Collaborative Group (1994). Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Lancet *343*, 311-322.

Fleiss, J., Levin, B. and Paik, M. (2003). Statistical methods for rates and proportions, 3rd edition. New York. Wiley & Sons.

Forster, M., Bailey, C., Brinkhof, MWG. et. al. (2008). Electronic medical record systems, data quality and loss to follow-up: survey of antiretroviral therapy programmes in resource-limited settings. Bull World Health Organ *86*, 939-47.

Fraleigh, JB., Beauregard, RA. and Katz, VJ. (1994). Linear algebra, 3rd edition. Boston. Addison-Wesley-Longman.

Gathara, D., Opiyo, N., Wagai, J. et al. (2011). Quality of hospital care for sick newborns and severely malnourished children in Kenya: a two-year descriptive study in 8 hospitals. BMC Health Serv Res *11*, 307.

Ghiselli, EE., Campbell, JP. and Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco, California: Freeman & Company.

Gibberd, R., Hancock, S., Howley, P. and Richards, K. (2004). Using indicators to quantify the potential to improve the quality of health care. Int J Qual Health Care *16*, i37-43.

Gilson, BS., Gilson, JS., Bergner, M. et. al. (1975). The sickness impact profile. Development of an outcome measure of health care. Am J Public Health *65*, 1304-10.

Goldstein, H. and Speigelhalter, DJ. (1996). League tables and their limitations: statistical issues in comparison of institutional performance. [cited 27 June 2014]. Available from http://www.bristol.ac.uk/cmm/team/hg/limitations-of-league-tables.pdf.

Haessel, W. (1978). Measuring goodness of fit in linear and nonlinear models. Southern Econ J *44*, 648-652.

Hammermeister, KE., Shroyer, AL., Sethi, GK. and Grover, FL. (1995). Why it is important to demonstrate the linkages between outcomes of care and processes and structures of care. Med Care *33*, OS5-16.

Hannan, TJ., Rotich, JK., Odero, WW. et. al. (2000). The Mosoriot medical record system: design and initial implementation of an outpatient electronic record system in rural Kenya. Int J Med Inform *60*, 21-8.

Hannan, TJ., Tierney, WM., Rotich, JK. et. al. (2001). The MOSORIOT medical record system (MMRS) phase I to phase II implementation: an outpatient computer-based medical record system in rural Kenya. Stud Health Technol Inform *84*, 619-22.

Harris, AHS., Kivlahan, DR., Bowe, T. et al. (2009). Developing and validating process measures of health care quality: an application to alcohol use disorder treatment. Med Care *47*, 1244-1250.

Harutyunyan, T., Demirchyan, A., Thompson, ME. et al. (2010). Patient satisfaction with primary care in Armenia: good rating of bad services?. Health Serv Manage Res *23*, 12-17.

Hayes, RJ. and Bennett, S. (1999). Simple sample size calculation for cluster-randomised trials. Int J Epidemol *28*, 319-326

Hays, WL. (1994). Statistics, 5th edition. Fort Worth Texas. Harcourt-Brace.

Higgins, JPT. and Thompson, SG. (2002). Quantifying heterogeneity in a meta-analysis. Stat Med *21*, 1539-1558.

Higgins, JPT., Thompson, SG., Deeks, JJ. and Altman, DG. (2003). Measuring inconsistency in meta-analyses. BMJ *327*, 557-560.

Hooper, D., Coughlan, J. and Mullen, MR. (2008). Structural equation modelling: guidelines for determining model fit. Electronic Journal of Business Research Methods *6*, 53-60.

Hosmer DW, Lemeshow S (2000). Assessing the fit of the model. In: Applied Logistic Regression. John Wiley & Sons, New Jersey, pp 143-202.

Hox, JJ. and Maas, CJM. (2001). The accuracy of multilevel structural equation modelling with pseudobalanced groups and small samples. Struct. Equ. Modelling *8*, 157-174.

Hu, LT. and Bentler, PM. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modelling, *6*, 1-55.

Huicho, L., Scherpbier, RW., Nkowane, AM. et al. (2008). How much does quality of child care vary between health workers with differing durations of training? An observational multicountry study. Lancet *372*, 910-916.

Hunt, SM., McEwen, J. and McKenna, SP. (1985). Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract *35*, 185-8.

Hussey, P., Anderson, G., Berthelot, J. et al. (2008). Trends in socioeconomic disparities in health care quality in four countries. Int J Qual Health Care *20*, 53-61.

Irimu, G., Wamae, A., Wasunna, A. et al. (2008). Developing and introducing evidence based clinical practice guidelines for serious illness in Kenya. Arch. Dis. Child. *93*, 799-804.

Irimu, GW., Gathara, D., Zurovac, D.. et al. (2012). Performance of health workers in the management of seriously sick children at a Kenyan tertiary hospital: before and after a training intervention. PLoS One 7:e39964. doi: 10.1371/journal.pone.0039964.

Jacobs, R., Goddard, M. and Smith, PC. (2005). How robust are hospital ranks based on composite performance measures? Med Care *43*, 1177-84.

Jöreskog, KG. and Sörbom, D. (1988). LISREL 7: A guide to the program and applications. Chicago: SPSS.

Kang, SH., Ahn, CW. and Jung, SH. (2003). Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. Drug Inf J *37*,109-14.

Kihuba, E., Gathara, D., Mwinga, S., Mulaku, M., Kosgei, R., Mogoa, W., Nyamai, R. and English, M. (2014). Assessing the ability of health information systems in hospitals to support evidence-informed decisions in Kenya. Global Health Action *7,* 24859.

Klassen, AF., Dix, D., Cano, SJ. et al. (2009). Evaluating family-centred service in paediatric oncology with the measure of processes of care (MPOC-20). Child Care Health Dev *35*, 16-22.

Klevsgard, R., Froberg, B., Risberg, B. and Hallberg, IR. (2002). Nottingham Health Profile and Short-Form 36 Health Survey questionnaires in patients with chronic lower limb ischemia: before and after revascularization. J Vasc Surg *36*, 310-7.

Kline, P. (1994). An easy guide to factor analysis. New York. Routledge.

Kline, RB. (2005). Principles and Practice of Structural Equation Modelling. New York: The Guilford Press.

Kline, TJB. (1960). Psychological testing: a practical approach to design and evaluation. California: Sage Publications.

Knapp, TR. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nurs Res *39*, 121-123.

Knaus, WA., Draper, EA., Wagner, DP. and Zimmerman, JE. (1985). APACHE II: a severity of disease classification system. Crit Care Med *13*, 818-29.

Kupeli, N., Chilcot, J., Schmidt, UH., Campbell, IC. and Troop, NA. (2013). A confirmatory factor analysis and validation of the forms of self-criticism/reassurance scale. Br J Clin Psychol *52*, 12-25.

Lambrechts, T., Bryce, J. and Orinda, V. (1999). Integrated management of childhood illness: a summary of first experiences. Bull World Health Organ *77*, 582-94.

Landis, JR. and Koch, GG. (1977). The measurement of observer agreement for categorical data. Biometrics *33*, 159-74.

LaVecchia, C., Levi, F., Lucchini, F. et al. (1998). Trends in childhood cancer mortality as indicators of the quality of medical care in the developed world. Cancer *83*, 2223-2227.

Leatherman, S., Ferris, TG., Berwick, D. et al. (2010). The role of quality improvement in strengthening health systems in developing countries. Int J Qual Health Care *22*, 237-243.

Lei, P. and Wu, Q. (2007). Introduction to structural equation modelling: issues and practical considerations. Education Measurement: Issues and Practice *26*, 33-43.

Lilford, R. and Pronovost, P. (2010). Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. BMJ *340*, c2016.

Linden, AF., Sekidde, FS., Galukande, M., Knowlton, LM., Chackungal, S. And McQueen, KA. (2012). Challenges of surgery in developing countries: a survey of surgical and anesthesia capacity in Uganda's public hospitals. World J Surg *36*, 1056-65.

Liu, L., Johnson, HL., Cousens, S. et. al. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet *379*, 2151-61.

Llewellyn, SK., Hamdan, AM. and Rock, WP. (2007). An index of orthodontic treatment complexity. Eur J Orthod *29*, 186-192.

Mabey, D., Peeling, RW., Ustianowski, A. and Perkins, MD. (2004). Diagnostics for the developing world. Nat Rev Bicrobiol *2*, 231-40.

Macinko, J. and Starfield, B. (2002). Annotated Bibliography on Equity in Health, 1980-2001. Int J Equity Health *1*, 1.

Mael, FA., O'Shea, PG., Smith, MA. et al. (2010). Development of a model and measure of process-oriented quality of care for substance abuse treatment. J Behav Health Serv Res *37*, 4-24.

Maher, D. (1996). Clinical audit in a developing country. Trop. Med. Int. Health *1*, 409-413.

Maloney, K. and Chaiken, BP. (1999). An overview of outcomes research and measurement. J Healthc Qual *21*, 4-9; quiz 9-10, 60.

Mann, SL., Marshall, MR., Woodford, BJ., Holt, A. and Williams, AB. (2012). Predictive performance of Acute Physiological and Chronic Health Evaluation releases II to IV: a single New Zealand centre experience. Anaesth Intensive Care *40*, 479-89.

Marcinowicz, L., Chlabicz, S. and Grebowski, R. (2009). Patient satisfaction with healthcare provided by family doctors: primary dimensions and an attempt at typology. BMC Health Serv Res *9*, 63.

Markgraf, R., Deutschinoff, G., Pientka, L. and Scholten, T. (2000). Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. Crit Care Med *28*, 26-33.

McArdle, JJ. and McDonald, RP. (1984). Some algebraic properties of the reticular action model. Brit J Math Stat Psychol *37*, 234-251.

McDowell, I. and Newell, C. (2006) Measuring Health: a guide to rating scales and questionnaires. New York, NY: Oxford University Press.

Ministry of Health (2013). Quality of care in 22 level 4 and 5 hospitals.

Ministry of Health. (2006). Basic paediatric protocols. Nairobi. Government of Kenya Printing Press.

Ministry of Health. (2007). Basic paediatric protocols. Nairobi. Government of Kenya Printing Press.

Ministry of Health. (2010). Basic paediatric protocols. Nairobi. Government of Kenya Printing Press.

Minne, L., Abu-Hanna, A. and de Jonge, E. (2008). Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Crit Care *12*, R161.

Montgomery, DC. & Woodall, WH. (1999). Research issues and ideas in statistical process control. J Qual Technol, *31*, 376-387.

Moons, K., Kengne, A., Woodward, M. et al. (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart *doi:10.1136/heartjnl-2011-301246.*

Mosteller, F., Tukey JW. (1977). Data analysis and regression: a second course in statistics. Michigan: Addison-Wesley.

M*plus* (Version 5.1). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Murphy-Filkins, R., Teres, D., Lemeshow, S. and Hosmer, D.W. (1996). Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. Crit Care Med *24*, 1968-73.

Mwakyusa, S., Wamae, A., Wasunna, A. et. al. (2006). Implementation of a structured paediatric admission record for district hospitals in Kenya – results of a pilot study. BMC Int Health Hum Rights *6*, 9.

Nair, SN., Czaja, SJ. and Sharit, J. (2007). A multilevel modelling approach to examining individual differences in skill acquisition for a computer-based task. J Gerontol B Psychol Sci Soc Sci *62* Spec No 1, 85-96.

Najjar-Pellet, J., Jonquet, O., Jambou, P. et al. (2008). Quality assessment in intensive care units: proposal for a scoring system in terms of structure and process. Intensive Care Med *34*, 278-285.

National Coordinating Agency for Population and Development (NCAPD) [Kenya], Ministry of Medical Services (MOMS) [Kenya], Ministry of Public Health and Sanitation (MOPHS) [Kenya], Kenya National Bureau of Statistics (KNBS) [Kenya], ICF Macro (2011). Kenya Service Provision Assessment Survey 2010. Nairobi, Kenya: National Coordinating Agency for Population and Development, Ministry of Medical Services, Ministry of Public Health and Sanitation, Kenya National Bureau of Statistics, and ICF Macro.

National Institute of Clinical Excellence (2002). Principles of best practice in clinical audit.[cited 12 April 2012]. Available from http://www.scie-socialcareonline.org.uk/profile.asp?guid=116C2200-BA53-49F5-81FA-1FED8DA387B7.

Nolan, T., Angos, P., Cunha, AJ. et al. (2001). Quality of hospital care for seriously ill children in less-developed countries. Lancet *357*, 106-110.

Noyez, L. (2009). Control charts, CUSUM techniques and funnel plots. A review of methods for monitoring performance in healthcare. Interact Cardiovasc Thorac Surg *9*, 494-499.

Nshakira, N., Kristensen, M., Ssali, F. and Whyte, SR. (2002). Appropriate treatment of malaria? Use of antimalarial drugs for children's fevers in district medical units, drug shops and homes in eastern Uganda. Trop Med Int Health *7*, 309-16.

Ntoburi, S., Wagai, J., Irimu, G. et al. (2008). Debating the quality and performance of health systems at a global level is not enough, national debates are essential for progress. Trop. Med. Int. Health *13*, 444-447.

Nunnally, JC. and Bernstein, IH (1994). Psychometric theory, 3rd edition. New York. McGraw-Hill

Nye, BR., Hyde, CE., Tsivgoulis, G. et. al. (2012). Slim stroke scales for assessing patients with acute stroke: ease of use or loss of valuable assessment data?. Am J Crit Care *21*, 442-7; quiz 448.

Opondo, C., Ayieko, P., Ntoburi, S. et al. (2011). Effect of a multi-faceted quality improvement intervention on inappropriate antibiotic use in children with non-bloody diarrhoea admitted to district hospitals in Kenya. BMC Pediatr *11*, 109.

Opondo, C., Ntoburi, S., Wagai, J. et al. (2009). Are hospitals prepared to support newborn survival? - An evaluation of eight first-referral level hospitals in Kenya. Trop. Med. Int. Health *14*, 1165-1172.

Oshikoya, KA. and Ojo, O.I. (2007). Medication errors in paediatric outpatient prescriptions of a teaching hospital in Nigeria. Nig Q J Hosp Med *17*, 74-8.

Osterholt, DM., Onikpo, F., Lama, M., Deming, MS. and Rowe, AK. (2009). Improving pneumonia case-management in Benin: a randomized trial of a multi-faceted intervention to support health worker adherence to Integrated Management of Childhood Illness guidelines. Hum Resour Health *7*, 77.

Pandey, VA., Kerle, MI., Jenkins, MP. and Wolfe, JH. (2007) AAA benchmarking by Dr Foster: a cause for concern? Ann R Coll Surg Engl *89*, 384-8.

Peabody, JW., Taguiwalo, MM., Robalino, DA. et al. (2006). Improving the Quality of Care in Developing Countries (in Disease Control Priorities in Developing Countries. 2nd edition. Jamison DT, Breman JG, Measham AR, et al., editors) Washington (DC): World Bank.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Phil Trans R Soc A *195*, 1-47.

Perkins, BA., Zucker, JR., Otieno, J. et. al. (1997). Evaluation of an algorithm for integrated management of childhood illness in an area of Kenya with high malaria transmission. Bull World Health Organ *75 Suppl 1*, 33-42.

Petersen, LA., Woodard, LD., Urech, T. et al. (2006). Does pay-for-performance improve the quality of health care?. Ann. Intern. Med. *145*, 265-272.

Pett, MA., Lackey, NR. and Sullivan, JJ. (2003). Making sense of factor analysis. Thousand Oaks California. Sage Publications.

Prieto, L. and Sacristan, J.A. (2003). Problems and solutions in calculating quality-adjusted life years (QALYs). Health Qual Life Outcomes *1*, 80.

Prudon, P. (2013). Confirmatory factor analysis: a brief introduction and critique. [cited 8 January 2014]. Available from http://home.kpn.nl/p.prudon/GOF.pdf.

Rabe-Hesketh, S. and Skrondal, A. (2008). Multilevel and Longitudinal Modelling Using Stata, 2nd Edition. Stata Press.

Raubenheimer, J. (2004) An item selection procedure to maximise scale reliability and validity. SA J Industr Psychol *30*, 59-64.

Reyburn, H., Mwakasungula, E., Chonya, S. et al. (2008). Clinical assessment and treatment in paediatric wards in the north-east of the United Republic of Tanzania. Bull. World Health Organ. *86*, 132-139.

Rosenthal, MB., Frank, RG., Li, Z. et al. (2005). Early experience with pay-for-performance: from concept to practice. JAMA *294*, 1788-1793.

Rowe, AK. (2013). The effect of performance indicator category on estimates of intervention effectiveness. Int J Qual Health Care *25*, 331-339.

Rowe, AK., Onikpo, F., Lama, M. et al. (2003). Risk and protective factors for two types of error in the treatment of children with fever at outpatient health facilities in Benin. Int J Epidemiol *32*, 296-303.

Rutten, GM., Degen, S., Hendriks, EJ. et al. (2010). Adherence to clinical practice guidelines for low back pain in physical therapy: do patients benefit? Phys Ther *90*, 1111-1122.

Rutterford, C., Eldridge, S. and Copas, A. (2011). A review of methodology for sample size calculations in cluster randomised trials. Trials *12*, A23

Saloojee, GM., Rosenbaum, PR., Westaway, MS. et al. (2009). Development of a measure of family-centred care for resource-poor South African settings: the experience of using a modified version of the MPOC-20. Child Care Health Dev *35*, 23-32.

Sarle, WS. (1995). Measurement theory: frequently asked questions. Disseminations of the International Statistical Applications Institute *1,* 61-66.

Sassi, F. (2006). Calculating QALYs, comparing QALY and DALY calculations. Health Policy Plan *21*, 402-8.

Scheiber, JB., Stage, FK., King, J., Nora, A. and Barlow, EA. (2006). Reporting structural equation modelling and confirmatory factor analysis results: a review. J Educ Res *99*, 323-337.

Schoenfelder, T., Klewer, J. and Kugler, J. (2011). Analysis of factors associated with patient satisfaction in ophthalmology: the influence of demographic data, visit characteristics and perceptions of received care. Ophthalmic Physiol Opt *31*, 580-587.

Shi L. (1992). The relationship between primary care and life chances. J Health Care for the Poor Underserved. *3*, 321-35

Shi L. (1994). Primary care, specialty care, and life chances. Int J Health Services *24*, 431-58.

Sibanda, T., Sibanda, N. (2007). The CUSUM chart method as a tool for continuous monitoring of clinical outcomes using routinely collected data. BMC Med Res Methodol *7,* 46.

Siebes, RC., Ketelaar, M., Wijnroks, L. et al. (2006). Family-centred services in The Netherlands: validating a self-report measure for paediatric service providers. Clin Rehabil *20*, 502-512.

Siebes, RC., Nijhuis, BJG., Boonstra, AM. et al. (2008). A family-specific use of the Measure of Processes of Care for Service Providers (MPOC-SP). Clin Rehabil *22*, 242-251.

Simoes, EAF., Peterson, S., Gamatie, Y. et. al. (2003). Management of severely ill children at first-level health facilities in sub-Saharan Africa when referral is difficult. Bull World Health Organ *81*, 522-31.

Sims, AJ, Keltie, K., Burn, J. and Robson, SC. Assessment of competency in clinical measurement: comparison for two forms of sequential test and sensitivity of test error rates to parameter choice. Int J Qual Health Care *25*, 322-30.

Sivaprakasam, J. and Purva, M. (2010). CUSUM analysis to assess competence: what failure rate is acceptable. Clin Teach *7*, 257-61.

Sixma, HJ., van Campen, C., Kerssens, JJ. et al. (2000). Quality of care from the perspective of elderly people: the QUOTE-elderly instrument. Age Ageing *29*, 173-178.

Smith, S., Sinclair, D., Raine, R. and Reeves, B. (2005). Health care evaluation. Maidenhead. Open University Press.

Snijders, T. and Bosker, R. (1999). Multilevel Analysis. SAGE Publications Ltd.

Sodemann, M., Jakobsen, MS., Molbak, K., Alvarenga, IC. and Aaby, P. (1997). High mortality despite good care-seeking behaviour: a community study of childhood deaths in Guinea-Bissau. Bull World Health Organ *75*, 205-12.

Spiegelhalter, D. (2002). Funnel plots for institutional comparison. Qual Saf Health Care *11*, 390-91

Spiegelhalter, D. (2005). Funnel plots for comparing institutional performance. Stat Med *8*, 1185-1202.

Srinivasan, V. and Basu, AK. (1989). The metric quality of ordered categorical data. Market Sci *8*, 205-230.

StataCorp. (2013). Stata: Structural equation modelling reference manual . Statistical Software. College Station, TX: StataCorp LP.

Steckler, A. and McLeroy, KR. (2008). The importance of external validity. Am J Public Health *98*, 9-10

Steiner, SH., Cook, RJ., Farewell, VT. (2001). Risk-adjusted monitoring of binary surgical outcomes. Med Decis Making *21*, 163-9

Stevens, SS. (1946). On the Theory of Scales of Measurement. Science *103*, 677-80.

Streiner, DL. and Norman, GR. (1989). Health measurement scales: a practical guide to their development and use. OUP Oxford.

Suhonen, R., Gustafsson, M., Katajisto, J. et al. (2010). Individualized care scale - nurse version: a Finnish validation study. J Eval Clin Pract *16*, 145-154.

Symmons, DP., Hassell, AB., Gunatillaka, KA. et al. (1995). Development and preliminary assessment of a simple measure of overall status in rheumatoid arthritis (OSRA) for routine clinical use. QJM *88*, 429-437.

Tabachnick, BG. and Fidell, LS. (2007). Using Multivariate Statistics. New York: Allyn and Bacon.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. J Diagn Med Sonog *6*, 35–39.

Teasdale, G. and Jennet, B. (1974). Assessment of coma and impaired consciousness: a practical scale. Lancet *2*, 81-84.

Thompson SG. (1994). Why sources of heterogeneity in meta-analysis should be investigated. Br Med J *309*, 1351-1355

Thompson SG. and Sharp SJ. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med *18*, 2693-2708

Tierney, WM., Achieng, M., Baker, E. et. al. (2010). Experience implementing electronic health records in three East African countries. Stud Health Technol Inform *160*, 371-5.

Tucker, LR. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika *38*, 1-10.

Uebersax, JS. (2006). The tetrachoric and polychoric correlation coefficients. Statistical Methods for Rater Agreement web site. Accessed 15th August 2013 at: http://john-uebersax.com/stat/tetra.htm.

Ullman, J. B. (2001). Structural equation modelling. In B. G. Tabachnick & L. S. Fidell (Eds.), Using multivariate statistics (4th ed.). Needham Heights, MA: Allyn & Bacon.

Van Prooijen, TW. and van der Kloot, WA. (2001). Confirmatory analysis of exploratively obtained factor structures. Educ Psychol Meas *61*, 777-792.

Vassar, MJ., Lewis, FRJ., Chambers, JA. et. al. (1999). Prediction of outcome in intensive care unit trauma patients: a multicenter study of Acute Physiology and Chronic Health Evaluation (APACHE), Trauma and Injury Severity Score (TRISS), and a 24-hour intensive care unit (ICU) point system. J Trauma *47*, 324-9.

Velleman, PF. (1993). Nominal, ordinal, interval and ratio typologies are misleading. Am Stat *47*, 65-72.

Vogel, RL. and Ackermann, RJ. (1998). Is primary care physician supply correlated with health outcomes? Int J Health Services *28*,183-96.

Walter, ND., Lyimo, T., Skarbinski, J. et. al. (2009). Why first-level health workers fail to follow guidelines for managing severe disease in children in the Coast Region, the United Republic of Tanzania. Bull World Health Organ *87*, 99-107.

Wenger, NS., Shekelle, PG. and ACOVE Investigators (2001). ACOVE-1 quality indicators. Ann Intern Med *135*, 653-67.

Weisberg, S. (2005). Simple linear regression. In: Applied Linear Regression. John Wiley & Sons, New Jersey, pp 19-44.

Wheeler, DJ. and Chambers, DS. (1992). Understanding statistical process control. Knoxville, TN: SPC press.

Wierenga, PC., Klopotowska, JE., Smorenburg, SM. et al. (2011). Quality indicators for in-hospital pharmaceutical care of Dutch elderly patients: development and validation of an ACOVE-based quality indicator set. Drugs Aging *28*, 295-304.

Wijdicks, E. F. M. (2006). Clinical scales for comatose patients: the Glasgow Coma Scale in historical context and the new FOUR Score. Rev Neurol Dis, 3, 109-17.

Williams, F. and Boren, SA. (2008). The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. Inform Prim Care *16*, 139-145.

Williams, SM., Parry, BR. and Schlup, MT. (1992). Quality control: an application of the cusum. BMJ *304*, 1359-61.

Wilson, IB. and Cleary, PD. (1995). Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA : the Journal of the American Medical Association *273*, 59-65.

Wobrock, T., Weinmann, S., Falkai, P. et al. (2009). Quality assurance in psychiatry: quality indicators and guideline implementation. Eur Arch Psychiatry Clin Neurosci *259 Suppl 2*, S219-26.

Wong, DT., Barrow, PM., Gomez, M. and McGuire, GP. (1996). A comparison of the Acute Physiology and Chronic Health Evaluation (APACHE) II score and the Trauma-Injury Severity Score (TRISS) for outcome assessment in intensive care unit trauma patients. Crit Care Med *24*, 1642-8.

World Health Organization (1997). Integrated management of childhood illness: conclusions. Division of Child Health and Development. Bull World Health Organ *75 Suppl 1*, 119-28.

World Health Organization (2006). Electronic health records: manual for developing countries. WHO Geneva.

World Health Organization (2008). Global burden of disease: 2004 update. WHO Geneva.

World Health Organization. (2011) World malaria report 2011. Geneva. WHO.

Zali, AR., Seddighi, AS., Seddighi, A. and Ashrafi, F. (2012). Comparison of the acute physiology and chronic health evaluation score (APACHE) II with GCS in predicting hospital mortality of neurosurgical intensive care unit patients. Glob J Health Sci *4*, 179-84.

Zaslavsky, AM. (2001). Statistical issues in reporting quality data: small samples and casemix variation. Int J Qual Health Care *13*, 481-488.

Zimmerman, JE., Wagner, DP., Draper, EA., Wright, L., Alzola, C. and Knaus, WA. (1998). Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit Care Med *26*, 1317-26.

Zuercher, M., Ummenhofer, W., Baltussen, A. and Walder, B. (2009). The use of Glasgow Coma Scale in injury assessment: a critical review. Brain Inj *23*, 371-84.

Zurovac, D. and Rowe, AK. (2006). Quality of treatment for febrile illness among children at outpatient facilities in sub-Saharan Africa. Ann Trop Med Parasitol *100*, 283-296.