The use of statistical models to estimate the timing and causes of neonatal deaths

Shefali Bharat Oza

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of London
October 2019

Department of Infectious Disease Epidemiology

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funding from CHERG/MCEE, UNICEF, and Save the Children USA

Research group affiliations: Maternal and Child Epidemiology Estimation group (MCEE); Maternal, Adolescent, Reproductive, and Child Health (MARCH) Centre at LSHTM

Declaration

I, Shefali Bharat Oza, 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.

Signature:

Date:

Abstract

Despite major reductions in child mortality, decrease in neonatal (first month of life) deaths has been substantially slower. To further reduce neonatal deaths, scale-up of relevant and timely health interventions is necessary. Such scale-up needs to be supported by evidence, but important gaps remain in our knowledge regarding the timing and causes of neonatal deaths.

Birth and the days immediately following carry the highest daily risk of death, yet standard life tables do not present daily mortality risks within the neonatal period. Around three-quarters of neonatal deaths occur during the first week, and most interventions to prevent these deaths must be delivered very quickly. Thus, understanding the neonatal day-of-death distribution is important for delivering appropriate and timely interventions. We fitted an exponential function to survey data to model the daily neonatal mortality risk, focusing on the first day and week after birth. Using this model and observed data, we estimated the daily risk of death in the neonatal period for 186 countries in 2013.

Targeted interventions also require reliable estimates of neonatal cause-of-death distributions. Cause-of-death estimation is challenging because of limited data quantity and quality in many countries. Previous work highlighted the need to expand the existing country-specific neonatal cause-of-death estimates and improve the methods. We developed a multinomial model to estimate the neonatal cause-of-death distribution by the early (days 0-6) and late (days 7-27) neonatal periods. We then focused on methodological improvements, including evaluating performance and developing a proof-of-concept Bayesian mixed effects model.

This thesis straddles two topics that are receiving increased attention: cause-of-death estimation and neonatal health. Ideally, the results from this work can help current neonatal health policies and programmes while contributing to the growing area of cause-of-death modelling. However, the longer-term aim should be to improve data collection to obviate the need for statistical modelling exercises.

Acknowledgements

I have had the great fortune of having wonderful family, friends, colleagues, and mentors in my life. They have inspired me and taught me, shown me kindness and brought me joy, and lifted my spirits in times of need. I am grateful to all of them for shaping me and my path, and for keeping me company along the way.

To my mentors and colleagues: I have found such pleasure in working with and learning from all of you. I am hugely grateful to Simon Cousens (my supervisor) for his years of mentorship and insights, and for leading by example on how to be a thoughtful, patient, and rigorous researcher. Thank you to Joy Lawn and Hannah Blencowe (my advisory committee members) for invaluable advice and inspiration, especially on how research can shape policy. I am deeply appreciative of the many discussions which helped guide the work in this thesis, especially with Dan Hogan, Colin Mathers, David Prieto-Merino, Amy Mulick, Jamie Perin, Pancho Villavicencio, Bob Black, and the wider CHERG/MCEE team. Thank you also to Majid Ezzati, my former mentor and current friend, who helped me find my place in epidemiology and research.

To my friends: I cherish our bonds – you have helped fill my days with meaning, joy, and laughter. I want to say a special thank you to Neha, Giorgia, Hoviyeh, Hannah, and Piali. So many others have a place in my heart, but to thank each of you properly would fill up the allocated word count of this thesis. So, let us meet again soon and I will thank you in person.

To my family: who needs an exponential when one can have constants like you? Mummy/Daddy, I am forever grateful for your unconditional love and support. Pappa, I hope you knew how much you have inspired me since I was a child. Shalin/Colleen and the Harrow clan, how lucky I feel to have you as my family. As for my two no-longer-neonates-but-still-little-ones: during these uncertain times, I do not know what kind of world you will inherit. But I will try my best to give you a sense of wonder, an appreciation for kindness and humility, and the spirit to fight for equality. Aram, I could not ask for a more perfect companion to walk alongside me in life.

I am looking forward to the next chapter (thesis-related pun) with all of you.

A brief postscript: These feel like tough times, with dispiriting news of authoritarianism, hate, climate change, and xenophobia/nationalism on the rise. I have felt confusion and sadness from this, but also hope and energy from the many who are fighting back. Let us urgently fight louder and harder for a fair, equal, and kind world. I take inspiration from Tennyson's Ulysses: "…to strive, to seek, to find, and not to yield." May we all strive, seek, find, and not yield together.

Table of Contents

List of panels

List of figures

List of tables

Abbreviations

ANC	Antenatal care
BCG	Bacillus Calmette-Guerin vaccine (i.e. tuberculosis vaccine)
BMGF	Bill and Melinda Gates Foundation
CHERG	Child Health Epidemiology Reference Group
CI	Confidence interval
COD	Cause of death
CRVS	Civil registration and vital statistics
DHS	Demographic and Health Surveys
DPT	Diphtheria/Pertussis/Tetanus vaccine
FE	Fixed effects
FLR	Female literacy rate
GBD	Global Burden of Disease
GFR	General fertility rate
GNI	Gross national income
GOF	Goodness of fit
ICD	International Statistical Classification of Diseases
IHME	Institute for Health Metrics and Evaluation
IMR	Infant mortality rate
IQR	Interquartile range
LAC	Latin America and the Caribbean
LBW	Low birth weight
LSHTM	London School of Hygiene and Tropical Medicine
MCEE	Maternal and Child Epidemiology Estimation group
MCMC	Markov chain Monte Carlo
MDGs	Millennium Development Goals
ME	Mixed effects
NMR	Neonatal mortality rate
PAB	Protected at birth (against neonatal tetanus)
RE	Random effects
SA	South Asia
SBA	Skilled birth attendance
SD	Standard deviation
SDGs	Sustainable Development Goals
SRS	Sample registration system
SSA	Sub-Saharan Africa
U5MR	Under-5 mortality rate
UN	United Nations
UNICEF	United Nations Children's Fund
UN-IGME	United Nations Inter-agency Group for Child Mortality Estimation
UR	Uncertainty range
VA	Verbal autopsy
VR	Vital registration
WHO	World Health Organization

Definitions

Neonatal period: first four weeks (days 0-27) of life

Early neonatal period: first week of life (days 0-6)

Late neonatal period: weeks 2-4 of life (days 7-27)

Neonatal mortality rate (NMR): probability of dying between birth and 28 days, expressed per 1,000 live births

Infant mortality rate (IMR): probability of dying between birth and 1 year, expressed per 1,000 live births

Under-5 mortality rate (U5MR): probability of dying between birth and 5 years, expressed per 1,000 live births

Preterm: a baby that is born at less than 37 weeks gestational age

Note: Throughout this thesis, I use the terms "observed data" or "empirical data" to refer to ground-level data collection (i.e. primary data collection). I use the terms "estimates" and "predictions" to refer to the outputs of statistical models.

1 Background

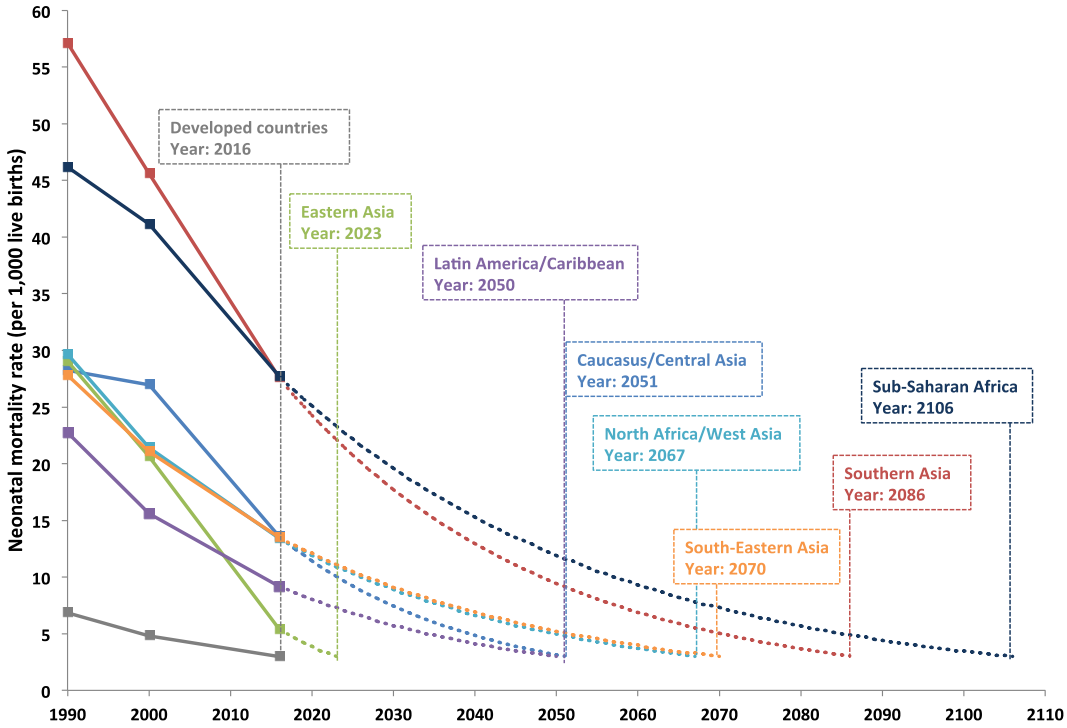
1.1 Motivation

This is a special time for global health. We stand at a crossroads as the Millennium Development Goals (MDGs) ended in 2015 and progress towards the Sustainable Development Goals (SDGs) is now underway. In the last few decades, we have seen remarkable progress in health worldwide, including substantial drops in child mortality across most countries and reductions in the incidence of diseases like HIV/AIDS, tuberculosis, and malaria [1]. But a staggering number of preventable deaths still occur each year, including most of the 2.5 million estimated deaths among newborns under one month old [2, 3].

This is also a special time for data. The quantity of data is exploding, with more than 2.5×10^{18} bytes created each day in 2018 and 90% of the data worldwide created in the last two years [4]. Yet in the era of "big data", we still have shockingly little information on some of the most basic but important topics regarding many of the world's citizens. For example, more than half of countries have no or poor quality systems for recording births and deaths, resulting in only about 40% of annual deaths being recorded globally [5]. The majority of countries lacking such systems are in Sub-Saharan Africa and Asia. Just as with inequality in general, data inequality means that the places with the greatest need but fewest resources have the least data available to help inform priorities, practice, and progress.

This thesis views these two topics – global health and data – through the lens of neonatal mortality (deaths in the first four weeks of life). This is an area of global health that has traditionally been high in burden but low in data. Although there have been major recent reductions in overall under-5 mortality, the decrease in deaths during the first month of life (neonatal period) has been substantially slower [6]. The estimated 2.5 million annual neonatal deaths, many from preventable causes, now account for about 47% of all under-5 child deaths [3]. One hundred and thirty-three countries were unable to achieve the fourth MDG target of a two-thirds reduction in under-5 mortality between 1990 and 2015 [7], at least partly due to limited reductions in neonatal deaths [8]. At the current rates of decline, it will take nearly 70 years for Southern Asia and 90 years for Sub-Saharan Africa to achieve neonatal mortality rates presently observed in most European and North American countries (Figure 1.1).

Figure 1.1: Projected time for regions to reach 3 neonatal deaths per 1,000 live births based on their average annual rate of reduction for neonatal mortality from 2000-2016



Note: 3 deaths per 1,000 live births was the average neonatal mortality rate in the MDG "developed countries" region in 2016; this an updated version of a figure created by the author for [6].

To further reduce neonatal deaths, scale-up of relevant and timely health interventions is necessary. Such scale-up needs to be supported by evidence, but important gaps remain in our knowledge regarding the timing and causes of neonatal deaths. This is particularly true for the highest burden countries, for which we have the fewest data. For such countries, statistical modelling exercises are necessary to provide estimates to inform decisions. To be credible and thus useful, such estimates need to be reliable. In this thesis, I aim to address some of these information and reliability gaps. Below is an overview of the topics that I will cover.

Birth and the days immediately following carry the highest daily risk of death, yet standard life tables do not present daily mortality risks within the neonatal period. Deaths on the day of birth (day 0) are particularly important because they account for a large number of deaths that can be targeted by interventions at the time of birth [9]. A dramatic fall in risk occurs even within hours of birth. For example, the risk of death (per 1,000 live births) in the first hour after birth in the US is 0.91 [10]. In contrast, the risk in the following 23 hours is 1.58, which translates to an average hourly risk of about 0.07, indicating a dramatic decline in risk. Additionally, the causes of day 0 deaths more closely resemble those of intrapartum stillbirths than those of deaths later in the neonatal period [11]. Around three-quarters of neonatal deaths occur during

the early period (days 0-6) [12], and most interventions to prevent these deaths need to be delivered within a very short window of time. Thus, understanding the day-of-death distribution within the neonatal period is important for delivering appropriate and timely interventions. Yet, there have been no systematic, nationally comparable estimates of risk during the first day or week of life, despite increasing programmatic focus on these important time periods.

Targeted interventions to prevent neonatal deaths also require reliable estimates of the neonatal cause-of-death distribution. Ideally, country-specific neonatal cause-of-death estimates would 1) provide accurate information about both levels and trends, 2) include all major programmatically relevant causes of death, 3) distinguish between deaths in the early (days 0-6) and late (days 7-27) neonatal periods, and 4) identify relevant subnational variation. Cause-of-death estimation is challenging because of the limited quantity and quality of data available for many countries, especially regarding trends. Since 2005, there have been regular estimates of neonatal causes of death by country [13, 14], as well as some recent time trend analyses [15]. While an important start, these estimates have highlighted the need to expand the estimates where possible (e.g. early/late neonatal period split), improve the statistical models, and test the reliability of the models against empirical data.

For my PhD, I present a body of work whose aim is to improve our understanding of the temporal and causal distribution of deaths within the neonatal period and to address some of the important remaining challenges in neonatal cause-of-death estimation. Much of the work presented here sits at the junction of epidemiology, statistics, and public health, with some influences from medicine and computer science.

1.2 Thesis objectives

My overall research goal is to improve our understanding of the temporal and causal distributions of deaths within the neonatal period. This includes both conducting analyses to fill knowledge gaps and making improvements to existing statistical models.

The key objectives are divided into three main themes, as follows:

<u>Theme 1: Filling knowledge gaps</u>

- Objective 1: Estimate risk of death by day within the neonatal period by country (chapter 3)

- Objective 2: Estimate cause-of-death distributions for the early and late neonatal periods (chapter 4)

Theme 2: Improving modelling techniques

- Objective 3: Identify limitations of the existing neonatal cause-of-death models (chapter 5)
- Objective 4: Investigate improvements to the neonatal cause-of-death models (chapters 6 and 7)

Theme 3: Discussion and recommendations

- Objective 5: Synthesize main findings and relevant lessons learned for researchers, policy makers, and programme implementers (chapter 8)

1.3 Structure of thesis

This thesis is divided into nine main chapters.

In **chapter 2**, I give an overview of available data sources for causes and levels of neonatal mortality, as well as for covariates used in our neonatal cause-of-death regression models. This includes a discussion of the strengths and limitations of each data source type.

I present our work on estimating the distribution of deaths by day within the neonatal period in **chapter 3**. For this, we developed and fitted a mathematical model to empirical data on daily deaths during the neonatal period. We then used this model to estimate the proportion of daily neonatal deaths for countries without adequate vital registration data, with a focus on deaths within the first day and week after birth.

Our work on estimating causes of death separately for the early and late neonatal periods is included in **chapter 4**. I describe how we modified the existing neonatal cause-of-death models to include this finer age gradation and discuss the implications of this.

In **chapter 5**, I describe the limitations we identified with the existing neonatal multinomial cause-of-death models. I investigate model performance issues, particularly focused on predictive accuracy and model stability, and consider whether and how to weight country-specific empirical data for a country's modelled estimates.

I then describe our work to improve the neonatal cause-of-death models in chapters 6 and 7. Our results of implementing a set of binomial models and comparing these to the multinomial model results are included in **chapter 6**. In **chapter 7**, I present our work on shifting our models to the Bayesian framework, including with covariate selection and country-level random effects, and how these results compare to the multinomial model results.

In **chapter 8**, I summarize the main findings of this thesis, compare our results with work done by others, and discuss overall strengths and limitations. Drawing from the body of work presented in this thesis, I also provide recommendations on some aspects of statistical modelling, including a discussion about core modelling principles and the uses and limitations of modelled estimates.

1.4 Contributions by the candidate

The work presented in this thesis is that of the author, with the following exceptions:

- Chapter 3 – the majority of text in this chapter has been published, and thus has been through revisions based on suggestions from co-authors and peer reviewers.
- Chapter 4 – the majority of text in this chapter has been published, and thus has been through revisions based on suggestions from co-authors and peer reviewers.

I worked closely with Professor Simon Cousens on various technical aspects described in this work. I also benefited from discussions with several colleagues (see Acknowledgements) regarding many of the topics presented in this thesis.

1.5 Ethical approval

All of the work presented in this thesis used only data that is fully available in the public domain, downloadable from published literature, the World Health Organization, and the Demographic and Health Surveys website.

1.6 Funding

I am grateful to have received funding for several aspects of the work presented in this thesis. Funding was provided from the following sources:

- Timing of neonatal deaths (chapter 3): Save the Children USA

- Neonatal cause-of-death estimates and model improvement (chapters 4-7): United Nations Children's Fund (UNICEF) and Child Health Epidemiology Reference Group (CHERG) (through the Bill and Melinda Gates Foundation [BMGF]); the Maternal and Child Epidemiology Estimation group (MCEE) (through BMGF).

2 Data sources for neonatal estimates on timing and causes of death

Here I describe the mortality, cause of death, and covariate data sources used for the work in this thesis, including some strengths and limitations of these sources. While many of these sources are generalizable to all ages, the focus of this chapter is on sources relevant for the neonatal age group. My aim is to provide some context to help the reader better understand the nuances of the input data used in this work. This, paired with the descriptions of methods in later chapters, should help the reader better gauge our modelled estimates, including their strengths and limitations.

2.1 Introduction

For the statistical models we use to produce various neonatal estimates, we need input data on all-cause mortality levels, cause-of-death (COD) distributions, and explanatory variables (covariates). These data come from a variety of sources and range widely in type, data collection methodology, and quality. We obtain and use some of these data in their unaltered raw form (e.g. COD distributions from studies) while others are provided to us already processed to deal with issues like missingness or the need for smoothing (e.g. nationally comparable covariate time-series data received from the World Health Organization [WHO]).

The various data collection and estimation methods (e.g. registries, household surveys, modelling) have different benefits and limitations, often balancing feasibility against quality and/or completeness. For example, improving quality may come at a financial or human resource cost that is untenable, or a study may only be possible for a small non-representative population sample. Understanding the nuances of these different methodologies is useful when interpreting the data coming out of them, as well as data that are indirectly related (e.g. modelled estimates – like ours – that rely on such input data).

In this chapter, I provide a brief background on the mortality rate, cause of death, and covariate data sources that are relevant for the neonatal estimates we developed. The complexities of each data source mentioned are greater than can fully be addressed in the short summary presented here; there are entire manuals and books dedicated to most of these data source types. Thus, this chapter should be viewed as a non-exhaustive overview that highlights the key points relevant to the specific work described in this thesis.

2.2 Mortality data sources

Registering births and deaths at the national level has wide-ranging implications for health, from population estimation and resource allocation to identifying demographic and mortality patterns. Mortality information is especially useful for understanding the overall health of a population and how well the health system is functioning. In this section, I discuss mortality data from civil registration and vital statistics (CRVS) systems, as well as alternative data sources for areas lacking adequate CRVS systems. I also discuss how we use neonatal mortality data in our work.

2.2.1 Civil registration and vital statistics systems

Recognition that official birth and death statistics are useful dates back to at least 1685, when churches in Norway reported aggregated birth, death, and marriage numbers from local parishes [16]. Over the ensuing 300+ years, such recording of major life events has evolved into formalized national CRVS systems. Vital registration (VR) data are the accumulated data on vital events collected through birth and death certifications in these systems. Defining characteristics of well-functioning CRVS systems are that the individual-level VR data within them are collected in real-time, continuously, and at the local level [17].

Well-functioning CRVS systems have several benefits, including health benefits for the population [18]. For individuals, these records are useful for a number of reasons, including providing documentation (e.g. birth certificates), conveying legal status, and allowing access to various services and social protections [19]. When aggregated (typically by national-level health or statistics offices), these data provide an important lens onto the population at local, regional, and national levels [17]. For example, age- and sex-specific death rates, regional mortality differences, and excess mortality can all be ascertained from these data. VR data can thus provide much of the information needed for determining policies and priorities relevant to mortality reduction. Importantly, the continuous nature of VR data allows the tracking of trends over time [20].

However, this seemingly basic task of registering births and deaths is in fact a complex process that requires reliable long-term financial, regulatory, and political/legislative commitments alongside adequate infrastructure, organisation, and training [20, 21]. A 2015 study estimated that only 65% of births and 38% of deaths worldwide are registered [5]. The majority of low-resource, high-mortality settings lack CRVS systems. Of countries with national records, many

suffer from incomplete or low-quality recording of deaths (e.g. misreporting of age or cause of death) [22], leaving important gaps in understanding even basic health outcomes. Additionally, the availability and quality of VR data in a country may change over time, especially in countries with newly emerging registration systems [23].

Several factors contribute to the quality of VR data. The WHO assesses the quality of death registration data by coverage, completeness, COD data quality, and (to a lesser extent) timeliness of data reporting [24]. Coverage refers to the percentage of a country's population covered by the CRVS system. Completeness refers to the percentage of recorded deaths (out of all deaths) in areas covered by the CRVS system in that country. The quality of COD VR data can be assessed by the proportion of deaths attributed to ill-defined and/or non-specific codes [24]. Recently, an alternative "vital statistics performance index" was developed [25]. This composite score assesses VR data using six factors consisting of completeness, COD data quality, level of cause-specific detail, quality of age and sex reporting, data availability and timeliness, and internal consistency.

Based on the WHO quality measures, only 25% (49/194) of WHO member states had high-quality death registration data whereas 42% (81/194) had no or very low-quality data between 2005-2015 [26]. The latter category was dominated by some of the poorest and highest mortality countries; 94% (43/47) of African countries and 73% (8/11) of Asian countries (based on WHO regions) had no or very low-quality VR data. However, there is growing momentum to implement new systems and improve existing ones in countries lacking adequate CRVS systems [27, 28]. Innovative solutions are also being deployed to overcome the challenges of implementing such systems in resource-limited settings. Some of these include trying to link health service records (e.g. hospital data) with CRVS systems [29], incorporating verbal autopsies (described in section 2.3.2) into CRVS systems [30, 31], and using community health workers to increase vital event registration [32].

For neonates, high-quality VR data are a key source of information on the number, timing, and causes of death [33, 34]. Unfortunately, neonatal deaths are among the most underreported (i.e. low completeness) in lower-quality CRVS systems [17, 35, 36], highlighting the fact that there are substantial barriers towards registering births and deaths within this first month. Other data quality issues for neonatal VR data include misrecording of deaths by day (i.e. age of death) and by cause (further discussed in section 2.3.1), as well as misreporting very early neonatal deaths as stillbirths.

2.2.2 Alternatives to CRVS systems

Given that the majority of countries lack adequate VR data, alternative data sources are needed to estimate mortality in these countries. Here, I describe some of the key alternatives.

Continuous data collection systems

Various methods that allow for continuous recording of vital events have been implemented in countries lacking adequate CRVS systems. Most of these options lack the coverage of a well-functioning CRVS system but can still provide much-needed mortality information for the covered populations.

Sample registration systems (SRSs) are similar to CRVS systems but only cover a nationally (or subnationally) representative sample of the population [37]. Alongside the continuous collection of demographic information in these systems, SRSs often also include regular household surveys about these demographic events [38]. This dual recording helps catch errors and acts as a form of quality control and evaluation between the two data collection methods. Advantages of an SRS include lower costs compared to a full CRVS system, (sub)nationally representative samples if done well, and acting as a step towards a more comprehensive CRVS system in the future. However, SRSs require substantial infrastructure and resources to maintain representative samples, and can suffer from similar completeness, coverage, and quality issues as CRVS systems (section 2.2.1). Countries currently with SRSs include India, Bangladesh, Indonesia, and China [37, 39, 40].

Health and demographic surveillance systems (HDSSs) are well-defined geographic areas which have continuous monitoring of health and demographic indicators/outcomes [20]. Numerous HDSSs have been established in low- and middle-income countries (largely in Africa, Asia, and Oceania) since the 1940s [20]. Some HDSSs are part of networks (e.g. INDEPTH) which result in better standardization and comparability across sites [41]. Advantages of HDSSs include that they are generally well-monitored sites with good completeness and coverage of key indicators (e.g. mortality) at the community-level and over time [42]. However, they typically cover small areas and are not (sub)nationally representative, thereby limiting the generalisability of their results.

Various programme- or disease-specific registers also collect continuous mortality data. For example, the Maternal Newborn Health Registry is a prospective, population-based registry of pregnancy outcomes across Global Network sites [43]. This allows the continuous collection of

demographic and health information on relevant topics, including on neonatal mortality. Such registers are usually limited in scope geographically (similar to HDSSs) and topic area. Thus, these registers have similar drawbacks to localized HDSS sites.

Health management and information systems (HMISs) are medical record systems which typically collect data on patients accessing health services. The data from these systems can then be aggregated at the subnational or national level for calculation of relevant health statistics. The patient-level information in HMISs can be relatively extensive, including, potentially, medically certified death records. However, many births and deaths in low-resource high-mortality settings do not occur within the healthcare system (e.g. at a hospital or clinic), and thus are not likely to be included in the HMIS. This limits the generalisability of HMIS records in many countries. Even for deaths with associated medical records, these records may not be aggregated, cross-checked, or analysed in settings with weak or fragmented HMISs [44]. Initiatives to improve the quality and completeness of HMISs [45], their reach beyond health facilities [46], and their integration with wider data collection systems (e.g. nascent CRVS systems) [47] have potential to yield better quality and higher coverage data in such settings.

One-off or periodic data collection methods

Retrospective household surveys such as the Demographic and Health Surveys (DHS) and UNICEF's Multiple Indicator Cluster Surveys (MICS) have been a vital source of mortality information for many countries with inadequate CRVS systems [48]. These surveys ask female respondents for birth histories including information about deceased children. These data can be used to calculate child mortality rates. Key advantages of such surveys are that they are standardized, feasible to conduct in many low-resource settings (e.g. DHS have been conducted in over 90 countries [49]), and can provide all-cause mortality statistics that are (at a minimum) nationally representative. Unfortunately, such surveys have well-documented limitations, including having various biases (e.g. recall), being retrospective instead of real-time, and not being useable for local information below the representative level (e.g. national or subnational) [50]. Additionally, nearly all of these surveys are one-offs or conducted periodically with several year gaps, making it infeasible to track time trends in real time [51].

Population censuses, which are typically done every five to ten years, are a source of periodic demographic data. These are similar to household surveys in requesting retrospective information from respondents, but collect much less detailed information. For example, a census may collect a summary birth history which requires various assumptions to be made in

order to estimate mortality rates. Censuses also have similar limitations to household surveys as well (e.g. recall biases, not continuous data collection) [22].

2.2.3 Modelling

The mortality data sources described in the previous section are collected at the local, subnational, or national level. Two main groups currently produce annual nationally comparable all-cause neonatal mortality rate estimates with time trends: the United Nations Inter-agency Group for Child Mortality Estimation (UN-IGME) [3, 52] and the Institute for Health Metrics and Evaluation (IHME) [53]. UN-IGME, as part of the UN, is led by UNICEF and includes membership from the World Bank, UN Population Division, and the WHO. These groups use various empirical data (including those described in sections 2.2.1 and 2.2.2) as inputs into statistical models to estimate mortality levels across countries and over time.

A key benefit of these modelled estimates is that they allow the comparison of mortality levels across countries, and also provide a foundation for producing nationally comparable estimated numbers of deaths by various sub-categories (e.g. cause, day of death). However, these estimates rely on statistical models and imperfect input data (as described in the previous sections) and thus may suffer from issues of predictive accuracy. Indeed, substantial differences have been documented between the results of the different modelling strategies [54] and between modelled estimates and empirical data (e.g. DHS) from countries [55].

2.2.4 How we use mortality data in our work

We use neonatal mortality data and estimates in several ways for the work presented in this thesis. Here, I have included a brief description of these uses; full details are provided in the relevant chapters later in the thesis.

Risk of death by day analysis (chapter 3)

- We used neonatal mortality data from DHS and high-quality VR as inputs into our models
- We applied day-of-death proportions to the nationally comparable UN-IGME total neonatal death numbers (or envelopes) to calculate day-specific numbers

Cause-of-death analysis (chapters 4-7)

- Neonatal mortality rate (NMR) levels were used to separate modelled countries into high (NMR~≥15) and low mortality (NMR~<15) countries

- We used neonatal, infant, and child mortality rates from UN-IGME as input and prediction covariates in our models (see section 2.4)
- We applied COD proportions to the nationally comparable UN-IGME total neonatal death numbers to calculate cause-specific numbers

2.3 Cause-of-death data sources

Beyond counting deaths, understanding the distribution of causes of deaths is important for identifying appropriate interventions and programme priorities. However, ascertaining causes of death is not a trivial task even in low-mortality high-resource settings. Here, I discuss various COD data sources and how we use neonatal COD data in our work.

2.3.1 CRVS systems

CRVS systems are one of the main sources of individual-level COD data. The importance of COD data in these systems is highlighted by the fact that the quality of these data is a key metric for gauging the overall quality of CRVS systems (see section 2.2.1 for details on CRVS systems). When aggregated and analysed, VR-based COD data provide valuable information on COD levels and trends to help inform policies and programmes.

In well-functioning CRVS systems, COD data include official death certification by a medical examiner with information on the immediate and underlying causes of death and the medical chain of events leading to the death [56]. Assigning standardized COD codes using the International Statistical Classification of Diseases and Related Health Problems (ICD) has made comparability of medically-certified deaths possible over time and across countries [23]. This intricate coding system (currently in its 10th revision; ICD-10), with multiple levels and thousands of codes, reflects the complexities of classifying causes of death. Further details on death certification are included in sections 8.3.2 and 8.4.

Cause-of-death data from CRVS systems can have several limitations. First, the patterns of COD data availability in CRVS systems mimic those of all-cause mortality data (section 2.2.1), but with even lower levels of completeness and quality. For example, a recent report estimated that only 27% of global deaths were registered with a cause of death, and an even smaller 13% were reported to the WHO with a meaningful ICD code [26]. There are also major differences in VR COD data by region. The completeness of COD data was estimated to range from 6% in the

WHO Africa region and 10% in South-East Asia to percentages in the mid-90s or higher for the Americas and Europe [57].

Additionally, differences in coding practices can affect accuracy and comparability of VR data over time and across countries. Correctly coding deaths using ICD classification rules requires extensive training. Problems with accurate COD coding are well-documented even in countries with high-quality VR data [58, 59], including substantial use of codes for unknown or ill-defined causes (sometimes referred to as "garbage codes"). Previous studies have found that coding practices for ascertaining causes of death can vary between and even within countries [60, 61], sometimes due to transitions between different ICD revisions [62]. Finally, ICD-10 codes are not always ideal for neonatal causes, particularly because several programmatically relevant causes are relegated to the often-unused fourth digit/level in the codes.

Well-functioning CRVS systems remain the best option for high-quality VR data that are useful at the local, subnational, and national levels. However, the limited availability and quality of these data in many high-mortality countries mean that alternative data sources and methods are also needed (sections 2.3.2-2.3.3).

2.3.2 Verbal autopsies

For settings where medical certification of cause of death is irregular or unavailable, alternative methods are required to provide insight into the probable cause(s) that led to an individual's death. The main alternative method for this is verbal autopsy (VA). Since its first uses in Asia and Africa in the 1950s and 1960s [63], VA has become increasingly common in areas where CRVS is not well established and/or deaths commonly occur outside of the health system. Typically, VA involves trained interviewers obtaining information on the signs, symptoms, and events surrounding a death from a family member of the deceased person [64]. The probable cause of death is then determined based on this information either through physician review or analytically using computer-based algorithms [65].

Although VAs provide a much-needed source of COD information, they remain an imperfect alternative to high-quality VR data [63]. VA data are known to be of variable quality, and the reported cause distributions depend heavily on factors like the causal hierarchies and case definitions used to attribute deaths to particular causes [66-69]. The lack of both standardized VA methods and full reporting of methods has made comparisons between studies even more challenging [68, 69]. This is true even for some of the (otherwise standardized) DHS which now

include neonatal COD assessments [70]. Accurate cause attribution using VA is especially problematic for causes that are difficult to distinguish between, such as neonatal sepsis and pneumonia, or difficult to identify, such as congenital disorders without external signs. Unsurprisingly, VA methods work better for causes with clear symptoms [71] and less well for illnesses like malaria which have non-specific signs and symptoms [72]. VA for neonatal deaths has the added complication that the sick baby is unable to communicate symptoms to a caretaker.

To address some of these VA concerns, several improvements have been proposed and tested in recent years, including standardization of VA tools [67, 73, 74] and probabilistic algorithms for assigning deaths [73, 75]. Several networks (e.g. INDEPTH [41]) and multi-country studies (e.g. AMANHI [76]) have also implemented VAs at their sites, which helps to enforce standardised methods and increases within-network comparability. Additionally, using computers to process large amounts of raw VA data has potential to greatly increase both efficiency and standardisation of results [77]. Portable electronic devices have also been found to improve effectiveness and timeliness of results when conducting VAs [77, 78].

Recently, more countries and donors are investing in large nationally representative VA studies in countries without adequate CRVS system. There are also increasing discussions on how best to integrate VA into CRVS systems [30, 79]. While not easy, such integration has potential to strengthen weak CRVS systems which currently lack COD information. Despite their limitations, verbal autopsies remain the only viable large-scale option for ascertaining causes of deaths in places without adequate medical certification of deaths. Further discussions of VA methods and limitations are included in sections 4.5 and 5.1.3, and additional details on methods to improve VA quality are included in section 8.5.

2.3.3 Alternatives to CRVS systems and verbal autopsies

In the event of death, patient medical records (preferably as part of an HMIS) can be a rich source of data even in countries without adequate CRVS systems. These records ideally include medical certification of death by a clinician with cause of death listed using standard ICD classifications. However, no or poor-quality medical certification of deaths even in hospital settings is common in many countries [17, 80]. Additionally, these data have limited generalisability in the many settings where the majority of deaths occur outside of the health system [42].

Mortality audits aim to assess factors contributing to individual deaths, with goals of improving quality of care and identifying ways to avoid similar deaths in the future where possible [81]. Of particular relevance to neonates are perinatal mortality audits, which are focused on the time period immediately before and after birth. However, perinatal audits are still far less common than maternal mortality audits in low-resource settings [81]. Mortality audits range in their scope, but may include investigations of factors related to clinical causes of death (through medical certification or verbal autopsy), access to and quality of care, and other societal and health system issues that may have contributed to a death [82]. These audits can provide deep insights into the multifaceted factors surrounding deaths while also helping to reduce undercounting and underreporting [81]. However, mortality audits have typically been conducted at the facility level, which has limited their scope and generalisability [83]. Recent increases in community-based mortality audits, integration of mortality audit systems into national health data, and expansion of mortality audits within national programmes have potential to improve the reach of mortality audit systems [81, 83].

Finally, minimally invasive autopsy (MIA; sometimes called minimally invasive tissue sampling [MITS]) is a promising recent method aimed at improving accuracy of COD ascertainment in low- and middle-income countries [84]. MIA/MITS typically involves using needle sampling to obtain post-mortem tissue samples from various organs, followed by histopathologic examinations and identification of infectious organisms [85]. These methods have equipment, training, and financial costs which are higher than clinical diagnoses or verbal autopsies, but lower than conventional autopsies. Recent studies have investigated the feasibility, acceptability, and diagnostic accuracy of MIA/MITS with generally positive results except where cultural/religious beliefs conflicted with post-mortem examinations [84-86]. While unlikely to replace VA studies at least in the short term, MIA/MITS could be used for evaluation of VAs to help improve the accuracy of VA data and better identify patterns of bias in VA COD ascertainment [87].

2.3.4 Modelling

Modelled estimates are another source of cause-of-death numbers. We do not use modelled COD estimates as an input source in our work, but they are a product of our work. I have therefore included a brief history and description of neonatal COD modelling here.

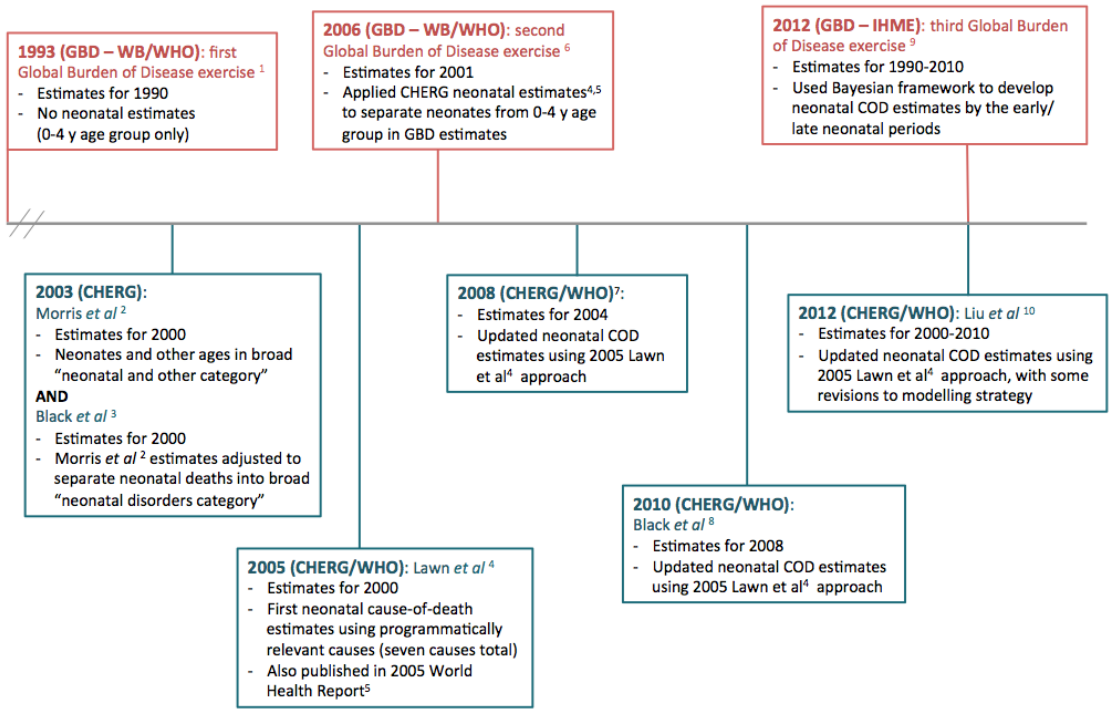
<u>History of COD modelling</u>

For countries which lack recent high-quality VR data, modelling exercises have remained necessary to provide relevant COD estimates. The first systematic exercise to develop nationally

comparable estimates of disease burden was released in 1993 for the year 1990 by the World Bank and WHO as part of the Global Burden of Disease (GBD) study, with several related publications through 1997 [88-92]. Prior to this work, different groups often produced estimates for single causes, which when summed together resulted in all-cause mortality estimates that far exceeded the actual global mortality totals [93]. A key goal of the GBD study was to estimate the COD distribution by country, including for those without high-quality data, while enforcing the condition that the cause-specific deaths in a country summed to the total national number of deaths. Since this seminal work, COD estimation has become an important area of research. While this first GBD exercise did not report neonates as a distinct age group, it established a foundation for developing such estimates.

Since the early 2000s, the Child Health Epidemiology Reference Group (CHERG) and the GBD project have released COD estimates pertaining to the neonatal period. A brief history of their efforts (up until the start of the work presented in this thesis) is summarized in Figure 2.1. Note that CHERG is now the Maternal Child Epidemiology Estimation (MCEE) group (as of 2013) and the GBD project has been based at the Institute for Health Metrics and Evaluation (IHME) since 2010. CHERG/MCEE and IHME work independently from each other. CHERG/MCEE collaborates closely with the WHO and other parts of the UN system, including UNICEF.

Figure 2.1: Major global cause-of-death estimation exercises pertaining to neonatal mortality from 1993-2013



Notes: WB = World Bank; references: [1][88-92]; [2][94]; [3][95]; [4][13]; [5][33]; [6][96-98]; [7][99]; [8][14]; [9][93]; [10][15]

In 2003, Morris et al. used a regression model to estimate the child COD distribution – with neonatal deaths grouped with other ages into a broad category of "neonatal and other" – for countries with poor VR data [94]. Subsequently, others estimated under-5 deaths while including neonates in a separate "neonatal disorders" category [95]. The first nationally comparable COD estimates of neonatal mortality using programmatically relevant cause categories were published by Lawn et al. in 2005 for the year 2000 [13]. These estimates, produced on behalf of CHERG, were developed using multi-cause multinomial logistic regression models based on high-quality VR data for low mortality countries and research studies from high mortality/low resource settings for high mortality countries. Although the second GBD project did not estimate COD for neonates as a distinct age group, they applied the 2005 CHERG neonatal results [13, 33] alongside their own adjustments to distinguish neonates within their 0-4 year age group estimates [98].

Updated neonatal COD estimates were subsequently published on behalf of CHERG for 2008 [14] and 2010 [15], with the latter including annual trends since 2000. None of these estimates distinguished between the early and late neonatal periods. Additionally, the analyses relied on regression strategies that are unlikely to account fully for the complexities in the data and modelling. I address several of these topics in this thesis (chapters 4-7).

IHME released neonatal COD estimates by country and neonatal period for 1990 and 2010 through the third major GBD exercise in 2012 [93]. IHME used a Bayesian framework to develop single-cause ensemble models (for most causes) to estimate death rates and then applied an algorithm to proportionally rescale cause-specific estimates to sum to overall age-sex envelopes [93]. They estimated distributions for 235 causes for 20 age groups and thus multi-cause models were not computationally feasible.

Since 2013, updated neonatal COD estimates have continued to be produced by IHME and CHERG/MCEE [100-103]. In this thesis, I describe some of our work (on behalf of CHERG/MCEE) of splitting neonatal causes of death by the early/late periods (chapter 4) and investigating improvements to the existing COD models (chapters 5-7). I also provide a brief comparison of the CHERG/MCEE versus IHME approaches in section 8.2.

2.3.5 How we use cause-of-death data in our work

As noted in the previous section, nationally comparable neonatal COD estimates are a product of our work. We use empirical neonatal COD data as model inputs to produce these estimates. Specifically, we use high-quality VR data (section 2.3.1) and data extracted from published VA

studies (section 2.3.2) as inputs to model the COD distributions for countries with inadequate CRVS systems. For countries with high-quality VR data, we report their VR COD data as-is. Further details of our neonatal COD modelling strategy, including on the COD input data, are provided in section 4.2. Selection criteria for inclusion of VA studies are included in section 4.2.2, and a list of literature review search terms conducted can be found in Appendix B.2.3.

2.4 Covariate data sources

The cornerstone of a regression analysis is the relationship between the covariate(s) (i.e. independent variables) and the outcome(s) (i.e. dependent variables). Thus, covariates play an essential role in regression-based models, including those presented in this thesis. Here, I discuss the various sources of covariate data used in this work, and how these data are processed to develop nationally comparable covariate time series.

2.4.1 Overview of covariates used in this thesis

We use covariate data in both the input and prediction datasets for our neonatal COD models. Specifically, we run regression models with covariates as the independent variables and the associated COD inputs as the dependent variables. We then apply the estimated regression coefficients for these covariates to a set of nationally comparable covariate time series data to obtain COD distribution predictions by country and year. Further details of our neonatal COD modelling strategy, including the use of covariates, are provided in section 4.2. Discussion of how covariate data availability and quality impact our work is included in section 5.1.3.

Since the goal of our work is to produce nationally comparable time series estimates, we were only able to include in our models those covariates which had annual, nationally comparable times series estimates available across 194+ countries. The WHO collates and processes various nationally comparable covariate time series data annually. For our neonatal COD estimation work, we were provided with a database of all relevant covariate time series data by the WHO. Table 2.1 includes a description of these covariates alongside the data sources used by the WHO. In the following sections, I describe the raw data sources relevant to these covariates, and give a brief description of the types of processing which are done to develop complete covariate time series.

Table 2.1: Description of covariates relevant to the neonatal cause-of-death models

Covariate	Source[1]	Description
Neonatal mortality rate (NMR)	UN-IGME	Number of deaths in first 28 days from birth per 1,000 live births
Infant mortality rate (IMR)	UN-IGME	Number of deaths in first 1 year from birth per 1,000 live births
Under-5 mortality rate (U5MR)	UN-IGME	Number of deaths in first 5 years from birth per 1,000 live births
Low birthweight (LBW)	World Bank	% of babies weighing <2500 g within first hours of birth
General fertility rate (GFR)	MMEIG[2]	Total number of live births divided by number of women 15-49 years old in a given year
Antenatal care (ANC)	World Bank	% of women with 1+ antenatal care visit during pregnancy
Female literacy rate (FLR)	World Bank	% of women aged 15+ who can read and write at least short simple statements
Diphtheria/Pertussis/ Tetanus vaccine (DPT)	WHO-UNICEF	% of children ages 12-23 months immunized with DPT vaccine
Bacillus Calmette-Guerin vaccine (BCG)	WHO-UNICEF	% of children ages 12-23 months immunized with BCG vaccine
Protected at birth (PAB) against tetanus	WHO-UNICEF	% of births by women of child-bearing age who are protected against tetanus
Skilled birth attendance (SBA)	WHO-UNICEF	% of births with a skilled attendant at delivery
Gross national income (GNI)	World Bank	Sum of money earned per year by people and business in a country (in US dollars)
GINI coefficient	World Bank	Statistical measure of economic inequality based on income distribution in a population

[1] Main source as listed in WHO database; [2] MMEIG = UN Maternal Mortality Inter-agency Group

2.4.2 From "raw" empirical covariate data to nationally comparable time series

As noted in the previous section, the covariates listed in Table 2.1 were provided to us in a database by the WHO as complete nationally comparable time series. For example, values for LBW were available in this database from 1980 to 2015 for 194 countries (i.e. 6,984 country-year datapoints). There is a multi-stage process involved in going from incomplete "raw" empirical data from multiple sources to a single complete nationally comparable covariate time series. A simplified example of the process is as follows: 1) data for a (somewhat) standardized covariate are collected over time in different countries through various means (e.g. household surveys, national reporting systems, censuses); 2) these data are collated and processed (e.g. weighted averages) by a secondary source (e.g. World Bank, UN-IGME; see "source" column in Table 2.1) to produce nationally comparable time series data for the covariate; and 3) this time series is further processed (e.g. for completeness, smoothing) by the WHO prior to being made available for use in estimation models. Thus, the 6,984 datapoints for a complete nationally

comparable 35-year covariate time series for 194 countries may be a combination of "raw" empirical data, slightly processed empirical data (e.g. weighted averages if there are multiple sources with data for same country/year), and modelled data (e.g. imputation, extrapolation, smoothing).

The "raw" data sources typically include a variety of standardized household surveys (e.g. DHS, MICS), other surveys (e.g. country- or topic-specific, facility-based), and data from national statistical systems (e.g. GNI data reported by countries). Each of these have strengths and limitations. For example, household surveys such as DHS allow for standardized implementation of questionnaires across many countries, but have important limitations (e.g. recall bias). A few covariate time series rely on only one or two sources of underlying raw data, but the majority have data from a combination of sources. For example, the World Bank reports that their SBA (i.e. "skilled birth attendance") covariate time series is collated from UNICEF, State of the World's Children, ChildInfo, and DHS data sources [104]. The large number of data sources has the potential to decrease comparability of the covariate. For example, primary data collection sources may apply different definitions for what a "skilled attendant" means for the SBA covariate.

The collation and processing steps use a diverse set of methods. For example, various methods used within or between covariates listed in Table 2.1 include weighted averages, Bayesian B-spline smoothing, Loess regression smoothing, flat-trend extrapolation, regression-based extrapolation, linear interpolation, backward/forward progression based on annual rates of change (ARC), and backward/forward progression based on average of flat trend and ARC trend [104-106]. This extensive list is indicative of the data gaps and quality issues that exist when trying to develop comparable time series. The main implication of this is that there can be substantial amounts of measurement error between and within national covariate time series data, both at the level of primary data collection and through the collation and processing of these data. These issues are discussed in greater detail in sections 2.5 and 5.1.3.

2.4.3 Local covariate data collection

An additional source of covariate data which we use in our work is locally collected data reported in VA studies. As noted in section 2.3.5, we extract cause-of-death data from verbal autopsy studies for inputs into our models. We also try to extract covariate information from the same geographic area, population, and time period as the extracted cause-of-death data. This is done because regression-based models inherently depend on the relationship between

outcome variables and covariates, which should ideally come from the same population and time period. For example, a study may report the local NMR, which may be substantially different from national- or subnational NMR estimates. However, we are not always able to find corresponding local-level covariate information, and thus use regional- or national-level covariates when such local-level data are missing. Additionally, studies often state a covariate value without explaining its source, thus making it difficult to gauge its accuracy and method of primary data collection.

2.5 Discussion

The data used in this thesis come from a diverse range of sources of varying availability and quality. This speaks to the dearth of strong data collection infrastructure in many countries and the complexities of the data being collected. High-quality mortality and cause-of-death data, such as those from well-functioning CRVS systems, are a rich source of information but are limited by the number of countries with such systems. Commonly used alternatives, such as household surveys and verbal autopsy methods, provide much-needed data in low-resource high-mortality settings but with important quality limitations.

There is also considerable heterogeneity in data availability and quality for the covariates used in our neonatal cause-of-death models. The covariate data in our models are a mixture of data reported by countries, determined from household surveys, modelled, and/or imputed. A particular covariate time series may have a combination of these sources across countries or over time for the same country. Such variation may not always be known or fully understood, and is usually impossible to specifically identify within the given time series data.

The data quality and quantity issues discussed here have the potential to affect predictive accuracy and stability of our models. In particular, the various types of measurement error – both for outcomes and covariates – can mask true covariate-cause relationships, which are the cornerstone of regression analyses. This would weaken our model's ability to produce robust and stable estimates. It is important to keep these data issues in mind when reading the rest of the thesis. The implications of various data sources and how they are used in our work are discussed throughout the thesis, particularly in chapters 5 and 8. In section 5.1, I more closely investigate how several data issues, including the noise from measurement error, may affect our estimates.

Theme 1: Filling knowledge gaps

3 Timing of neonatal deaths

In this chapter, I present our work on estimating the distribution of deaths by day in the neonatal period. Preliminary results were published in Save the Children's State of the World's Mothers 2013 report [12], and final results were published in the *Lancet Global Health* under the following citation: *Oza S, Cousens SN, Lawn JE. Estimation of the daily risk of neonatal death, including the day of birth, in 186 countries in 2013: a vital-registration and modelling-based study. Lancet Global Health 2014; 2(11): e635-644* [107]. The text in this chapter is adapted from this journal article and its associated web appendix. The open-access article can be found at: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(14)70309-2/

3.1 Introduction

Birth and the following few days are biologically and emotionally remarkable, and yet also the riskiest. While the Millennium Development Goals (MDGs) galvanised efforts leading to major reductions in maternal and child mortality in recent years, deaths during the neonatal period (first four weeks after birth) have decreased at a substantially slower rate [108, 109]. The estimated average annual rate of mortality reduction for neonates was 2.2% from 1990-2013 [110], compared to 4.0% for children aged 1-59 months [110] and 2.6% for maternal deaths [111]. The high risk of death in the first days after birth is striking. In 2013, 45% of the estimated 6.2 million deaths in children under the age of five occurred during the neonatal period [110], along with an estimated 1.2 million intrapartum stillbirths [112]. Around three-quarters of neonatal deaths are estimated to occur during the early neonatal period (see Panel 3.1 for neonatal time period definitions) [13, 113].

Panel 3.1: Definitions for the neonatal period

- The *neonatal period* refers to the first four weeks (28 days) after birth.
- The *early neonatal period* consists of the first week (7 days) and the *late neonatal period* consists of the last three weeks (21 days) of the neonatal period.
- The *first day* is typically called "day 0" in household survey and vital registration data or "day 1" in clinical practice. The neonatal period ranges from days 0-27 when using "day 0" as the first day, and from days 1-28 when using "day 1" instead.

Here, we use the term "day 0" to refer to the first day. Thus, we use days 0-6 for the early neonatal period (week 1), days 7-27 for the late neonatal period (weeks 2-4), and days 0-27 for the full neonatal period.

Deaths on the day of birth (day 0) are particularly important to assess because they account for a large number of deaths that can be targeted by interventions at birth. A dramatic fall in risk occurs even within hours of birth. The risk of death (per 1,000 live births) in the first hour after birth in the US is 0.91 [10]. In contrast, the risk in the following 23 hours is 1.58 which translates to an average hourly risk of about 0.07, indicating a dramatic decline in risk. The causes of day 0 deaths more closely resemble those of intrapartum stillbirths than of deaths later in the neonatal period. This has led some to propose an indicator combining intrapartum stillbirths and day 0 deaths as a marker of the quality of intrapartum care [11], as these are far more common than maternal deaths and thus offer a more practical indicator. Yet, there have been no systematic, nationally comparable estimates for risk during the first day or week of life, despite increasing programmatic focus on these important time periods.

Vital registration (VR) data, collected through birth and death certificates, are available for more than half of the 193 UN member states [114]. However, these data are reliable for only about half of these countries, which are generally the wealthiest and account for fewer than 5% of the world's 3 million estimated neonatal deaths in 2013 [115]. For the majority of countries in the world, day-of-death data for the neonatal period are either unavailable or are derived from cross-sectional surveys, such as the Demographic and Health Surveys (DHS), which ask women of reproductive age how many of their children have died and the child's age at death [116]. Such data are susceptible to error with possible underreporting of deaths, including stillbirths, and misreporting of the day of death [117-119]. Particularly important for the present work is the potential for misrecording of deaths between day 0 and day 1, and misclassification between stillbirths and very early neonatal deaths (Panel 3.2).

Panel 3.2: Measurement challenges in the neonatal period

Misclassification between stillbirths and live births: The probability of recording the baby as being alive at birth has been shown to be associated with the perception of viability of survival. For example, if babies are not assessed at birth and resuscitated, a live term baby that is not breathing may be misclassified as an intrapartum stillbirth [112, 120]. This is typically the direction in which misclassification between live births and stillbirths occurs. Note: stillbirths are categorized as antepartum (those occurring before the onset of labour) and intrapartum (those occurring after the onset of labour).

Variation of registration with gestational age: As complexity of care increases, and even very preterm babies under 25 weeks of gestation are given intensive care, the registration of live very preterm babies under 28 weeks increases, as documented in Denmark [121]. Some countries without care for extremely preterm babies still may not count some live born babies as live births.

Misrecording of day of death: Importantly, the perception of what time period this "first day" refers to can lead to critical differences in recording practices. Some different ways in which the first day is perceived include 1) the first 24 hours after birth, 2) until sundown of the day of birth, or 3) the calendar date. These various perceptions can lead to differences in how deaths are classified between "day 0" and "day 1" and can result in misclassification if survey interviewers or respondents misunderstand "day 1" as meaning the day of birth. Even in the absence of such misunderstanding, deaths on "day 0" will not include all deaths within 24 hours of birth if, for example, calendar dates are used. Additionally, misclassification between days is possible if only the dates but not the exact times of birth and death are reported in vital registration systems.

Hence, to provide estimates for countries lacking recent VR data of adequate quality, modelling exercises remain necessary. Regular estimates of under-five and neonatal mortality by country are provided by the United Nations (UN) and the Institute for Health Metrics and Evaluation (IHME) [102, 110, 115, 122]. However, while it is well recognised that mortality is highest during the first few days, we were unable to identify any multi-country analyses of the daily risk of death within the first week or month after birth, with current standard life tables and survival curves grouping the time periods in months or years instead of by day [123]. Although the literature on modelling survival curves within the neonatal period is sparse, previous work has suggested that exponential functions are suitable for modelling the neonatal period, while other functions, like the Gompertz and Weibull, are better for later periods in the lifespan [124, 125].

Exponential functions have been used to smooth day of death heaping (preference for reporting deaths on certain days, e.g. “at one week”) within the neonatal period [117]. Previous analyses of DHS suggested that up to 50% of neonatal deaths occur in the first 24-48 hours, but highlighted data limitations including misclassification between day 0 and day 1, and “heaping” of deaths on particular days [13].

Here, we report results obtained using a mathematical model with a good fit to the available empirical data and present global, regional, and national estimates of risk and number of deaths (with uncertainty ranges) for the day of birth, first week of life, and the late neonatal period for 186 countries in 2013.

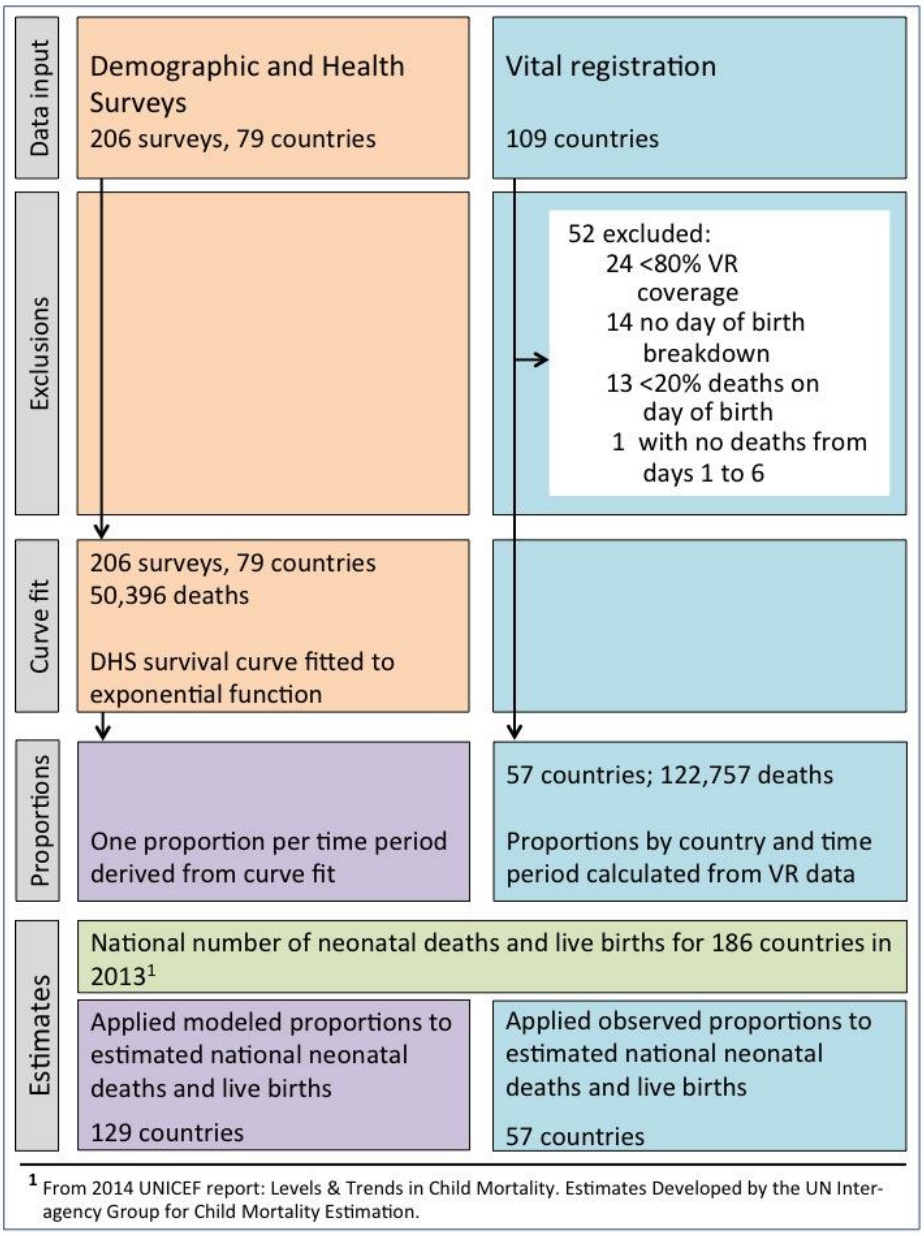
3.2 Methods

3.2.1 Data inputs

First, we searched PubMed, Web of Knowledge, Medline, and Google Scholar with various search terms that covered timing of deaths in the neonatal period. The search terms included “neonatal” or “newborn” and one or more of “day of birth”, “day 0”, “day 1”, “deaths”, “mortality”, “risk of death by day”, “daily risk”, “survival”, “survival curves”, “day 0 risk”, “day 1 risk”, “temporal distribution”, “day of death”, and “time of death”. We found no systematic national estimates of daily risk of death during the neonatal period.

We then reviewed data on the timing of neonatal deaths from two main sources. First, we obtained the most recent publicly available VR data from the World Health Organization (WHO) for the years 2006 to 2010. These VR data are reported by countries to the WHO and thus may not reflect the full extent of relevant data recorded in a country’s CRVS system, especially for countries with multiple registration systems. For Canada, we used data from Statistics Canada rather than WHO due to the availability of more recent data. Second, we acquired DHS data from 1986-2011 using the STATcompiler tool from MEASURE DHS (http://www.measuredhs.com/). Finally, to derive estimates of the absolute risk of death and numbers of deaths by time period, we applied our results to the 2013 estimates of neonatal deaths and live births produced by the UN Inter-agency Group for Mortality Estimation (UN-IGME) [110]. An overview of the data inputs, exclusion criteria, and estimation techniques is shown in Figure 3.1.

Figure 3.1: Strategy for neonatal day-of-death analysis



3.2.2 Data inclusion and exclusion criteria

We used the reported VR data to generate national risk estimates if the country had 1) VR coverage of adult mortality of 80% [126] or higher and 2) the VR data available classified neonatal deaths into the following time periods: day 0, days 1-6, and days 7-27. We used these three categories because the VR dataset from the WHO did not provide a more detailed breakdown of neonatal deaths by day. For the 50 countries with more than 50 neonatal deaths in the most recent year with available data, we used the data from the most recent year. For seven countries with fewer than 50 neonatal deaths in the most recent year with data available, we combined deaths from the previous two to five years to avoid instability due to small

numbers, stopping as soon as the total number of neonatal deaths was 50 or higher. We excluded data from countries with ≤20% of deaths on day 0 (n=13) or no deaths on days 1-6 (n=1) on the grounds that these proportions are implausibly low based on the data from all countries and thus likely indicate poor data quality (see section 3.4 for further details).

We obtained DHS data on neonatal deaths and live births for 206 surveys in 79 countries. While 15 surveys had fewer than 50 neonatal deaths, we did not exclude these because we combined deaths across surveys in our mathematical model, as described below. Similarly, we sought to account for misclassification of deaths between days 0 and 1 in these surveys through our model rather than discarding implausible data.

Lists of included and excluded data are provided in Appendix A.1.

3.2.3 Model fitting

All statistical analyses were undertaken using Stata version 12 (www.stata.com). We postulated a model for the risk of neonatal death by day of life. We then applied this model to the DHS data to estimate the proportion of neonatal deaths occurring on each day of the neonatal period for countries without adequate VR data. Our model assumes that the probability of dying on day t conditional on surviving until that day declines exponentially. In addition, the model allows the probability of dying on day 0 to differ from this pattern. Mathematically, this can be expressed as:

$$h_t = \begin{cases} \alpha & t = 0 \\ \beta\gamma^{t-1} & 1 \le t \le 27 \end{cases} \qquad \text{(Eq. 3.1)}$$

where h_t is the probability of dying on day t given survival until that day. Given this model, the unconditional probability of dying on day t of the neonatal period, p_t, is:

$$p_t = \begin{cases} \alpha & t = 0 \\ (1-\alpha) * \beta & t = 1 \\ (1-\alpha) * (\beta\gamma^{t-1}) \prod_2^t (1-\beta\gamma^{\tau-2}) & t \ge 2 \end{cases} \qquad \text{(Eq. 3.2)}$$

The probability of surviving the neonatal period, p_s, is then given by $(1-\alpha) * \prod_1^{27}(1-\beta\gamma^{\tau-1})$. Thus, the probability of dying in the neonatal period is $1-(1-\alpha) * \prod_1^{27}(1-\beta\gamma^{\tau-1})$. So the probability of dying on a given day conditional on dying in the neonatal period (i.e. the expected proportion of neonatal deaths on day t) is:

$$m_t = \begin{cases} \frac{\alpha}{1-(1-\alpha)*\prod_1^{27}(1-\beta\gamma^{\tau-1})} & t = 0 \\ \frac{(1-\alpha)*\beta}{1-(1-\alpha)*\prod_1^{27}(1-\beta\gamma^{\tau-1})} & t = 1 \\ \frac{(1-\alpha)*(\beta\gamma^{t-1})\prod_2^{t}(1-\beta\gamma^{\tau-2})}{1-(1-\alpha)*\prod_1^{27}(1-\beta\gamma^{\tau-1})} & t \geq 2 \end{cases} \quad \text{(Eq. 3.3)}$$

Based on the multinomial distribution, we can write an expression for the likelihood of observing n_0, …, n_{27} deaths in the neonatal period (days 0 to 27) given N live births and the proportion surviving the neonatal period p_S as:

$$p_0{}^{n_0} * p_1{}^{n_1} * p_2{}^{n_2} \dots p_{27}{}^{n_{27}} * p_S{}^{N-\sum_0^{27} n_t} = p_S{}^{N-\sum_0^{27} n_t} * \prod_0^{27} p_t{}^{n_t} \quad \text{(Eq. 3.4)}$$

To deal with potential misclassification between days 0 and 1 in the DHS data, we combined observed deaths on days 0 and 1 and re-wrote the likelihood as:

$$(p_0 + p_1)^{n_0+n_1} * p_2{}^{n_2} \dots p_{27}{}^{n_{27}} * p_S{}^{N-\sum_0^{27} n_t} = p_S{}^{N-\sum_0^{27} n_t} * (p_0 + p_1)^{n_0+n_1} * \prod_2^{27} p_t{}^{n_t}$$
(Eq. 3.5)

Maximum likelihood estimation

We used maximum likelihood to estimate the parameters α, β, and γ. For this, we used maximum likelihood estimation implemented by the built-in Stata command "ml". To do this, we first wrote a log-likelihood function based on the mathematical function described in Equation 3.5 and assumed that the number of deaths observed on each day together with the number of survivors follows the multinomial probability distribution. The log-likelihood function L(α,β,γ) is built from the following functions L_t(α,β,γ) which indicate the log-likelihood of observing n_t deaths on day t with n_s survivors of the neonatal period:

$$L_t(\alpha,\beta,\gamma) = \begin{cases} \ln(\alpha + (1-\alpha)*\beta) * n_t & t = 0.5 \\ \ln\left((1-\alpha)*(\beta\gamma^{t-1})\prod_2^t 1-\beta\gamma^{\tau-2}\right) * n_t & 2 \leq t \leq 27 \\ \ln\left((1-\alpha)*\prod_1^{27}(1-\beta\gamma^{\tau-1})\right) * n_s & \text{contribution of survivors} \end{cases} \quad \text{(Eq. 3.6)}$$

where *t* is the day of death (0.5 represents days 0 and 1 combined) and the arguments for the maximum likelihood equation are α, β, and γ. This is obtained by taking the log of the probability of observing n_t deaths of day t when the probability of dying on day t is p_t, where p_t is a function of α, β, and γ. Finally, the overall log-likelihood L(α, β, γ) is obtained by summing the log likelihoods for the different days: L(α, β, γ) = $\sum_t L_t\,(\alpha,\beta,\gamma)$.

The input dataset contained information on the number of deaths by day (with deaths on days 0 and 1 combined as deaths on "day 0.5"). We then used the "ml model lf" and "ml maximize" commands in Stata, which maximize the equation L(a,b) above using Stata's default method, which is a modified Newton-Raphson algorithm.

Daily proportions of neonatal deaths

This model allows us to estimate a "corrected" proportion of neonatal deaths on day 0 under the assumption encoded in the model that the probability of dying on subsequent days declines exponentially. Using these estimates, we calculated the expected proportion of neonatal deaths on a given day (Equation 3.3) and during various time periods, including days 0-6, days 1-6, and days 7-27. We initially applied the model to the aggregated DHS data. We also fitted the model to subsets of the data categorized by neonatal mortality rate (NMR), national income category, geographic region, and study period to assess whether there was any evidence that the proportional distribution of deaths varied by any of these factors.

We compared this model to a simpler 2-parameter model, which assumes that $\gamma=1$, using a likelihood ratio test (see section 3.4). Because we wished to correct for misreporting between day 0 and 1, we needed to use a model in which the relationship between day 1 deaths and deaths on subsequent days was constrained. Otherwise, the proportions of deaths occurring on each of days 0 and 1 are not identifiable with the likelihood in Equation 3.5.

3.2.4 Estimation of daily neonatal mortality risk and numbers by country

We generated estimates for 186 countries with 1,000 or more live births in 2013. These countries account for an estimated 2.76 million neonatal deaths, greater than 99% of the global total.

For the 57 countries with adequate VR data, we calculated the proportion of deaths on day 0 and days 0-6 directly from the data to produce the final risk estimates. For the remaining 129 countries, we used the day 0 and days 0-6 proportions estimated using our model. We calculated the number of deaths by applying the day 0 and days 0-6 proportions to the UN 2013 neonatal death estimates. We then derived the day 0, days 0-6, and days 7-27 risks by dividing these time period-specific numbers of deaths by the 2013 live births in the country.

3.2.5 Uncertainty estimation

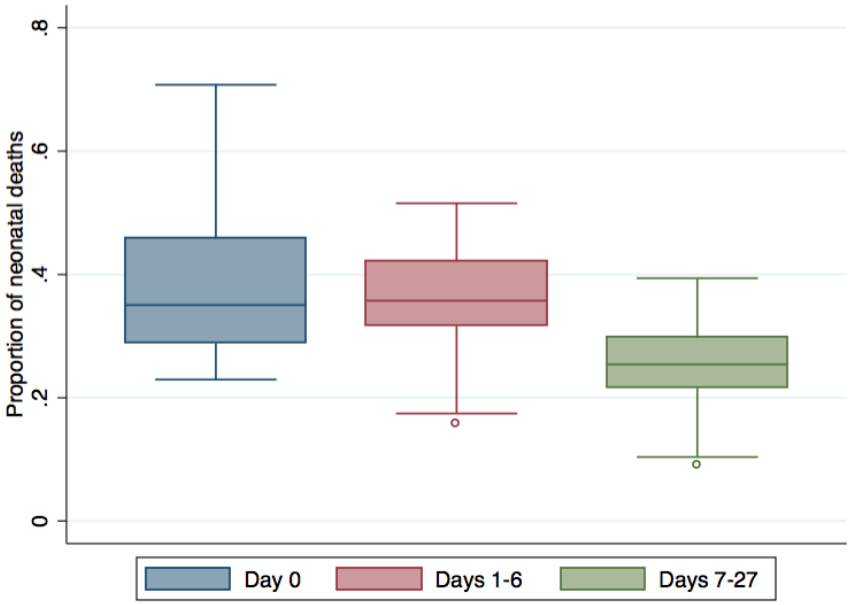
We developed uncertainty estimates for the modelled proportions by drawing 1,000 bootstrap samples with replacement from the 206 DHS in the input dataset. We then re-ran the analysis to estimate the model parameters and used these to estimate the proportion of deaths by day for each of these 1,000 datasets. Finally, we took the 2.5^{th} and 97.5^{th} centiles from the resulting distributions of these proportions for our uncertainty range. These uncertainty estimates do not include uncertainty in total neonatal deaths. For countries where we used VR data, we calculated the uncertainty for the proportions by assuming a Poisson distribution for the number of deaths during those periods (i.e. the standard error is equal to the square root of the reported number of deaths).

3.3 Main results

3.3.1 Description of data inputs

The input dataset, after applying the exclusion criteria, consisted of data from 57 VR countries (median year: 2010) with 122,757 neonatal deaths and 206 DHS (median year: 1999) with 50,396 neonatal deaths in the 5 years prior to the survey for 79 countries. Nine countries had both VR and DHS data. Thus, the final input dataset contained information from 127 countries and included 173,153 neonatal deaths (Figure 3.1). For VR countries, the median proportions of neonatal deaths occurring in different periods were as follows: 0.35 (interquartile range [IQR]: 0.29-0.46) on day 0; 0.36 (IQR: 0.32-0.42) on days 1-6; and 0.25 (IQR: 0.22-0.30) on days 7-27 (Figure 3.2). The median proportion of deaths in the first week (i.e. days 0-6) was 0.75 (IQR: 0.70-0.78). The proportion of deaths on day 0 was relatively variable compared to the proportions for the other two time periods (Figure 3.2). The three countries with the highest proportions (with 95% confidence intervals) of deaths on day 0 were Switzerland (0.71 ± 0.04), Canada (0.69 ± 0.02), and Austria (0.62 ± 0.07), and the three with the lowest proportions were the Czech Republic (0.23 ± 0.06), Belize (0.24 ± 0.10), and Macedonia (0.25 ± 0.07).

Figure 3.2: Proportion of neonatal deaths for VR countries on days 0, 1-6, and 7-27

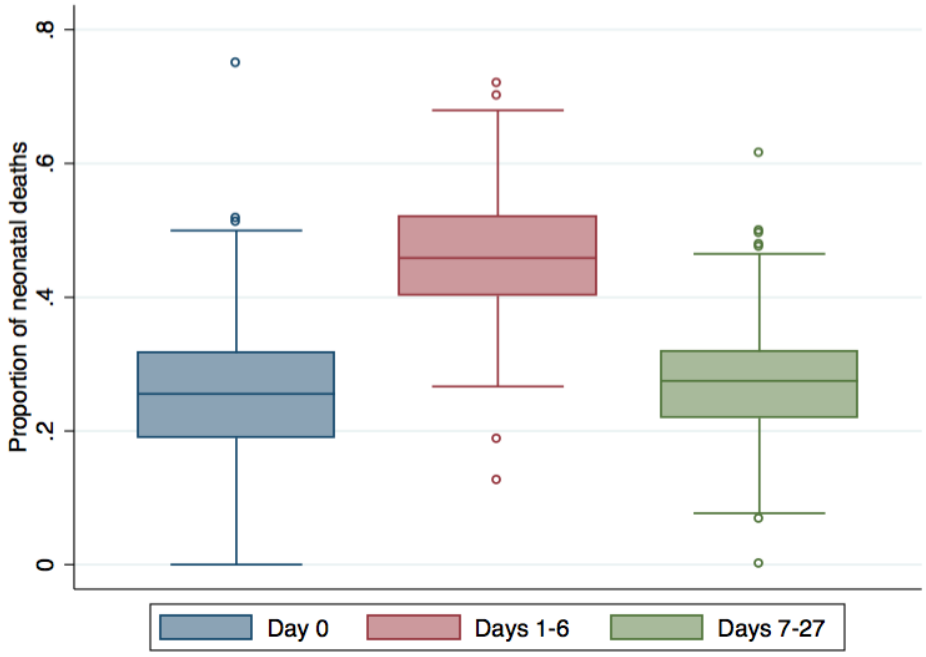


Note: VR = vital registration

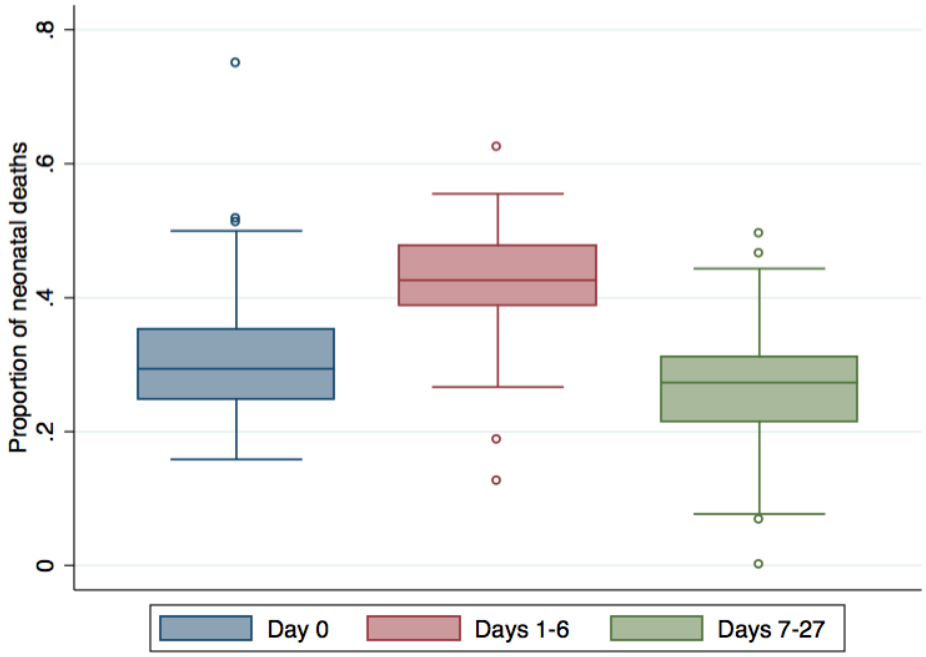
Of the 79 countries with DHS, 29 countries had one survey while 50 had between two and six surveys each. Across all of the surveys, the median proportion of reported deaths was 0.26 (IQR: 0.19-0.32) on day 0, 0.19 (IQR: 0.15-0.24) on day 1, 0.46 (IQR: 0.40-0.52) on days 1-6, and 0.72 (IQR: 0.68-0.78) for the first week (Figure 3.3a). The median proportion for days 0 and 1 combined was 0.46 (IQR: 0.39-0.52). Sixty-six DHS (32%) had higher proportions of deaths on day 1 than day 0, suggesting substantial misclassification of deaths between these days. For surveys with higher day 0 than day 1 proportions, the median proportions of reported deaths were 0.29 (IQR: 0.25-0.35) on day 0, 0.16 (IQR: 0.13-0.21) on day 1, 0.43 (IQR: 0.39-0.48) for days 1-6, and 0.73 (IQR: 0.69-0.78) for the first week (Figure 3.3b).

Figure 3.3: Proportion of neonatal deaths for DHS on days 0, 1-6, and 7-27

a) all 206 DHS



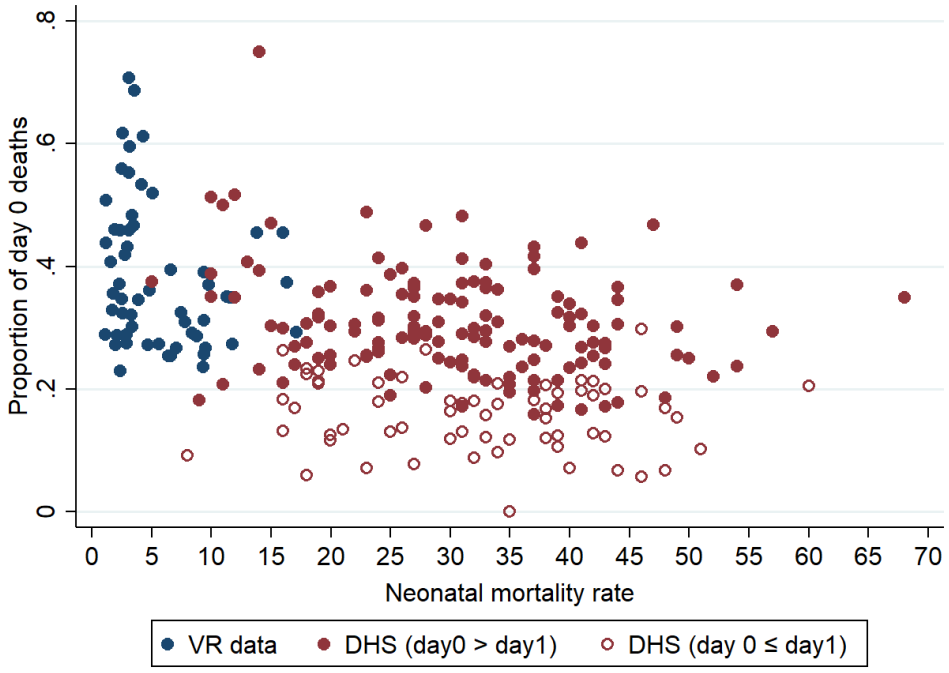
b) 140 DHS that reported more deaths on day 0 than day 1



Note: DHS = Demographic and Health Surveys

Figure 3.4 shows a plot of the reported proportion of deaths on day 0 against NMR for VR countries and DHS. The DHS data are likely to include day 0/1 misclassification, ranging from severe to mild or none depending on the survey. The only discernible pattern is that several countries with very low NMR (<5), which are all countries with high-quality VR data, have higher reported day 0 proportions.

Figure 3.4: Proportion of day 0 deaths reported in VR and DHS by neonatal mortality rate

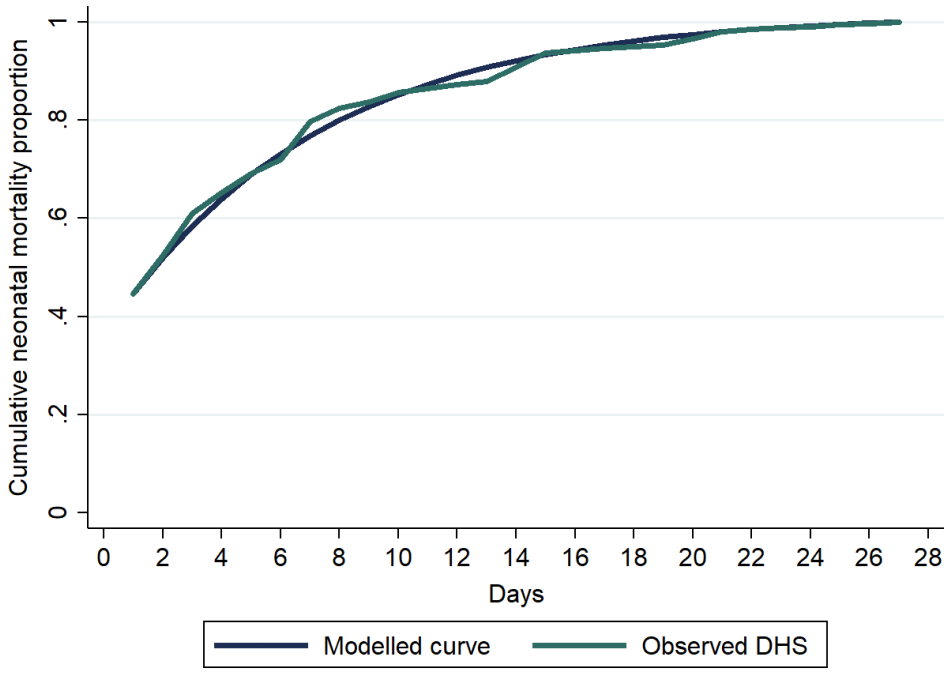


Note: severe underreporting of day 0 deaths or misrecording between day 0 and 1 in DHS (fewer day 0 than day 1 deaths) is noted by hollow circles; VR = vital registration; DHS = Demographic and Health Surveys.

3.3.2 Model fitting

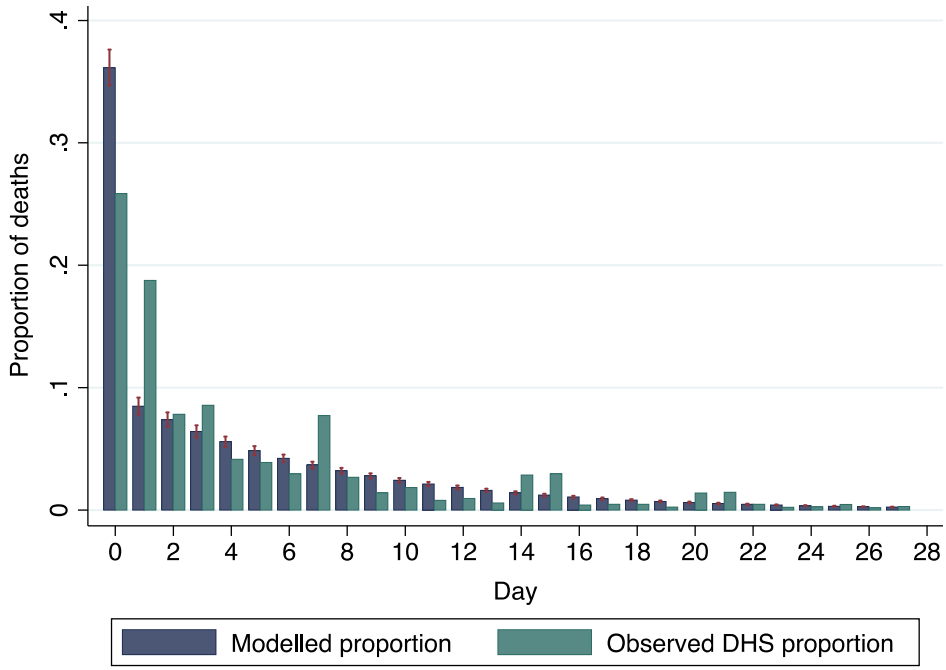
We summed the reported births and neonatal deaths by day across the 206 DHS. The 3-parameter model fitted the observed DHS data better than the 2-parameter model ($p<0.0001$). Visual inspection of the modelled distribution of deaths by day compared to the observed distribution in the DHS data also indicated a good fit (Figure 3.5), with the poorest fit occurring on days 7, 14, and 15. This likely reflects day-of-death heaping at 1 and 2 weeks of age (Figure 3.6). Based on these results, we used the 3-parameter model in the remainder of this work. The estimated parameter values (from Equation 3.1) were $\alpha=0.012$ (uncertainty range [UR]: 0.010-0.014), $\beta=0.003$ (UR: 0.002-0.003), and $\gamma=0.872$ (UR: 0.868-0.875).

Figure 3.5: Observed and modelled cumulative mortality in the neonatal period (conditional on dying in the neonatal period)



Note: the combined proportion of day 0 and day 1 deaths is about 0.45 in both the observed data and modelled estimates; DHS = Demographic and Health Surveys.

Figure 3.6: Observed and modelled proportions of deaths in the neonatal period



Note: red bars show 95% confidence intervals; DHS = Demographic and Health Surveys.

The modelled proportions for deaths during the neonatal period were 0.36 (UR: 0.34-0.38) on day 0 and 0.73 (UR: 0.72-0.74) for days 0-6. These estimates are very similar to the median proportions observed in the VR data: 0.35 for day 0 and 0.75 for days 0-6. The probability of dying on a given day and the cumulative probability of surviving that day conditional on 1) survival up to that day and 2) dying during the neonatal period, is given in Table 3.1.

Table 3.1: Probability of dying and cumulative probability of surviving by day in the neonatal period

Day	Probability of dying on the given day	Cumulative probability of surviving	Day	Probability of dying on the given day	Cumulative probability of surviving
0	0.36	0.64	**14**	0.01	0.08
1	0.08	0.55	**15**	0.01	0.07
2	0.07	0.48	**16**	0.01	0.06
3	0.06	0.42	**17**	0.01	0.05
4	0.06	0.36	**18**	0.01	0.04
5	0.05	0.31	**19**	0.01	0.03
6	0.04	0.27	**20**	0.01	0.02
7	0.04	0.23	**21**	0.01	0.02
8	0.03	0.20	**22**	0.00	0.01
9	0.03	0.17	**23**	0.00	0.01
10	0.02	0.15	**24**	0.00	0.01
11	0.02	0.13	**25**	0.00	0.00
12	0.02	0.11	**26**	0.00	0.00
13	0.02	0.09	**27**	0.00	0.00

If we do not correct for day 0 and 1 misclassification (i.e. by running our model with day 0 and 1 separate), the estimated proportions for the DHS data are 0.26 on day 0 (UR: 0.24-0.27) and 0.73 (UR: 0.72-0.74) for days 0-6.

Analysis across subsets of data

We also summed births and neonatal deaths by day across subsets of surveys based on NMR level, income level, region, and time period and fitted the model to these data (see Appendix A.2 for a list of countries by region and income category). The estimated day 0 proportion did not vary importantly by NMR level or income category (Table 3.2, Figures 3.7a-b). There does appear to be some variation between regions and by survey period (Table 3.2, Figures 3.7c-d). We combined data from the Western Asia, Northern Africa, and Caucasus/Central Asia MDG regions to form a "mid-east" region for this DHS analysis to avoid small numbers and because of the similarities in the data and the health system context for these regions. Among the 28 datasets from this combined region, 75% (n=21) had fewer day 0 deaths than day 1 deaths,

suggesting widespread undercounting of day 0 deaths and/or misreporting of day 0 deaths as day 1 in surveys from this region. In comparison, 26% of surveys from Latin America and the Caribbean, the region with the next highest percentage of such surveys, had more day 1 than day 0 deaths. The estimated day 0 proportions for the other regions do not show much variation. If the "mid-east" region is excluded from the analysis, the overall results remain largely unchanged with 0.36 (UR: 0.35-0.38) for the day 0 proportion and 0.73 (UR: 0.72-0.74) for the week 1 proportion. For survey period, earlier surveys had, on average, lower proportions of deaths in the first few days than later surveys. If only surveys from 2000 or later are included in the analysis, the proportions are 0.39 (UR: 0.37-0.41) for day 0 and 0.75 (UR: 0.73-0.76) for week 1.

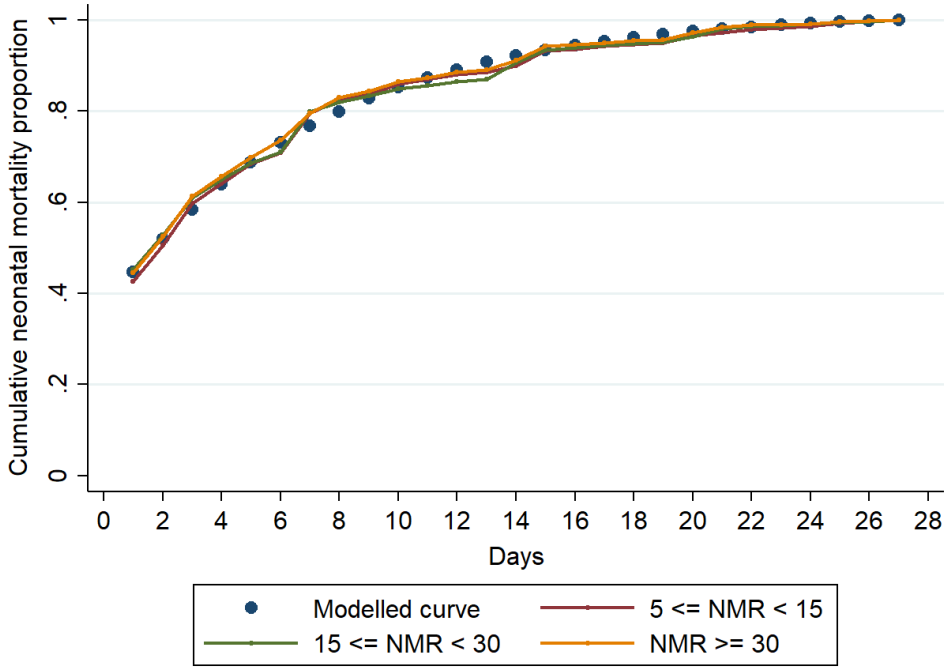
Table 3.2: Estimated proportions of day 0 and week 1 deaths (with uncertainty ranges) by neonatal mortality rate, income, region, and survey period

		Day 0	Week 1
Neonatal mortality rate	5 ≤ NMR < 15	0.34 (0.31-0.38)	0.72 (0.71-0.74)
	15 ≤ NMR < 30	0.37 (0.35-0.39)	0.73 (0.71-0.74)
	NMR ≥ 30	0.36 (0.33-0.38)	0.74 (0.72-0.76)
Income	Low	0.36 (0.34-0.39)	0.73 (0.72-0.74)
	Lower middle	0.36 (0.33-0.38)	0.73 (0.72-0.74)
	Upper middle	0.38 (0.35-0.41)	0.73 (0.70-0.75)
Region	East Asia & Southeast Asia	0.39 (0.33-0.43)	0.74 (0.72-0.78)
	Southern Asia	0.36 (0.33-0.39)	0.73 (0.71-0.75)
	Sub-Saharan Africa	0.37 (0.34-0.39)	0.74 (0.72-0.75)
	Latin America/Caribbean	0.39 (0.36-0.42)	0.72 (0.70-0.74)
	North Africa/West & Central Asia	0.28 (0.25-0.32)	0.70 (0.68-0.71)
Period	1986-1995	0.32 (0.30-0.35)	0.71 (0.69-0.73)
	1996-2005	0.37 (0.35-0.39)	0.73 (0.72-0.75)
	2006-2011	0.41 (0.37-0.43)	0.76 (0.73-0.78)
Overall		0.36 (0.34-0.38)	0.73 (0.72-0.74)

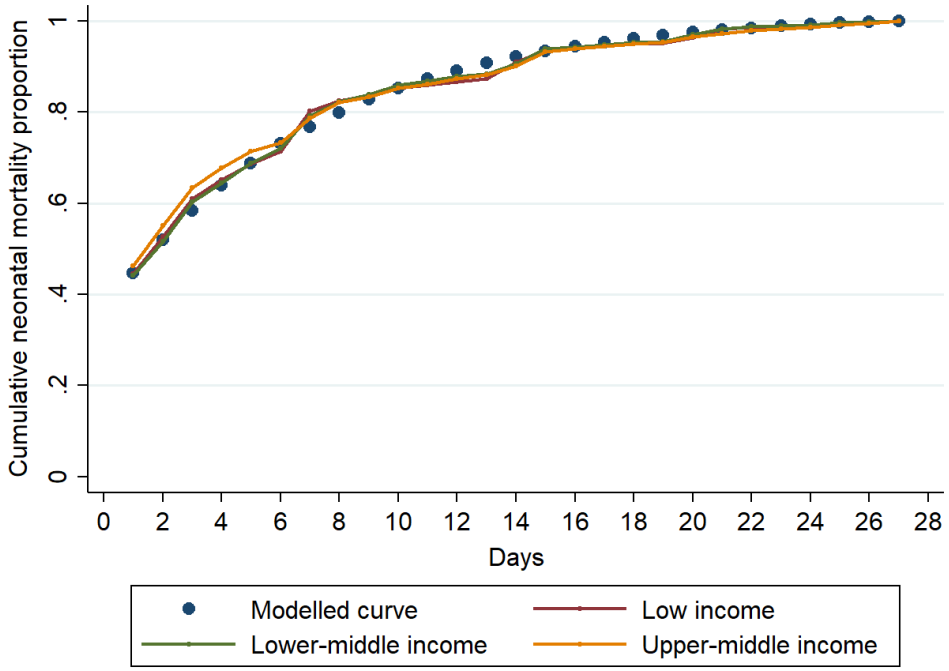
Figure 3.7: Observed and modelled cumulative mortality in the neonatal period (conditional on dying in the neonatal period) by neonatal mortality rate, income, region, and survey period for DHS

Note: These include data from 206 DHS (50,396 neonatal deaths) from 1986-2011.

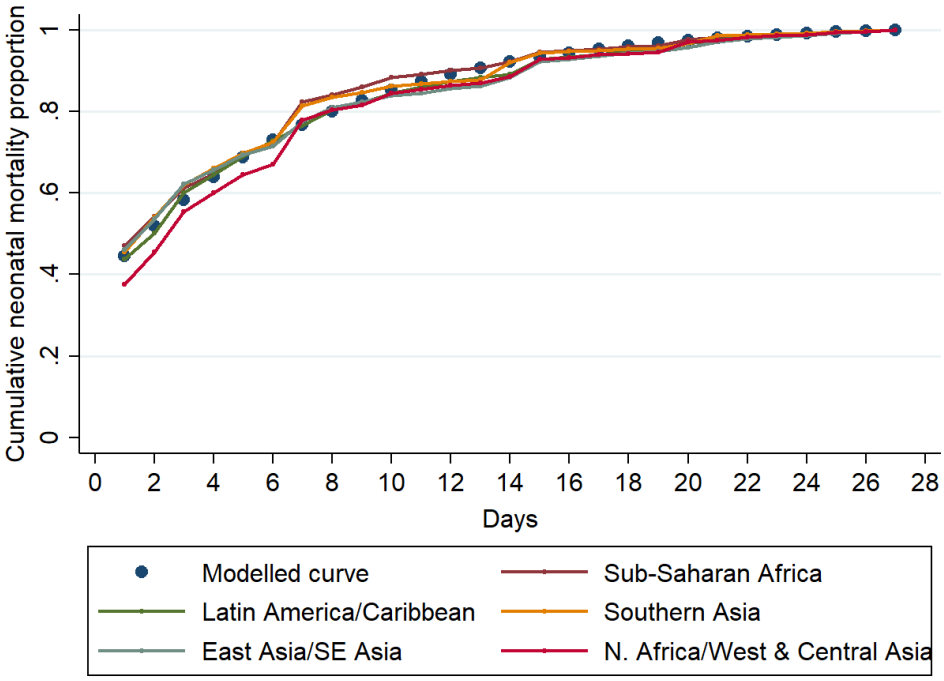
a) Neonatal mortality rate



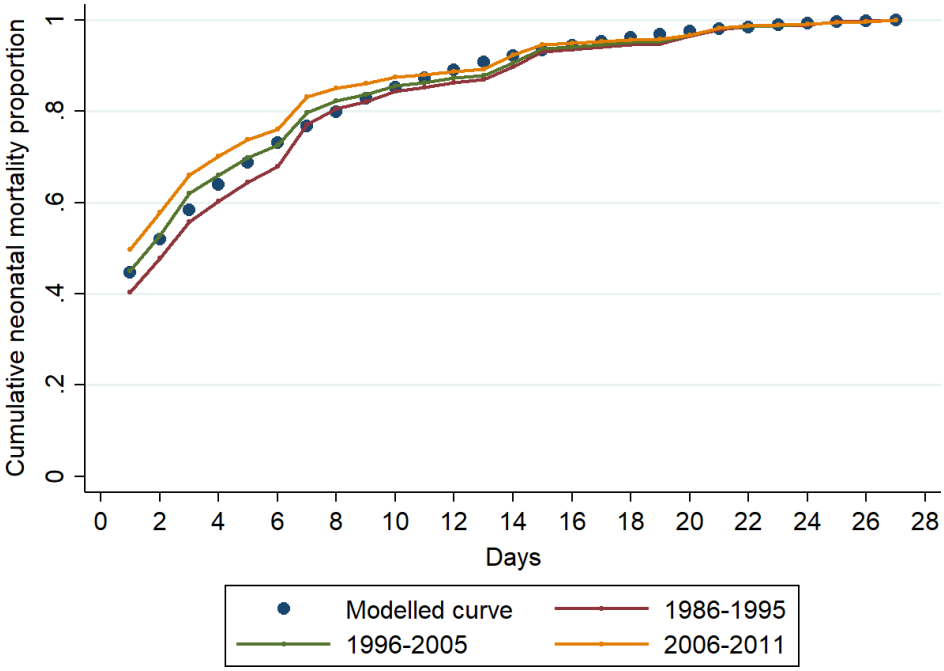
b) Income category



c) Geographic region



d) Survey period



In theory, our model requires the total number of live births to be known as well as neonatal deaths per day. However, we found that varying the number of live births while keeping the number of deaths fixed across wide ranges of NMR (from 1 to 1,000) resulted in negligible changes (<0.5 percentage points) to the estimated day 0 and week 1 proportions. Thus, in practice, the results do not appear to be sensitive to the number of live births.

3.3.3 Risk and numbers by neonatal period

An estimated 2.76 million neonatal deaths occurred in the 186 countries included in this analysis [110]. Of these, an estimated 1.00 million (36.3%) (UR: 0.94 million – 1.05 million) occurred on day 0 and 2.02 million (73.2%) (UR: 1.99 million – 2.05 million) occurred within the first week (including on day 0).

Regional estimates are provided in Table 3.3. Sub-Saharan Africa had the highest risk (deaths per 1,000 live births), with 11.2 (UR: 10.6-11.8) on day 0 and 21.5 (UR: 21.2-21.8) in week 1. Southern Asia had a lower risk but a larger number of births, and thus the largest number of deaths, with 392,300 (UR: 369,100-412,500) on day 0 and 793,300 (UR: 781,500-803,300) during week 1.

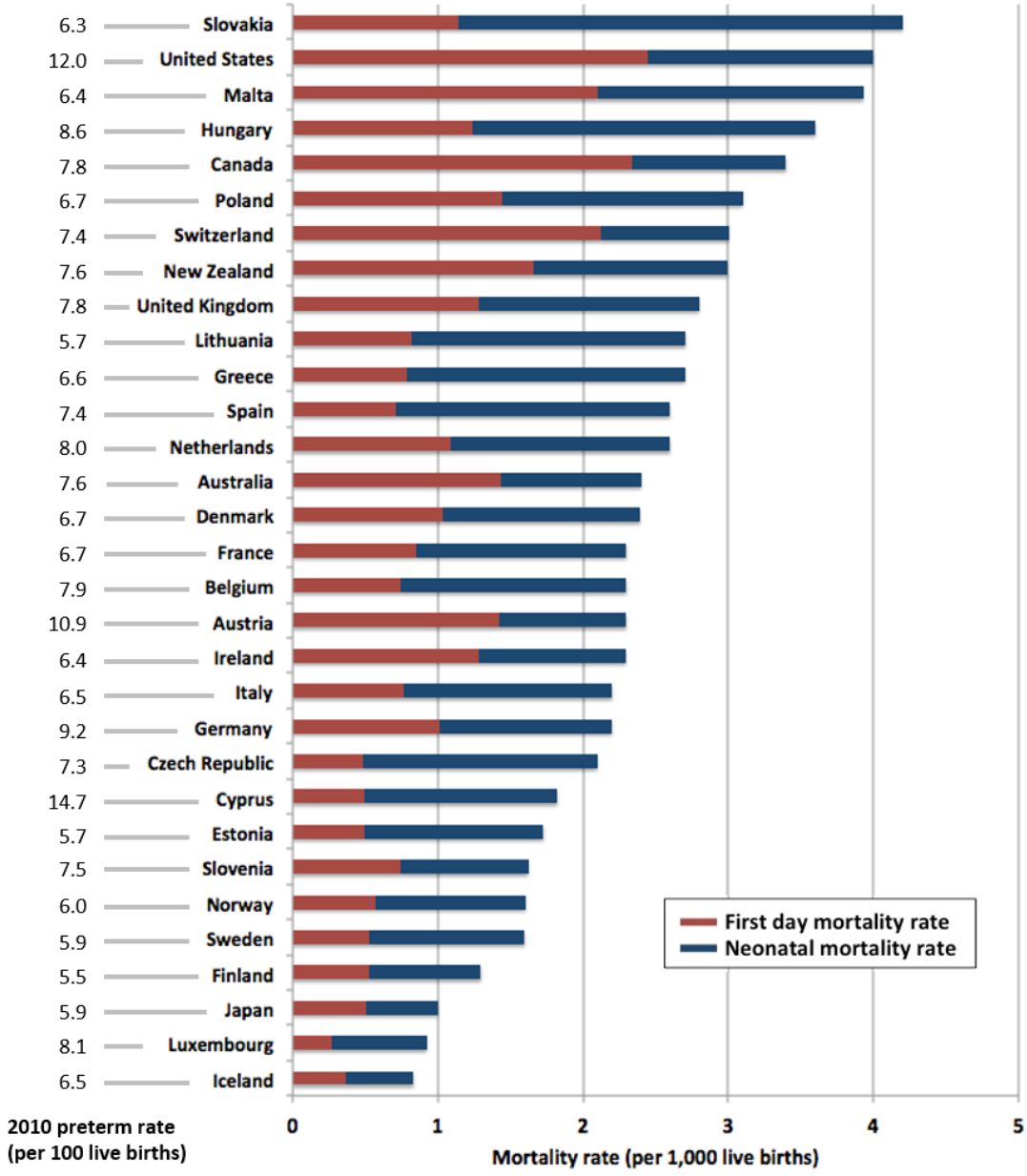
Table 3.3: Risk and number of deaths (with uncertainty ranges) by MDG region for day 0, week 1, and weeks 2-4 in 2013

	Day 0		Week 1		Weeks 2-4	
	Risk[1,2]	**Deaths[2] (1000s)**	**Risk**	**Deaths (1000s)**	**Risk**	**Deaths (1000s)**
Sub-Saharan Africa	11.2 (10.6-11.8)	385.2 (362.6-404.9)	22.7 (22.4-23.0)	779.1 (767.5-788.9)	8.4 (8.1-8.7)	286.2 (276.6-297.8)
Southern Asia	10.6 (10.0-11.2)	392.3 (369.1-412.5)	21.5 (21.2-21.8)	793.3 (781.5-803.3)	7.9 (7.7-8.2)	292.0 (282.2-303.9)
Oceania	7.7 (7.3-8.1)	2.0 (1.9-2.1)	15.6 (15.4-15.8)	4.1 (4.1-4.2)	5.7 (5.6-6.0)	1.5 (1.5-1.6)
Caucasus / Central Asia	5.4 (5.1-5.7)	9.5 (8.9-10.0)	11.0 (10.8-11.2)	19.2 (18.8-19.5)	3.8 (3.6-3.9)	6.6 (6.3-6.9)
South-eastern Asia	5.2 (4.9-5.5)	58.0 (54.5-60.9)	10.5 (10.4-10.6)	117.2 (115.5-118.7)	3.9 (3.7-4.0)	43.1 (41.7-44.9)
Western Asia	4.9 (4.6-5.2)	24.0 (22.6-25.3)	10.0 (9.8-10.1)	48.6 (47.9-49.3)	3.7 (3.6-3.8)	18.0 (17.3-18.7)
Northern Africa	4.8 (4.5-5.1)	19.3 (18.1-20.3)	9.7 (9.6-9.8)	39.0 (38.4-39.5)	3.6 (3.5-3.7)	14.3 (13.9-14.9)
Latin America / Caribbean	3.2 (3.1-3.4)	35.4 (33.8-36.9)	6.8 (6.7-7.0)	74.3 (72.7-75.9)	2.4 (2.3-2.5)	26.2 (25.1-27.3)
Eastern Asia	2.8 (2.6-2.9)	54.3 (51.1-57.1)	5.6 (5.5-5.7)	109.8 (108.2-111.2)	2.1 (2.0-2.2)	40.4 (39.1-42.1)
Developed regions	1.6 (1.5-1.7)	22.9 (21.6-24.2)	2.6 (2.5-2.7)	36.6 (35.2-38.1)	0.8 (0.7-0.9)	11.5 (10.6-12.4)
World	7.3 (6.9-7.6)	1002.7 (944.2-1054.1)	14.7 (14.4-14.9)	2021.3 (1989.7-2048.5)	5.4 (5.2-5.6)	739.8 (714.3-770.6)

[1] risk is deaths per 1,000 live births; [2] uncertainty ranges do not include uncertainty in total neonatal deaths

In the USA and Canada, the risk of death on day 0 was 2.4 and 2.3 per 1,000 live births, respectively, whereas in several Northern European countries (e.g. Norway, Sweden, and Finland), the risk was 0.6 or lower. Figure 3.8 shows the risk on day 0 and the overall neonatal period, alongside the preterm birth rate, for 31 industrialized countries with high-quality VR.

Figure 3.8: Risk of death on day 0 and during the neonatal period in 2013 (alongside preterm birth rates) for 31 countries with high-quality VR data



Note: the preterm birth rates are for 2010 from Blencowe et al [121].

The risk of death (per 1,000 live births) ranged widely for the 186 countries, from 1 to 47 for the full neonatal period, 1 to 34 for the first week, and <1 to 17 for day 0. Nine of the ten countries with the highest risk were in Sub-Saharan Africa. The risk of death for these ten countries

ranged from 14 to 17 on day 0 and 29 to 34 in the first week (Table 3.4). Full details of country-specific estimates are in the web appendix of the published paper [107].

Table 3.4: Risk of death per 1,000 live births (with uncertainty ranges) within the neonatal period for the ten countries with highest risks in 2013

Country	Day 0	Week 1	Weeks 2-4
Angola	17 (16-18)[1]	34 (34-34)	13 (12-13)
Somalia	17 (16-18)	34 (33-34)	12 (12-13)
Sierra Leone	16 (15-17)	32 (32-33)	12 (12-12)
Guinea-Bissau	16 (15-17)	32 (32-33)	12 (11-12)
Lesotho	16 (15-17)	32 (31-32)	12 (11-12)
Central African Republic	16 (15-16)	31 (31-32)	12 (11-12)
Pakistan	15 (14-16)	31 (30-31)	11 (11-12)
Mali	15 (14-15)	29 (29-30)	11 (10-11)
Chad	14 (14-15)	29 (29-29)	11 (10-11)
Zimbabwe	14 (13-15)	29 (28-29)	11 (10-11)

[1] Uncertainty estimates do not include uncertainty in total neonatal deaths

The number of deaths during each time period also varied widely. The ten countries with the largest numbers of deaths are dominated by populous countries, but are also affected by the level of risk. The number of deaths in these ten countries ranged from 14,300 to 270,100 on day 0 and 28,900 to 546,300 in the first week (Table 3.5).

Table 3.5: Number of deaths in the thousands (with uncertainty ranges) within the neonatal period for the ten countries with the most neonatal deaths in 2013

Country	Day 0	Week 1	Weeks 2-4
India	270.1 (254.1-284.0)[1]	546.3 (538.2-553.1)	201.1 (194.3-209.3)
Nigeria	94.4 (88.8-99.3)	191.0 (188.1-193.4)	70.3 (67.9-73.2)
Pakistan	70.0 (65.9-73.6)	141.6 (139.5-143.3)	52.1 (50.4-54.2)
China	51.8 (48.7-54.4)	104.7 (103.1-106.0)	38.5 (37.2-40.1)
DRC[2]	37.8 (35.6-39.7)	76.4 (75.3-77.4)	28.1 (27.2-29.3)
Ethiopia	30.5 (28.7-32.1)	61.7 (60.8-62.4)	22.7 (21.9-23.6)
Bangladesh	27.7 (26.1-29.2)	56.1 (55.2-56.8)	20.6 (19.9-21.5)
Indonesia	23.8 (22.4-25.0)	48.1 (47.4-48.7)	17.7 (17.1-18.4)
Angola	15.4 (14.5-16.2)	31.1 (30.7-31.5)	11.5 (11.1-11.9)
Kenya	14.3 (13.4-15.0)	28.9 (28.5-29.3)	10.6 (10.3-11.1)
Total[3]	*649.7 (611.3-683.2)*	*1314.0 (1294.5-1330.5)*	*483.7 (467.5-503.4)*

[1] Uncertainty estimates do not include uncertainty in total neonatal deaths; [2] DRC = Democratic Republic of the Congo; [3] the total is based on rounded estimates.

3.4 Sensitivity and validation analyses

In this section, I describe sensitivity and validation exercises we performed.

3.4.1 Inclusion of excluded vital registration data

As noted in section 3.2, the decision to exclude VR data for countries with implausibly low proportions of deaths on day 0 (≤20% of neonatal deaths) or days 1-6 (≤10% of neonatal deaths) was based on evaluating the data from all countries. Using these criteria, thirteen countries had ≤20% of neonatal deaths on day 0 and one country had 0% deaths on days 1-6. Such low proportions are likely an indicator of poor data quality due to issues such as underreporting, misclassification between live births and stillbirths, and/or misrecording of deaths between days. Therefore, we used the modelled proportions rather than actual VR data for the final estimates for these countries. Since the VR data were not used in the model themselves, these exclusion criteria only affect the data for the individual countries rather than the overall model. In Table 3.6 we have included the day 0 and week 1 proportions from the actual VR data for these countries for comparison.

Table 3.6: Vital registration data for countries with implausibly low early neonatal death proportions

Country	Year	Day 0 proportion	Days 1-6 proportion	Week 1 proportion
Azerbaijan	2007	0.02	0.84	0.86
Bahamas	2008	0	0.64	0.64
Barbados	2008	0.09	0.54	0.63
Cuba	2010	0.16	0.51	0.67
Fiji	2009	0	0.69	0.69
Grenada	2010	0.05	0.77	0.82
Guatemala	2008	0.12	0.52	0.63
Latvia	2010	0.70	0	0.70
Montenegro	2009	0.20	0.55	0.75
Romania	2010	0.16	0.51	0.67
Saint Lucia	2008	0.04	0.77	0.81
Saint Vincent / Grenadines	2010	0.02	0.77	0.78
Sri Lanka	2006	0	0.75	0.75
Suriname	2009	0.01	0.71	0.71

Using the proportions above for these countries instead of applying our modelled proportions would result in 3,800 fewer day 0 deaths and 600 fewer week 1 deaths in total. Note that these proportions, as stated above, were not used in the final estimates as we believed their values

were implausibly low, suggesting poor data quality (and potentially noisy data from the smaller countries).

3.4.2 Validation exercises

We conducted validation exercises, including out-of-sample validation and the addition of VR data to the model.

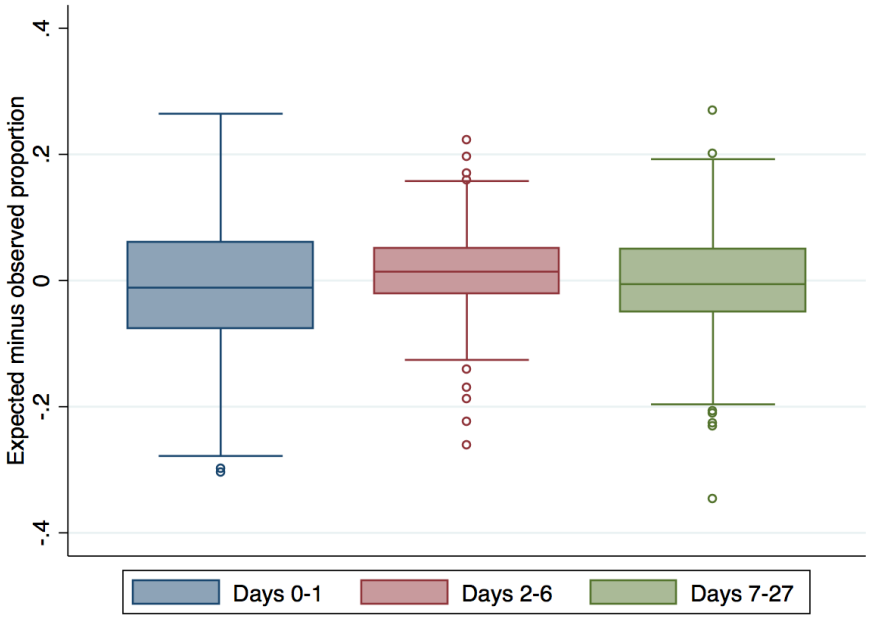
Out-of-sample validation

We performed out-of-sample validation to see how well our model predicted the observed DHS data. Our model predicts a higher proportion of day 0 deaths than the observed because of our day 0 correction, and thus we do not expect very good "agreement" between the observed data and modelled estimates for day 0. Therefore, we performed the validation for days 0-1 (combined), 2-6, and 7-27. We performed this out-of-sample validation using the jackknife approach. For this, we left out one survey at a time (n=206) and re-estimated the α, β, and γ parameters using maximum likelihood with the remaining data points. We then compared the modelled results for days 0-1, days 2-6, and days 7-27 to those observed in the DHS that was left out.

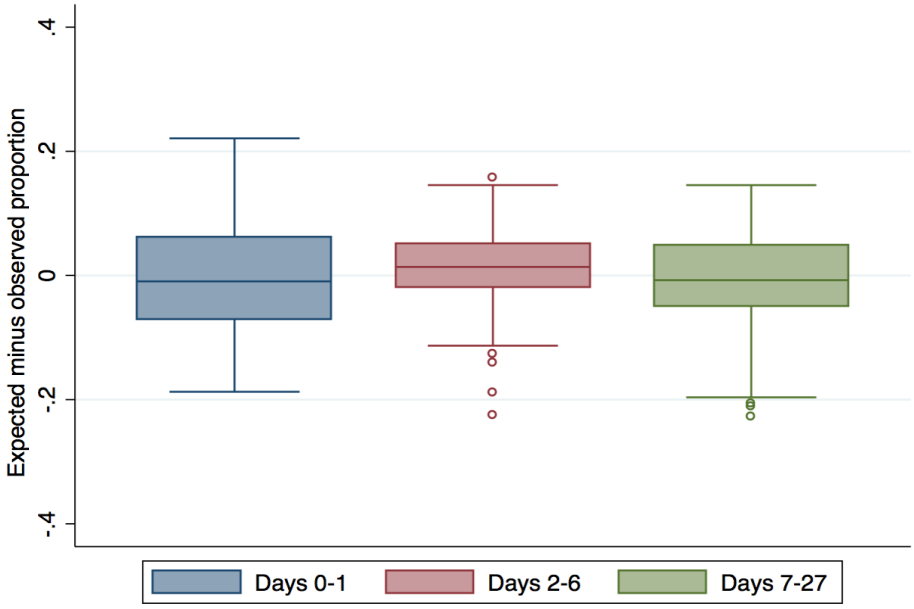
Figure 3.9a shows the distribution of the differences in the predicted and observed proportions based on the out-of-sample validation exercise. The median difference in predicted and observed proportions is as follows: 1) -1 percentage point for days 0-1 combined, 2) 1.4 percentage points for days 2-6, and 3) -0.8 percentage points for days 7-27. Dropping small surveys (i.e. those with <50 neonatal deaths) resulted in similar median differences but fewer "outliers" (Figure 3.9b). None of the time periods had a 25^{th} or 75^{th} percentile difference between predicted and observed proportions greater than 7 percentage points. The mean absolute differences between the predicted and observed proportions are 3.4 percentage points for days 0-1, 2.7 percentage points for days 2-6, and 3.1 percentage points for days 7-27.

Figure 3.9: Differences between predicted versus observed day 0-1, 2-6, and 7-27 values for out-of-sample validation

a) *all 206 DHS*



b) *191 surveys with ≥50 reported neonatal deaths*



Given that the observed proportions are themselves subject to sampling error, we also tested whether the predicted proportions lay inside or outside the 95% confidence intervals (CI) of the observed proportions. We found that the predicted proportions falling outside of the 95% CIs for observed proportions were as follows: 1) 28 out of 206 studies (14%) for days 0-1; 2) 26 (13%) for week 1; and 3) 26 (13%) for days 7-27. Since underreporting of early neonatal deaths

would lead to lower observed proportions in the first days and week, we repeated this analysis after removing surveys with potentially severe underreporting (as discussed in section 3.4.1). After doing this, we found that the predicted proportions falling outside of the 95% CIs for observed proportions were as follows: 1) 12 out of 180 (7%) for days 0-1; 2) 12 (7%) for week 1; and 11 (6%) for days 7-27.

These results suggest that for a high proportion of surveys there is good agreement between the observed and modelled results. It is also important to note that the DHS are of varying quality, which may affect the comparison of observed and modelled results.

Including vital registration data in the model

We also performed an analysis in which we added contributions from the VR data to the likelihood function. We ran two versions of this analysis, one with the USA included and one with it excluded. The reason we excluded the USA is because it is an extreme outlier (95% percentile) for day 0 deaths (61%) while contributing far more deaths than any other country/survey. The day 0 and week 1 proportions were as follows: 1) USA excluded – 0.36 (day 0) and 0.75 (week 1); 2) USA included – 0.39 (day 0) and 0.75 (week 1). These proportions, particularly when the USA is excluded, are quite similar to the results we found in our DHS-only analysis (day 0: 0.36 and week 1: 0.73).

3.4.3 Alternative mathematical models

To determine if simpler models performed as well as the 3-parameter model (Equation 3.1) we used, we conducted the additional analyses described here. As noted in section 3.3, the 2-parameter model fitted less well than the 3-parameter model, and we therefore rejected using this simpler model in our analysis.

Two-parameter model

In this model, which is a simplification of the 3-parameter model, we assumed that $\gamma = 1$. Therefore, the probability of dying on a day $t > 0$ equals the constant β. The likelihood ratio test comparing this model to the 3-parameter model provided strong evidence that the 3-parameter model fitted the data better than the 2-parameter model ($p<0.0001$).

More complex models

We were unable to test more complex models because high-quality daily day-of-death data for the neonatal period were unavailable. While we did use high-quality VR data in our analysis, these data only contained three data points: day 0, days 1-6, and days 7-27 and these data could therefore not be used to test models with greater complexity than our 3-parameter model.

3.4.4 Assumption of Poisson error for vital registration data

In estimating uncertainty intervals for countries with high-quality VR data, we assumed a Poisson distribution for the number of deaths during the neonatal time periods. Others have found that, at least for older ages, the Poisson distribution may underestimate the actual variance and thus a negative binomial distribution may be more appropriate [127].

In theory, we should be able to assess the variance in our dataset to determine if our Poisson assumption is appropriate. This is difficult to do with the data we have since we effectively have one dataset per country instead of multiple samples of mortality data for a country. We do, however, have VR data from previous years. We tested the Poisson distribution assumption for each VR country using this data (for years 2000 and later) under the assumption that there was no real change in the day of death distribution over time. We found that there was no evidence against the Poisson distribution for 52 of the 57 countries. For the other five countries, it is difficult to gauge whether the data are truly overdispersed or if our assumption about the lack of time trend is incorrect. However, since variation in the day of death distribution over time should appear as overdispersion, the lack of evidence against the Poisson distribution for 52 of 57 countries suggests that our assumption that the neonatal day of death distribution data are Poisson distributed is reasonable.

3.5 Identifying data quality issues in Demographic and Health Surveys

In our paper, we identified some data quality issues in DHS that may affect the interpretation around timing of neonatal deaths. Here, we suggest analytical checks for gauging whether a given DHS may have misrecording of deaths between days 0 and 1 and/or underreporting of very early neonatal deaths.

3.5.1 Misrecording of day 0 and day 1 deaths

As discussed earlier, a number of DHS appear to have misrecording of deaths between days 0 and 1. While the most severe cases are identifiable through having more day 1 deaths than day

0 deaths, other surveys may have milder misrecording. Based on the analysis in which we sought to correct for this misrecording, we estimated that, on average, 80% of day 0 + day 1 deaths occur on day 0. Based on this ratio, the following equation can be used to determine if a DHS may have substantial misrecording issues:

$$0.8(d_0 + d_1) - d_0 > 1.2\sqrt{d_0 + d_1} \qquad \text{(Eq. 3.7)}$$

where d_0 is the number of deaths on day 0 and d_1 is the number of deaths on day 1. If Equation 3.7 holds for a given dataset, this suggests that some day 0 deaths have been misclassified as having occurred on day 1. We derived Equation 3.7 as follows: 1) the left side of the equation is the difference between the expected day 0 deaths (out of day 0 + day 1 deaths) and the observed day 0 deaths, and 2) the right side is 3 standard deviations for the number of day 0 deaths (i.e. 3*sqrt(0.8*0.2)*sqrt(d0 + d1)), which yields 1.2*sqrt(d0+d1). We chose to use 3 standard deviations to account for real variation in the d0:d1 ratio. If the true ratio of d0:d1 deaths is 0.8:0.2, we would only expect 5 surveys in 1,000 to breach this cut-off in the absence of misclassification.

When this equation was applied to the surveys in our analysis, 74% (152/206) of surveys showed some evidence of misclassification.

3.5.2 Underreporting of very early neonatal deaths

If there is underreporting of very early neonatal deaths, we expect that the proportion of day 0 and day 1 deaths (summed together because we are assuming misrecording between those days may exist) will be low. Based on the analysis in our paper, we found that, on average, 45% of neonatal deaths occurred on days 0 and 1. Using a similar approach to that for Equation 3.7, the following equation can be used to determine if a DHS may have underreporting of very early neonatal deaths:

$$0.45d - (d_0 + d_1) > 1.5\sqrt{d} \qquad \text{(Eq. 3.8)}$$

where d_0 is the number of deaths on day 0, d_1 is the number of deaths on day 1, and d is the sum of deaths across all 28 days. If Equation 3.8 holds for a given dataset, then this suggests evidence of underreporting of very early neonatal deaths. Specifically, it means that this distribution of deaths would arise by chance only 0.5% of the time in the absence of

underreporting. We derived Equation 3.8 (using a similar approach to Equation 3.7) as follows: 1) the left side of the equation is the difference between the expected day 0+1 deaths (out of all neonatal deaths) and the observed day 0+1 deaths, and 2) the right side is 3 standard deviations for the number of day 0+1 deaths (i.e. 3*sqrt(0.45*0.55)*sqrt(*d*)) where *d* is the # of neonatal deaths. This yields 1.5*sqrt(*d*) on the right side. We chose to use 3 standard deviations to account for real variation in the (d_0+d_1):d ratio. Underreporting of very early neonatal deaths is a serious issue, as it may result in an underestimate of the neonatal mortality rate.

When this equation was applied to the surveys in our analysis, 13% (26/206) surveys showed some evidence of underreporting of very early neonatal deaths. Of these surveys, 85% (22/26) also showed some evidence of misclassification between day 0 and day 1 deaths (section 3.5.1).

3.6 Discussion

We estimated the risk of dying and number of deaths for the day of birth, first week of life, and the late neonatal period for 186 countries in 2013. Along with our preliminary results [12], to our knowledge this was the first systematic multi-country analysis of the daily risk of neonatal death around the world. Of the estimated 2.76 million neonatal deaths worldwide, approximately 36.3% of deaths occurred on the day of birth and 73.2% occurred within the first week. Hence, more than one million babies die on their day of birth, in addition to 1.2 million intrapartum stillbirths estimated to occur each year. This highlights the fact that the hours just before birth and the first few days of life are the riskiest in the human lifespan, a period which also carries an elevated risk for the mother.

For countries without high-quality VR data, we estimated the proportional distribution of neonatal deaths by day by aggregating DHS data from countries with a range of NMRs. The parameter estimates we obtained (for α, β, γ) thus represent "average" values of these parameters, which we used to estimate the average distribution of deaths by day. We then applied this average distribution to all countries without high-quality VR data. We examined whether it was appropriate to apply the same average distribution to countries with different NMRs by fitting the model to subsets of the data defined by NMR level. The proportion of deaths on day 0 appears remarkably consistent across countries with different NMR (and income) levels (Table 3.2, Figure 3.4, Figures 3.7a-b). The proportion of day 0 deaths showed some regional variation, but this may reflect underreporting or misclassification of day 0 deaths. The proportion of day 0 deaths varied slightly with time period of the survey, with evidence of

earlier surveys having a lower proportion of day 0 deaths than later surveys. However, we chose to include all survey years in our model because it is not clear whether this is a real change in proportion over time. A 2008 study concluded that in a number of countries, enumeration of child deaths was poorer in the more recent surveys compared to earlier ones for the same country [128]. Additionally, several countries with multiple surveys have wide fluctuations in their proportion of deaths on day 0, which are not consistently in an upward direction. Since the day 0 proportion is slightly higher when restricting the analysis to surveys from 2000 or later, we chose a conservative approach to estimating day 0 proportions.

We noted variation in the proportion of day 0 deaths for some very low mortality countries (NMR <5 per 1,000) (Figure 3.4). We might expect that countries with comprehensive neonatal intensive care would have a higher proportion of neonatal deaths on day 0 and the first week since many late neonatal deaths – which are classically due to infections – should have already been prevented [129, 130]. However, one effect of neonatal intensive care is also to shift first day deaths to later days, for example through ventilation of very preterm babies who die later from complications like intracranial haemorrhage or infection. Thus, while the overall risk is lower, the proportion may remain similar due to some deaths shifting to later days or even beyond the neonatal period.

While there is a greater than 30-fold difference in risk of death on day 0 between the poorest and richest countries, there is also, surprisingly, an almost 10-fold difference in day 0 risk across the richest countries. Given the very high quality of data collection and intensive care in these countries, it is likely that this variation is real and not an artefact of underreporting. The much higher preterm birth rate in North America may be a contributor, especially in the USA, where 12%, or over half a million, of all births are preterm each year (Figure 3.8) [121].

The exponential function we used as our model fitted the DHS data well. We applied the modelled estimates to countries with no day-of-death data, and also to those with DHS because of the substantial day 0/day 1 misclassification biases evident in some individual DHS. One-third of DHS reported more deaths on day 1 than day 0, which is biologically implausible. In fact, when comparing countries that had both VR and DHS data (with 50+ neonatal deaths), no DHS with more day 1 than day 0 deaths were supported by the VR data in that country. Others are likely to have misclassification that is less obvious (i.e. more day 0 than day 1 deaths, but with some misrecorded as day 1). We tried to correct for this using our mathematical model on the combined surveys. However, we did not account for misclassification between stillbirths and

early neonatal deaths, which is another well recognized issue in DHS [131]. If neonatal deaths in the first minutes of life are recorded as stillbirths (which is the most common direction of misclassification), very early neonatal deaths will be undercounted, and we would expect the proportion of deaths during week 1 to be lower than the average regardless of day 0/day 1 misclassification. Nineteen of the 206 DHS had observed week 1 proportions with uncertainty ranges that fell outside our estimated uncertainty range. Of these, 7 had more day 1 than day 0 deaths. The remaining 12 all had low proportions for week 1, with a median week 1 proportion of 0.60 (IQR: 0.56-0.63) versus 0.73 (IQR: 0.69-0.78) for the other 194 surveys. This pattern is consistent with undercounting of early neonatal deaths. Finally, several of the countries with multiple DHS had fluctuations in day 0 and week 1 proportions that are unlikely to be explained by real changes. For example, the day 0 proportion in Ethiopia varied from 0.30 in 2000, to 0.19 in 2005, to 0.42 in 2011. Due to these data quality issues (and random error) within individual surveys, we chose to apply our model to countries with DHS data to predict their day 0 and week 1 proportions instead of using their raw DHS data. In section 3.5, I presented simple analytical methods to identify DHS with substantial misclassification of deaths between days 0 and 1 and underreporting of very early neonatal deaths.

We hope that our estimates will be improved upon as better data become available. While our DHS-based model appears to be robust, it is clearly not ideal to apply a single model to all countries lacking reliable VR data of their own. Although we believe that, on average, our results likely represent the day 0 and week 1 proportions for many of the modelled countries, this approach will mask any variation that does exist between countries. Since the same proportions for day 0 and week 1 were applied to all non-VR countries, the rankings are tied to the variation in NMR and in the number of neonatal deaths occurring in the country. Also, our uncertainty ranges do not reflect the uncertainty in NMR, which was unavailable for the most recent NMR estimates. Thus, as with all modelled estimates, our results represent a starting point for understanding the burden of day 0 and week 1 deaths in each country. If relevant high-quality VR data were available for individual days, we would be able to compare our DHS-based model against these survival curves, and test more complex models. For example, since we attempted to correct for misreporting of deaths between days 0 and 1, we assumed that from day 1 onwards the daily hazard declines exponentially. While the model appears to fit the data well, external validation would require high-quality day-of-death data. Currently, the VR data available through the WHO are limited to three time periods (day 0, days 1-6, and days 7-27) and therefore cannot be used to construct neonatal survival curves. Additionally, characteristics such as income and NMR level of countries that currently have high-quality VR are substantially

different from those that are being modelled, thus making a comparison with existing data difficult.

Another desirable improvement would be subnational estimates, particularly for large countries with decentralised systems and high variability such as India and Nigeria. Subnational estimates are not available even for many countries with adequate national VR data, but are important for priority setting and sharing lessons from within the same health system. For example, the risk of day 0 death per 1,000 live births in the United States ranged from 1.3 in Alaska to 4.8 in the District of Columbia from 2007-2010 [10]. The county-level differences were even wider, from a risk of 0.9 in Hidalgo County, Texas to 6.2 in Baltimore City, Maryland [10]. Additionally, a few studies have evaluated differences in the neonatal cause-of-death distribution by day, and found marked differences between not only the early and late neonatal periods, but also day 0 and later days [132, 133]. While a better understanding of this distribution by day is needed in order to improve care, the availability of relevant data from high-mortality settings are currently too insufficient to support such an analysis.

In the post-MDG era, we need to accelerate the impressive but unfinished recent progress in reducing preventable child deaths, especially for the immense burden of nearly 3 million neonatal deaths, which is highest on the first day and week of life. Effective and low-cost interventions exist but are still not reaching every woman and newborn, especially during the time around birth, the most vulnerable for both death and long-term disability [109]. A Lives Saved Tool analysis estimated that simple and cost-effective intervention (e.g. resuscitation devices, chlorhexidine cord cleansing, and injectable antibiotics) could save up to one million newborn lives each year [12]. A more comprehensive approach with full obstetric and improved newborn care, linked to community based programmes [134, 135], could prevent almost all of these deaths, as well as many of the estimated 1.2 million intrapartum stillbirths (which are rarely seen in rich countries) and 289,000 maternal deaths [111] each year [9, 136]. The Every Newborn Action Plan, endorsed at the 2014 World Health Assembly by over 190 countries, was an important step towards accelerating progress [137]. Subsequently, the Sustainable Development Goals set explicit targets that all countries should reach 12 or fewer neonatal deaths per 1,000 live births by 2030 [138]. Around the world, a marker of development is when a society no longer accepts that stillbirths and neonatal deaths are inevitable, that babies can be named at birth and counted in national data systems, and that a baby's birth day should not be his or her last day.

4 Cause-of-death estimation by neonatal period

In this chapter, I present our work on estimating neonatal causes of deaths in the early and late neonatal periods for 194 countries. Preliminary estimates were published in the WHO's World Health Statistics 2014 report [139], and final estimates were used in UNICEF's "A Promise Renewed Progress Report 2014" [140] and the 2014 Liu et al. *Lancet* article [100]. We published detailed methods and results in the *Bulletin of the World Health Organization* under the following citation: Oza S, Lawn J, Hogan D, Mathers C, Cousens S. Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000-2013. *Bulletin of the World Health Organization* 2015; 93:19-28 [141]. The text in this chapter is adapted from this journal article and its associated web appendix. The open-access article can be found at: https://www.who.int/bulletin/volumes/93/1/14-139790/en/

4.1 Introduction

Most of the estimated 2.76 million neonatal (first month of life) deaths in 2013 occurred from preventable causes [142]. While the global neonatal mortality rate (NMR) is decreasing, its rate of reduction has been substantially slower than the decreases in under-5 and maternal mortality [108, 143]. Neonatal deaths constituted 44% of all deaths in children under 5 years old in 2014 [142]. One hundred and thirty-three countries were unable to achieve the fourth Millennium Development Goal (MDG) target of a two-thirds reduction in child mortality (U5MR) between 1990 and 2015 [7], at least partly due to limited reductions in neonatal deaths. The Every Newborn Action Plan, launched in June 2014 [144], provided a stimulus to accelerate progress by implementing effective cause-specific interventions to rapidly reduce neonatal mortality. Subsequently, the Sustainable Development Goals (SDGs) included a target of 12 or fewer neonatal deaths per 1,000 live births for all countries by 2030 [138].

Understanding the neonatal cause-of-death (COD) distribution is important for identifying appropriate interventions and programme priorities. Ideally, such a distribution would be as local as possible, current, and distinguish programmatically relevant causes. Moreover, separate COD estimates are required for the early (days 0-6) and late (days 7-27) neonatal periods since both our understanding of pathology and empirical data suggest that the COD distribution differs substantially between these periods [145, 146]. Around three-quarters of neonatal deaths occur during the early period [12], and most interventions to prevent these deaths need to be delivered within a very short window of time.

For countries with high-quality vital registration (VR) data by cause and age at death, such data provide the information needed to determine policies and priorities. VR data quality is dependent on the completeness of reporting and the quality of COD coding [24, 66]. Unfortunately, high-quality VR data are available for only about one-third of countries [147], which account for only about 4% of neonatal deaths. Thus, statistical modelling remains necessary to estimate COD distributions in the majority of countries.

Systematic estimates of neonatal deaths classified into programmatically relevant cause categories were first published in 2005, for the year 2000 [13], by the Child Health Epidemiology Research Group (CHERG). These estimates were developed using data from high-quality VR systems and from research studies in high-mortality/low-resource settings in which high-quality VR data were lacking. Updated neonatal COD estimates using this approach were subsequently published for 2008 [14] and 2010 [15].

The current work goes beyond these previous exercises by estimating neonatal causes of death separately for the early and late neonatal periods, and by adding injuries as a distinct cause for low mortality countries. The separation of early and late neonatal deaths is an important advance which will ideally aid policy makers and programme managers. The input data have also been updated and modifications made to the modelling strategy, particularly for the split of neonatal infections between pneumonia and sepsis.

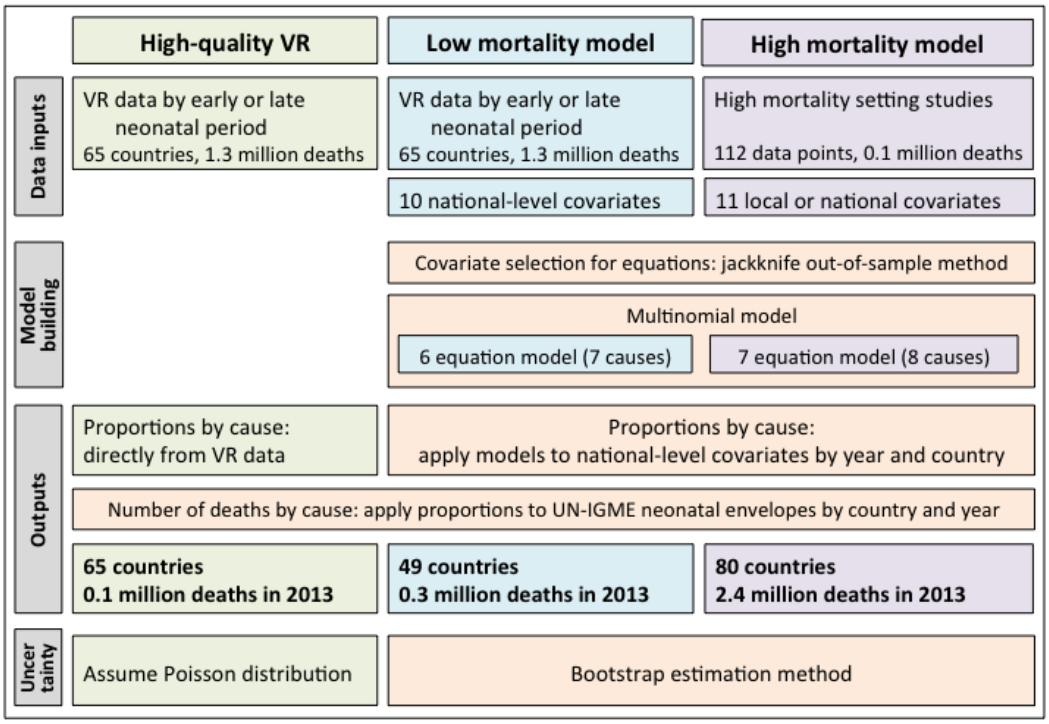
Here, we present global, regional, and national estimates (with uncertainty) of proportions, risks, and numbers of deaths for key programmatically relevant neonatal causes of death by the early and late neonatal periods for 194 countries for 2000-2013.

4.2 Methods

4.2.1 Overview of cause-of-death estimation

We divided 194 countries into three groups based on the quality of their VR data and their child mortality rates (Appendix B.1.1). Different methods were used to estimate the proportional COD distributions for countries in each group (Figure 4.1). For the 65 countries with high-quality VR data, the proportional COD distribution from 2000-2013 was obtained directly from the country's VR data. For the 49 countries without high-quality VR but with low child mortality, this distribution was estimated using a multi-cause model ("low mortality model") with input data from high-quality VR countries.

Figure 4.1: Strategy for cause-of-death estimation by neonatal period



Note: VR = vital registration

For the high-quality VR countries and low mortality model, we used seven cause categories: complications of preterm birth ("preterm"), intrapartum-related complications ("intrapartum"), congenital disorders, pneumonia, sepsis and other severe infections ("sepsis"), injuries, and other causes. For the 80 countries with inadequate VR and high child mortality, we used studies that identified neonatal COD in high mortality settings as input data in a separate multi-cause model ("high mortality model"). The eight cause categories for this model were preterm, intrapartum, congenital disorders, pneumonia, diarrhoea, neonatal tetanus, sepsis, and other causes. Table 4.1 includes the case definitions used for each of these causes.

Table 4.1: Case definitions for neonatal causes of death

	Used in VR and preferred in study data	Alternative definition accepted in study data
Preterm birth complications	○ Specific complications of preterm birth, such as surfactant deficiency (Respiratory Distress Syndrome), intraventricular haemorrhage, necrotizing enterocolitis ○ Prematurity (<34 weeks) at which level preterm complications occur for most babies ○ Neonatal death with birth weight < 2000 g with unknown gestational age	○ "Prematurity" ○ "Very low birth weight"
Intrapartum-related complications	○ Neonatal encephalopathy with criteria suggestive of intrapartum events ○ Early neonatal death in a full-term baby with no congenital malformations and a specific history of acute intrapartum insult or obstructed labour	○ "Birth asphyxia" with Apgar-based definition but excluding preterm infants ○ Fits and/or coma in the first two days of life in a baby born at baby ○ Acute intrapartum complications
Congenital disorders	○ Major or lethal congenital abnormalities ○ Specific abnormality listed (e.g. neural tube defect, cardiac defect)	○ Congenital abnormality or malformation
Sepsis	○ Sepsis/septicaemia, meningitis, or neonatal infection	○ Neonatal infection
Pneumonia	○ Pneumonia or acute respiratory tract infection	○ Pneumonia
Neonatal tetanus	○ Tetanus	○ Spasms and poor feeding after age of 3 days
Diarrhoea	○ Diarrhoea	---
Injuries	○ Injuries (in VR data only)	N/A
Other	Specific causes not included in the above-listed causes, including: ○ Neonatal jaundice ○ Haemorrhagic disease of the newborn ○ Term baby dying due to in-utero growth restriction ○ Injuries (only searched for in study data)	○ Author grouping of "other" causes (excluding unknown)
Notes: Table is slightly modified from Lawn [148], which adapted it from Wigglesworth [149, 150] and NICE [150]; ICD codes for VR COD data are included in Appendix B.2.1.		

4.2.2 Data inputs

Cause-of-death data from vital registration

For the 65 high-quality VR countries, we obtained publicly available VR COD data by the early and late neonatal periods from the World Health Organization (WHO) for years 2000 and later. We mapped the reported causes of death to our cause categories (Appendix B.2.1).

We then generated a proportional cause distribution by dividing the number of deaths attributed to each cause by the total deaths across the seven causes. To create a full time series, we imputed the cause-specific proportions for years with missing VR data. For years with missing data that were between years with existing data, we used linear interpolation to impute the missing proportions. For years that were before the earliest or after the latest available data year, we applied the proportions from the nearest year with available data to the missing data. The imputed proportions were only used as estimates for the high-quality VR countries; the low mortality model input dataset only included the non-imputed data. A list of the missing data years for the high-quality VR countries is included in Appendix B.2.2. Note that these data years were missing in the data received by the WHO from the given country, but the relevant data may not be missing within the country itself.

Cause-of-death data from high mortality setting studies

The high mortality model input data consisted of neonatal COD distribution data from studies in high mortality settings. We updated a previously developed database of neonatal COD studies [151] by conducting an extensive literature review for relevant research published from January 2011 to May 2013. We used the same criteria for selecting new studies to include in the input dataset as used in previous iterations of this work [151]. The criteria were as follows:

- Publication in 1980 or later
- Study set in one of nine (of a total f 14) subregions with no or few countries with >90% VR coverage
- Community-based study or hospital based in populations with over 90% hospital delivery and defined catchment population
- Case ascertainment: follow up of newly born infants from birth to at least 7 or 28 days
- Number of deaths with known cause ≥ 20
- Study duration ≥ 12 months
- Included four or more of the eight selected causes of neonatal death
- Deaths of unknown cause ≤ 25% of total deaths
- Cause attribution based on skilled clinical investigation, post-mortem, or verbal autopsy
- Case definitions specified and comparable with other studies

Our search strategy involved doing a literature review in ten databases for articles published between January 2011 and May 2013. Previous searches covered periods from 1980 to December 2010. We searched Pubmed, Embase, Web of Knowledge, Medline, Global Health, Popline, and region-specific indices (LILACS, Africa-Wide Information, Western Pacific Region, Eastern Mediterranean, IndMed). Full search terms are in Appendix B.2.3.

For each study, we extracted COD data and, when necessary, re-categorized the causes into our cause categories. While we included injuries as a separate category in the low mortality model, the study data lacked enough information on injuries for separate estimation in the high mortality model. We recorded deaths separately by early or late neonatal period whenever possible. The full list of studies is in Appendix B.2.4.

Covariate data

We chose for investigation covariates that we believed might partly predict variation in the COD distribution across countries. We expect that these covariates act either directly (e.g. increased skilled birth attendance may lead to decrease in intrapartum-related complications) or as proxies (e.g. gross national income [GNI] for system-level factors like healthcare availability). We could only use covariates for which national time series are publicly available. The full set of covariate options in the model are listed in Table 4.2.

Table 4.2: List of potential covariates in the low and high mortality cause-of-death models

Covariate[1]	Included in models?[2]	
	Low mortality	High mortality
Neonatal mortality rate (NMR)	Yes	Yes
Infant mortality rate (IMR)	Yes	Yes
Under-5 mortality rate (U5MR)	Yes	Yes
Low birth weight (LBW)	Yes	Yes
General fertility rate (GFR)	Yes	Yes
Antenatal care (ANC)	Yes	Yes
Female literacy rate (FLR)	Yes	Yes
Diphtheria/Pertussis/Tetanus vaccine (DPT)	Yes	Yes
Bacillus Calmette-Guerin vaccine (BCG)	No	Yes
Protected at birth (PAB) against neonatal tetanus	No	Yes
Skilled birth attendance (SBA)	No	Yes
Region (LAC, SSA, or SA)[3]	No	Yes
Neonatal period	No	Yes
Premature versus LBW distinction	No	Yes
Gross national income (GNI)	Yes	No
GINI coefficient	Yes	No

[1] See section 2.4 for descriptions and sources of covariates; [2] "Included in models" means the covariate was part of the covariate selection process for the model, and does not indicate that it was selected for the final model; [3] LAC = Latin America and the Caribbean, SSA = Sub-Saharan Africa, SA = South Asia (see Appendix A.2.1 for countries in each region)

Note that not all of these are included in the final models (see "Statistical modelling" section below for further details). The majority of covariates were included in both models. GNI and the GINI coefficient were excluded from the high mortality model because these covariates are

only available at the national level, whereas nearly all of the input data come from local studies. Covariates were excluded from the low mortality model based on lack of relevance (e.g. preterm vs LBW distinction; neonatal period) or due to lack of variability for prediction countries (e.g. SBA, PAB).

The covariates were used in two ways: 1) as inputs into the multinomial models, and 2) as predictor variables to which the final model coefficients were applied to estimate the national proportional COD distributions. We used national-level covariates as inputs to the low mortality model since the input COD data are national, and, whenever possible, local-level data extracted from the studies for the high mortality model. When local-level data were unavailable, we used subnational- or national-level covariate data instead. For the predictions, we used national-level covariate data, with the exception of India for which we used state-level data to produce state-level estimates. We applied the same rules for imputing missing covariate data as for the VR data. For prediction purposes, we restricted covariate values to the input data ranges, and performed a sensitivity analysis without this restriction.

4.2.3 Statistical modelling

All statistical analyses were done using Stata (version 12). For each model, a baseline cause was chosen; this cause was preterm for the low mortality model (the most common cause) and intrapartum for the high mortality model (reported in all studies). Our overall estimation process had two stages. First, we selected covariates. Then, we estimated the log of the ratio of each of the other causes to the baseline cause (the "log-cause ratio") as a function of the selected covariates using a multinomial logistic regression model. For both the low and high mortality models, we ran separate models for the early and late neonatal periods. Since not all studies in the high mortality model reported deaths by period, we included the studies reporting only overall neonatal deaths in both the early and late high mortality models, and included a binary covariate for period in these models.

Covariate selection for models

For each of the four models, we used a previously developed jackknife procedure [15] to select the set of covariates that minimized the jackknife out-of-sample prediction error for each log-cause ratio separately. The jackknife process involves conducting the analysis on n-1 observations, using the results to predict the value of interest for the left-out observation, and repeating this for all n observations. This allows a comparison between the out-of-sample

predictions and the observed data. For our covariate selection, we first determined if the relationship between each covariate and log-cause ratio in the input data was best represented by a linear, quadratic, or restricted cubic spline relationship. We did this by choosing the covariate relationship which yielded the smallest chi-squared goodness-of-fit statistic (sum of the squared differences between observed and expected deaths divided by expected deaths) for the given log-cause ratio with the jackknife process. We then selected the covariate with the best goodness of fit (i.e. smallest chi-squared statistic) as the first covariate in the model. Finally, we added one covariate at a time, retaining it in the model only if the chi-squared statistic decreased and cycling through all the remaining covariates again. In this way, we selected a set of covariates for each non-baseline cause in each model.

Multi-cause models

The multi-cause multinomial logistic regression models fit the data for all causes simultaneously. Each input observation received a weight inversely proportional to the square root of the total deaths contributed by that observation. This weighting is intermediate between giving equal weight to each death and equal weight to each study or country year in the input data. We made assumptions about the cause category into which deaths from an unreported cause would have been assigned and re-wrote the likelihood function accordingly. If preterm, congenital, or sepsis were unreported, we assumed deaths from these to be in the "other" category. If pneumonia, diarrhoea, or tetanus were unreported, these were assumed to have been in the "sepsis" category.

The coefficient values from the multinomial logistic regression models were applied to the country- and year-specific national level predictor covariates to estimate the proportional COD distribution for each modelled country from 2000-2013.

4.2.4 Estimation of cause-specific deaths and risks

We applied our estimated COD fractions to the neonatal deaths and live births estimated for each country from 2000-2013 by the United Nations Inter-agency Group for Child Mortality Estimation (UN-IGME) [142]. We split the overall neonatal deaths from UN-IGME into early and late neonatal deaths. For high-quality VR countries, we took this split directly from the data. For the other countries, we assumed that 74% of neonatal deaths occurred during the early period and 26% in the late period (based on work presented in chapter 3). We also performed a sensitivity analysis in which we assumed the early proportion was 65% or 85% instead of 74%.

To determine the number of deaths by cause, neonatal period, and year for each country, we applied the cause-specific proportions derived from the VR data (for high-quality VR countries) or the relevant models to the period-specific neonatal deaths for each country and year. We then divided these by the relevant country-specific live births to obtain the risk (per 1,000 live births) for each cause by neonatal period and year. We aggregated cause-specific proportions, risks, and numbers of deaths for regional and global estimates (see Appendix B.1.2 for regional categories).

4.2.5 Uncertainty estimation

For the modelled estimates, we generated uncertainty estimates by drawing 1,000 bootstrap samples with replacement from the input data and re-running the multi-cause models to produce new proportional cause distributions. We took the 2.5th and 97.5th percentile values for each cause as the uncertainty bounds. For the high-quality VR data, we developed uncertainty estimates by assuming a Poisson distribution for the number of deaths (i.e. the standard error equalled the square root of the reported number of deaths).

4.2.6 Methodological differences between this work and previous estimates

Appendix B.3 contains a detailed overview of how the COD estimation methods have changed since the first CHERG neonatal COD estimates were published in 2005 for the year 2000. Key methodological changes since the last round of neonatal cause-of-death estimates [15] are described below.

Early/late neonatal period estimates

We are now reporting the results by the early and late neonatal periods. As described above, we run 4 separate models (early and late separately for both the low and high mortality models).

Changes to country groupings

We have now included Kuwait, Macedonia, Montenegro, Republic of Korea, and Saint Lucia in the high-quality VR countries instead of the low mortality model. We made this change because VR time series data recently became available from the WHO for these countries. Additionally, due to improved VR data collection in South Africa, we included South Africa in the group of high-quality VR countries instead of the high mortality model.

Covariate selection

Previously [15], we allowed the relationship between each covariate and the log of the cause/baseline cause ratio to be described either linearly or quadratically. In this work, we also included the possibility of a restricted cubic spline relationship to include potentially more flexible non-linear relationships. We also included more covariates in the low mortality model than were previously used.

Additional cause categories within the multinomial

In the low mortality model, we added injuries as a separate cause, while in the previous work injuries were included in the "other" category. In the high mortality model, we included sepsis, pneumonia, and tetanus as separate cause categories. In the previous work, the high mortality model included a broader "infections" category that included sepsis and pneumonia, and these were then split in a separate regression using the results from the multinomial. Additionally, tetanus deaths were previously estimated using a single-cause model, but tetanus is now included in the multinomial itself.

India estimates

Similar to the 2010 estimates [15], we produced national estimates for India by aggregating state-level estimates. For the 2010 estimates, a separate multinomial was developed using India-only input data. This time, however, we estimated the state-level proportional cause distribution for each Indian state/year within the overall high mortality model. We did this because there were not enough Indian studies that reported sepsis and pneumonia separately in order to estimate these causes in an India-only multinomial model, and because the COD distribution between Indian studies and non-Indian studies appeared to be quite similar.

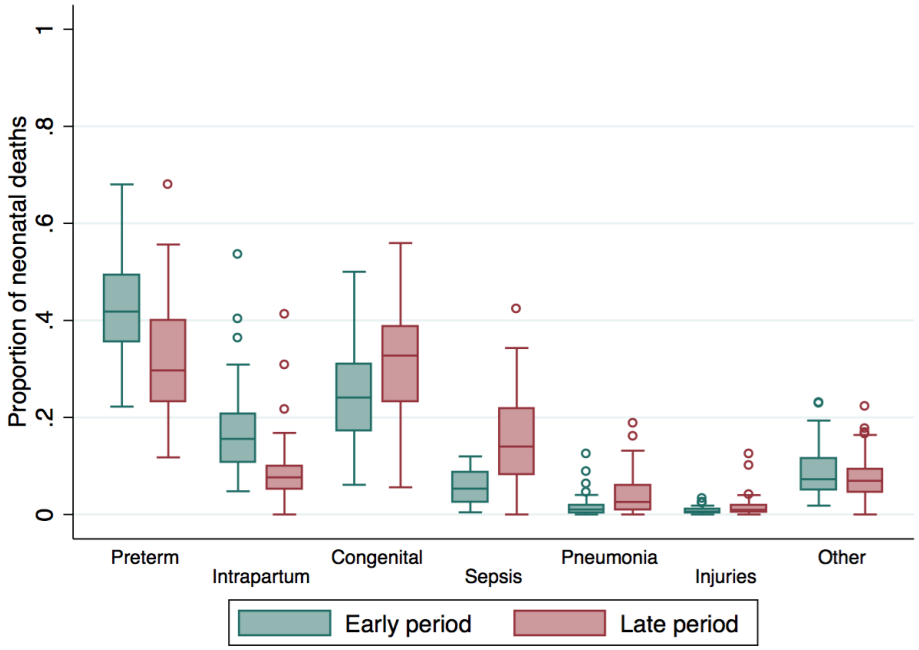
4.3 Main results

4.3.1 Description of data inputs

Cause-of-death data from vital registration

The high-quality VR input dataset, which was also used in the low mortality model, included 65 countries with 1,267,404 neonatal deaths and 665 country-years for each neonatal period from 2000-2013 (Figure 4.1). Of these deaths, 75.8% occurred in the early period. None of the seven causes we modelled for low mortality countries were missing in these data. Preterm and congenital were the most common causes during both neonatal periods (Figure 4.2).

Figure 4.2: Proportional COD distribution by period in the low mortality model input data



Cause-of-death data from high mortality setting studies

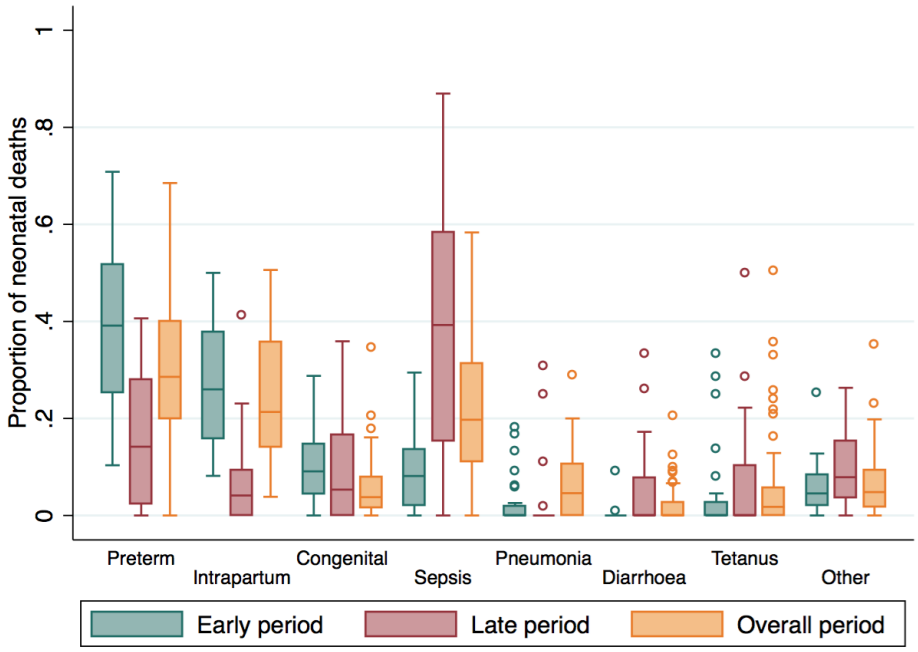
The high mortality model input dataset included 112 data points consisting of 98,222 deaths from 36 countries (Figure 4.1). This includes the addition of nearly 4,100 neonatal deaths from 15 new studies, representing 10 countries across five MDG regions. The overall dataset had 31 observations for the early period, 18 for the late period, and 63 for the overall neonatal period. Thirty-four observations had no causes missing, 37 had one cause missing, 13 had two causes missing, 23 had three causes missing, and 5 had four causes missing. No observations were missing data for the "intrapartum" or "other" cause categories; pneumonia and diarrhoea were the causes most commonly unreported (Table 4.3). All observations missing "sepsis" were also missing all other infection categories (i.e. diarrhoea, pneumonia, and tetanus). See section 4.2.3 for a description of the assumptions we made regarding which cause an unreported cause would be contained within.

Table 4.3: Number of observations with a given cause missing

Cause	# of observations[1]	Cause	# of observations[1]
Intrapartum	0 (0%)	**Pneumonia**	62 (55.4%)
Preterm	1 (0.9%)	**Diarrhoea**	42 (37.5%)
Congenital	13 (11.6%)	**Tetanus**	31 (27.7%)
Sepsis	3 (2.7%)	**Other**	0 (0%)
[1] out of 112 total observations			

Preterm and intrapartum were the most common causes of death during the early period, while infections (sepsis and pneumonia) dominated during the late period (Figure 4.3).

Figure 4.3: Proportional COD distribution by period in the high mortality model input data



Covariate data

Most covariate values were within the ranges of the input data (low mortality model, Table 4.4; high mortality model, Table 4.5). However, a few prediction covariates had values substantially outside the input data range, especially in the low mortality model. Notable examples were Monaco with an average GNI about 3.5 times larger than the maximum input data and Egypt with female literacy and ANC coverage 15-20 percentage points below the input ranges. An India-specific comparison is included in Appendix B.4.1.

Table 4.4: Comparison of input and prediction covariates in the low mortality model

	Input data			Prediction data[1]		
	Mean (SD[2])	**Median (IQR[3])**	**Range (min-max)**	**Mean (SD)**	**Median (IQR)**	**Range (min-max)**
NMR[4]	6.2 (4.3)	4.7 (2.8-8.8)	1.1-24.4	9.2 (4.5)	9.1 (5.4-12.4)	1.1-22.5
IMR	9.6 (7.2)	7.0 (4.2-13.6)	1.8-49.9	15.0 (7.7)	15.1 (8.4-20.6)	2.2-35.9
U5MR	11.4 (8.8)	8.2 (5.0-15.6)	2.3-76.9	17.8 (9.3)	17.6 (9.9-24.3)	3.0-44.8
LBW	7.3 (2.5)	7.0 (5.7-8.2)	3.8-23.2	8.7 (3.6)	8.4 (6.3-10.2)	**0.0**-22.3
GNI	21026 (13292)	18130 (11021- 30090)	1490-67970	20450 (38356)	7302 (4450-19600)	**1400-327346**
GFR	0.055 (0.018)	0.050 (0.040-0.065)	0.033-0.119	0.076 (0.029)	0.075 (0.050-0.096)	**0.032-0.171**
GINI	37.6 (9.7)	34.3 (30.9-45.3)	24.2-67.4	39.3 (7.7)	39.4 (32.6-42.1)	25.6-**69.2**
ANC	96.7 (4.3)	97.9 (96.3-99.4)	71.3-100	93.6 (7.2)	96.3 (90.3-98.9)	**52.9**-100.0
DPT	94.2 (5.6)	96.0 (93.0-98.0)	62.0-99.0	93.1 (8.6)	96.0 (92.0-98.0)	**41.0**-99.0
FLR	93.0 (4.7)	92.4 (90.7-97.0)	70.3-99.8	87.8 (9.0)	90.1 (81.9-93.5)	**50.6**-99.7

[1] bolded values are those outside the input data range; [2] SD = standard deviation; [3] IQR = interquartile range; [4] see Table 4.2 for covariate acronym definitions

Table 4.5: Comparison of input and prediction covariates in the high mortality model

	Input data			Prediction data[1]		
	Mean (SD[2])	**Median (IQR[3])**	**Range (min-max)**	**Mean (SD)**	**Median (IQR)**	**Range (min-max)**
NMR[4]	33.0 (15.9)	30.2 (18.8-47.2)	10.5-70.1	30.2 (10.2)	29.7 (21.8-37.3)	**8.8**-55.1
IMR	62.1 (30.0)	58.7 (35.7-81.6)	14.7-142.0	60.3 (24.9)	57.2 (40.0-76.5)	**14.4**-141.3
U5MR	88.6 (45.8)	89.1 (52.9-125.4)	17.1-227.0	89.2 (45.1)	83.0 (52.0-116.7)	**16.3**-**231.5**
LBW	18.9 (11.3)	15.9 (10.6-27.6)	2.5-50.0	14.2 (6.7)	12.8 (10.2-16.7)	4.2-35.6
GFR	0.127 (0.043)	0.118 (0.092-0.158)	0.057-0.235	0.140 (0.046)	0.145 (0.101-0.175)	**0.054**-**0.246**
ANC	67.5 (26.5)	73.9 (50.0-92.0)	5.0-98.3	78.6 (18.5)	84.1 (71.0-92.7)	16.1-**100.0**
DPT	67.9 (24.6)	73.5 (60.5-83.5)	0.0-99.0	76.3 (18.5)	80.0 (66.0-91.0)	3.0-99.0
BCG	78.8 (23.9)	87.0 (73.0-93.0)	0.0-100.0	85.1 (14.2)	90.0 (78.0-96.0)	24.0-99.0
PAB	63.3 (25.0)	68.0 (51.4-83.5)	0.0-98.5	76.0 (13.1)	79.7 (68.2-85.0)	24.0-97.0
FLR	51.9 (24.7)	48.7 (34.6-77.3)	4.0-94.0	62.2 (23.6)	61.8 (43.9-83.6)	9.4-**100.0**
SBA	48.7 (33.1)	45.3 (18.9-83.7)	0.0-100.0	60.1 (25.3)	60.2 (42.0-82.0)	5.6-100.0
Region	East Asia and Pacific: n = 4; Europe and Central Asia: n = 5; Latin America/Caribbean: n = 16; Middle East and North Africa: n = 8; South Asia: n = 52; Sub-Saharan Africa: n = 26 High income: n = 1			East Asia and Pacific: n = 14; Europe and Central Asia: n = 6; Latin America/Caribbean: n = 4; Middle East and North Africa: n = 6; South Asia: n = 5 Sub-Saharan Africa: n = 44; High income: n = 0		

[1] bolded values are those outside the input data range, results do not include Indian states (see Appendix B.4.1 for comparisons of Indian state data); [2] SD = standard deviation; [3] IQR = interquartile range; [4] see Table 4.2 for covariate acronym definitions

4.3.2 Statistical modelling

Covariate selection for models

Table 4.6 lists the covariates that were selected for each of the four models, as well as the performance of each equation in reducing the out-of-sample residuals. The latter was based on calculating the relative difference of the equation's chi-squared statistic with that of the null model with no covariates (herein "% reduction from null"). The model equations varied substantially in their performance, from 0% reduction in the chi-squared statistic (injuries, low mortality model, late period) to 87% reduction (diarrhoea, high mortality model, early period) (Table 4.6). Overall, equations in the high mortality model appeared to have better performance, with an average of 50% reduction from null the equations compared to an

average of 30% in the low mortality model equations. The poorer performance for some causes may be due to a number of factors, including the limited range of covariates available for inclusion, inaccurate measurement of included covariates, or the possibility that there is no pattern that can be predicted based on the input data.

Table 4.6: Selected covariates in cause equations of the low and high mortality models and % reduction from null

	Early neonatal period		Late neonatal period	
	Selected covariates[1]	% red.[2]	Selected covariates	% red.
Low mortality model[3]				
Intrapartum	**L:** FLR	16%	**S:** FLR, DPT	21%
Congenital	**L:** GINI, DPT, FLR **S:** IMR, U5MR, LBW	62%	**L:** NMR, DPT	10%
Sepsis	**L:** GNI, GINI, ANC; **S:** IMR, DPT	66%	**L:** FLR, GINI; **Q:** IMR; **S:** DPT	67%
Pneumonia	**L:** GNI	25%	**L:** GNI; **S:** ANC	40%
Injuries	**Q:** GFR	17%	none	0%
Other	**Q:** GFR	19%	**L:** LBW; **Q:** DPT; **S:**NMR	16%
High mortality model[4]				
Preterm	**L:** BCG, PAB, SBA, DPT **S:** LBW, GFR	47%	**S:** LBW, PAB, GFR **B:** SSA	61%
Congenital	**L:** LBW; **Q**: NMR, U5MR; **S:** BCG; **B**: per	77%	**Q:** SBA, U5MR **B:** per, SSA	71%
Sepsis	**L:** LBW; **Q:** BCG; **B:** per, SA	81%	**Q:** PAB; **S:** LBW; **B:** per	13%
Pneumonia	**L:** U5MR, LBW; **B:** per	16%	**Q:** PAB	23%
Diarrhoea	**L:** DPT, GFR; **Q:** NMR **B:** per, SA, SSA	87%	**L:** DPT, BCG, GFR, FLR **S:** LBW; **B:** LAC	45%
Tetanus	**L:** PAB, ANC, NMR; **B:** per	86%	**L:** NMR, IMR, U5MR, PAB **B:** period	44%
Other	**S:** GFR; **B:** SSA	6%	**S:** GFR; **B:** per, SSA	49%

[1] L = linear; Q = quadratic; S = restricted cubic spline; B = binary; per = neonatal period; reg = region; see Table 4.2 for covariate acronym definitions; [2] % red. = % reduction from null; [3] preterm as baseline cause; [4] intrapartum as baseline cause

Multi-cause models

Model regression coefficients are in Appendix B.4.2.

4.3.3 Overall cause-specific deaths and risks

Globally, the leading causes of neonatal death in 2013 were estimated to be preterm (35.7%), intrapartum (23.4%), and sepsis (15.6%), accounting for 2.1 (uncertainty range [UR]: 1.4-2.8) of the estimated 2.8 million neonatal deaths (Table 4.7). The proportional cause distribution varied by both neonatal period and NMR level. In the early period, preterm (40.8%) and intrapartum (27.0%) accounted for the majority of deaths while in the late neonatal period

nearly half of all deaths occurred from infectious causes (sepsis, pneumonia, tetanus, and diarrhoea) (Table 4.7). The proportion of deaths from congenital disorders was relatively stable across the periods. A comparison of the estimated COD proportions in this round of estimation versus previous rounds is presented in Appendix B.4.3.

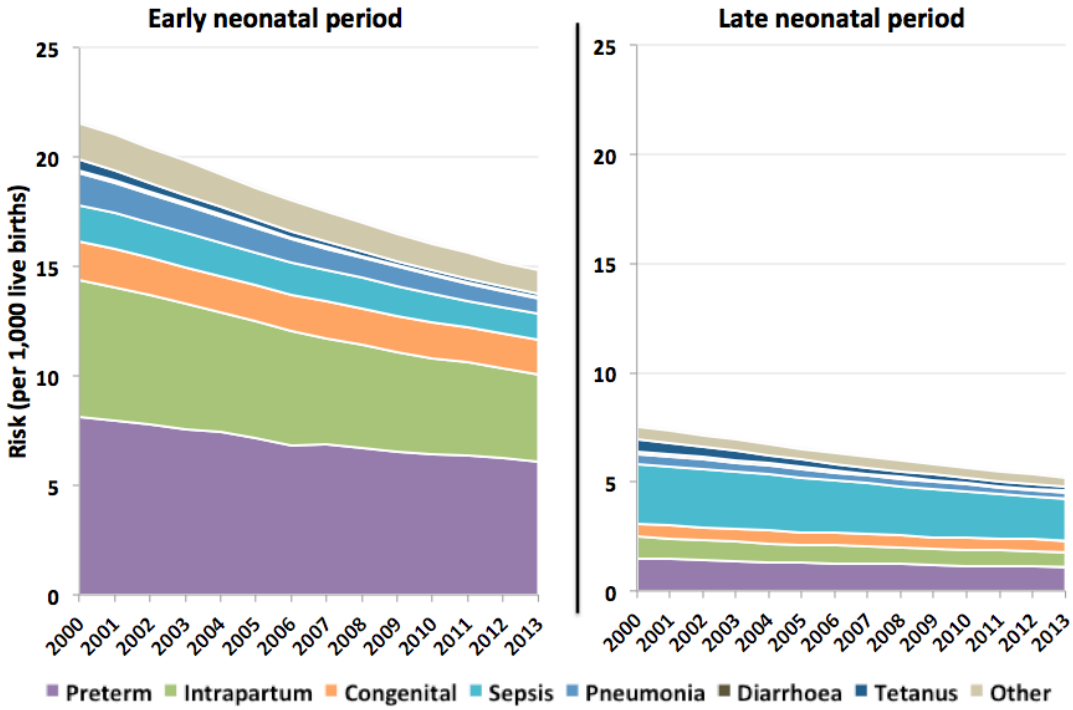
Table 4.7: Global cause-specific proportions, risks, and numbers of neonatal deaths (with uncertainty) in 2013

	Early period		Late period		Overall period		
	%	# of deaths in 1000s[1]	%	# of deaths in 1000s	%	# of deaths in 1000s	Risk[2]
Preterm	40.8	834.8 (608.1-1083.5)	21.2	152.1 (91.0-229.0)	35.7	986.9 (699.1-1312.5)	7.2
Intrapartum	27.0	552.7 (407.6-711.4)	12.9	92.1 (54.8-133.4)	23.4	644.8 (462.4-844.7)	4.7
Congenital	10.6	217.0 (140.9-325.9)	10.2	72.8 (42.5-124.5)	10.5	289.8 (183.3-450.4)	2.1
Sepsis	8.0	163.7 (62.4-271.6)	37.2	266.7 (156.5-393.2)	15.6	430.4 (218.9-664.8)	3.1
Pneumonia	4.8	98.9 (48.8-200.3)	5.2	37.6 (21.5-58.7)	4.9	136.4 (70.3-259.0)	1.0
Diarrhoea[3]	0.3	6.7 (0-57.4)	1.4	10.0 (3.2-25.6)	0.6	16.6 (3.2-83.0)	0.1
Tetanus[3]	1.0	21.1 (7.4-53.2)	3.8	27.1 (8.1-67.2)	1.7	48.2 (15.5-120.4)	0.3
Other[4]	7.3	149.9 (72.7-250.3)	8.1	57.9 (26.3-117.2)	7.5	207.8 (99.0-367.4)	1.5

[1] uncertainty ranges are in parentheses; [2] risk is per 1,000 live births; [3] estimated only for the 80 high mortality model countries; [4] injuries are included within the "other" category

While the absolute risks of death due to intrapartum and preterm were predicted by the model to have fallen in the early period, they decreased less in the late period. Risk of tetanus declined in both periods (Figure 4.4).

Figure 4.4: Global cause-specific risks of death from 2000-2013 for the early and late neonatal periods



4.3.4 Cause-specific deaths and risks by neonatal mortality rate, income, and region

Higher NMR (Figure 4.5a) and lower income (Figure 4.5b) were associated with higher proportions of deaths attributable to intrapartum and infectious causes. Globally, risks for all causes decreased as the global NMR decreased over time. The risk of death from each cause was substantially higher in higher mortality settings, even for causes that dominated proportionally in low mortality settings (e.g. preterm and congenital disorders). The risks of death due to preterm, intrapartum, and sepsis were 10, 36, and 34 times greater in settings with NMR≥30 compared to NMR≤5 (Figure 4.5a). In every MDG region, preterm was the leading cause of neonatal death, with the highest risks in Southern Asia (11.9 per 1,000 live births) and Sub-Saharan Africa (9.5 per 1,000 live births) (Figure 4.6).

Figure 4.5: Cause-specific risk of death by a) neonatal mortality rate and b) income groupings

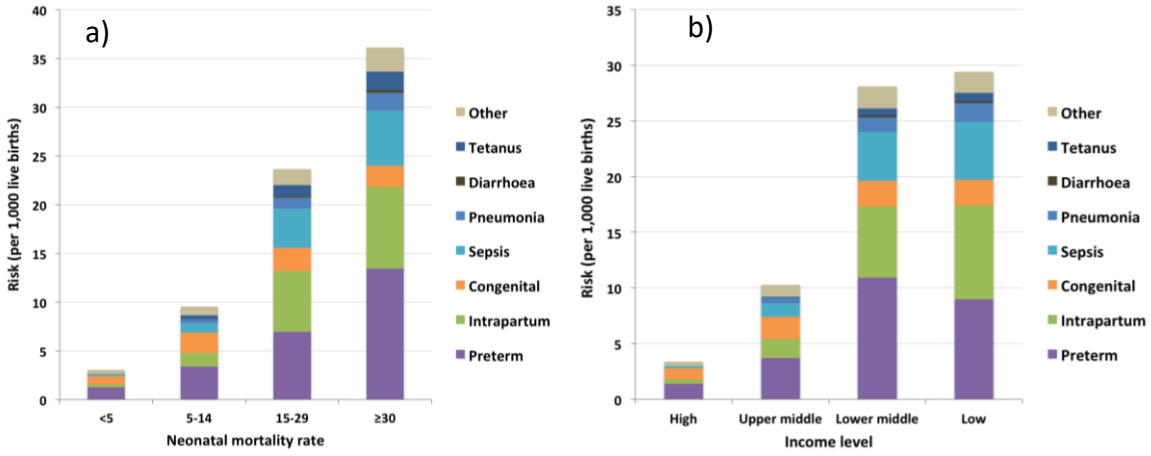
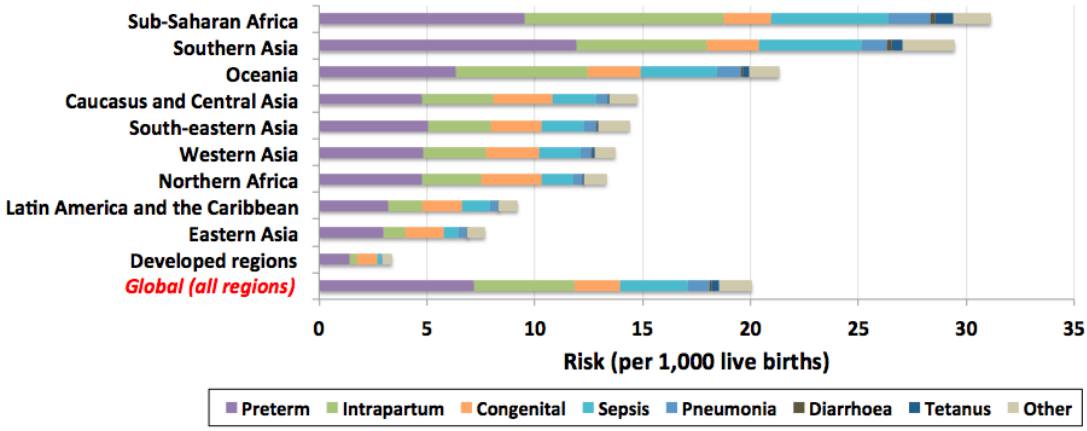


Figure 4.6: Cause-specific risk of neonatal death by MDG region in 2013



4.3.5 Cause-specific deaths and risks by estimation method

Table 4.8 includes the cause-specific proportions, risks, and numbers of deaths by estimation method (i.e. high-quality VR data, low mortality model, and high mortality model). As expected, the proportional COD distribution for the low and high morality models mimics the differences seen across low and high NMR areas (e.g. higher intrapartum and infection in the high mortality model; higher congenital and other in the low mortality model and high-quality VR data). These differences are unsurprising because the estimation groups are determined by mortality level and availability of high-quality data. The risk of death from most causes, including preterm and intrapartum, decreased at a slower rate for high compared to low mortality countries (Figures 4.7a-c).

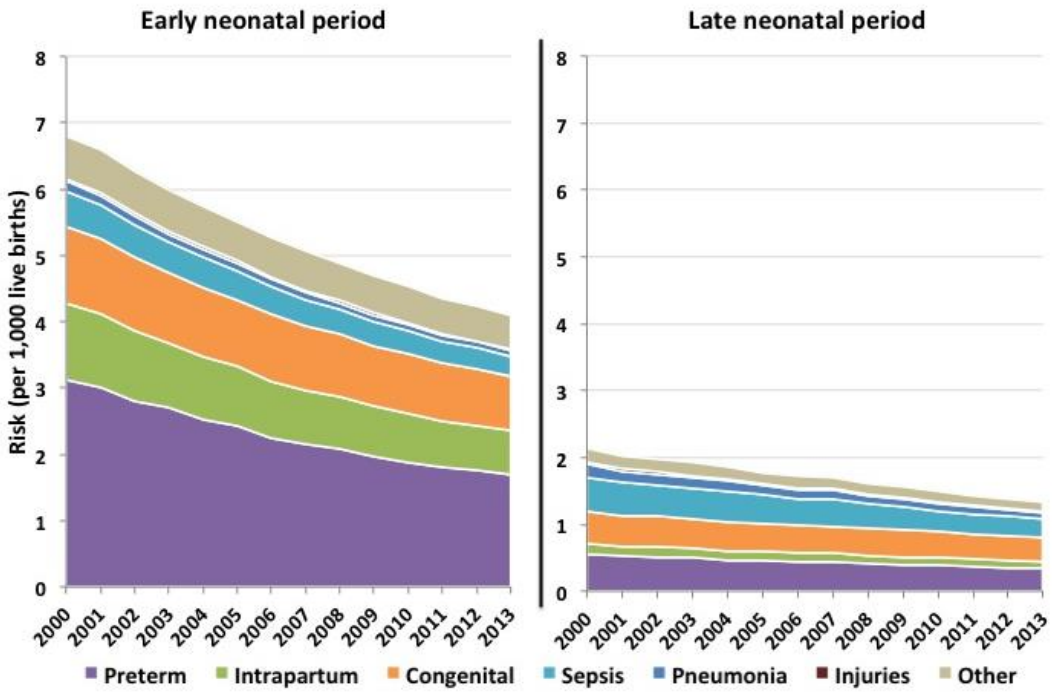
Table 4.8: Cause-specific proportions, risks, and numbers of deaths (with uncertainty) in 2013 by estimation method

	Early period			Late period		
	%	# of deaths in 1000s[1]	Risk[2]	%	# of deaths in 1000s	Risk
High-quality VR countries						
Preterm	41.5	35.9 (34.0-37.7)	1.7	25.9	7.3 (6.4-8.2)	0.3
Intrapartum	16.1	13.9 (12.8-15.0)	0.7	7.5	2.1 (1.7-2.6)	0.1
Congenital	20.2	17.5 (16.1-18.8)	0.8	26.4	7.5 (6.6-8.4)	0.4
Sepsis	7.3	6.3 (5.6-7.0)	0.3	21.3	6.0 (5.4-6.7)	0.3
Pneumonia	2.2	1.9 (1.5-2.2)	0.1	6.8	1.9 (1.6-2.3)	0.1
Injuries	0.5	0.4 (0.3-0.6)	<0.05	1.6	0.4 (0.3-0.6)	<0.05
Other	12.3	10.6 (9.7-11.5)	0.5	10.5	3.0 (2.5-3.5)	0.1
Total	*100*	*86.4*	*4.1*	*100*	*28.2*	*1.3*
Low mortality model countries						
Preterm	43.0	80.8 (68.8-91.8)	2.6	27.9	18.4 (16.0-22.2)	0.6
Intrapartum	15.8	29.7 (22.3-36.5)	0.9	9.0	5.9 (3.4-8.1)	0.2
Congenital	21.8	41.0 (30.3-57.1)	1.3	27.6	18.2 (15.9-21.6)	0.6
Sepsis	5.8	11.0 (7.0-14.5)	0.4	17.6	11.6 (7.6-14.8)	0.4
Pneumonia	2.9	5.4 (3.5-8.3)	0.2	9.9	6.5 (4.1-9.7)	0.2
Injuries	0.7	1.2 (0.9-1.8)	<0.05	1.5	1.0 (0.8-1.3)	<0.05
Other	10.1	19.0 (13.8-23.3)	0.6	6.7	4.4 (2.6-7.4)	0.1
Total	*100*	*188.1*	*6.0*	*100*	*66.1*	*2.1*
High mortality model countries						
Preterm	40.6	718.1 (505.3-954.0)	8.4	20.3	126.4 (68.6-198.6)	1.5
Intrapartum	28.8	509.2 (372.6-659.9)	6.0	13.5	84.0 (49.7-122.7)	1.0
Congenital	9.0	158.6 (94.5-250.0)	1.9	7.6	47.1 (20.0-94.6)	0.6
Sepsis	8.3	146.5 (49.7-250.2)	1.7	40.0	249.0 (143.6-371.7)	2.9
Pneumonia	5.2	91.5 (43.8-189.7)	1.1	4.7	29.2 (15.8-46.7)	0.3
Tetanus	1.2	21.1 (7.4-53.2)	0.2	4.4	27.1 (8.1-67.2)	0.3
Diarrhoea	0.4	6.7 (0-57.4)	0.1	1.6	10.0 (3.2-25.6)	0.1
Other	6.7	118.6 (48.1-213.1)	1.4	7.9	49.1 (20.2-104.5)	0.6
Total	*100*	*1770.2*	*20.8*	*100*	*622.0*	*7.3*

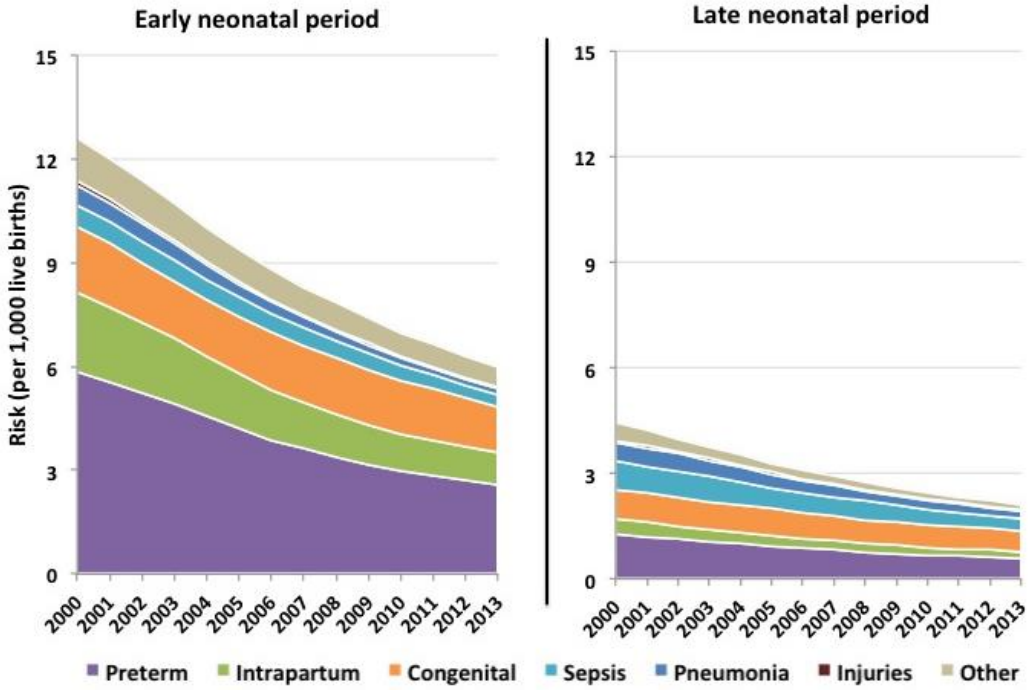
[1] uncertainty ranges are in parentheses; [2] risk is per 1,000 live births

Figure 4.7: Cause-specific risk from 2000-2013 by neonatal period and estimation method a) high-quality VR, b) low mortality model, and c) high mortality model. Note: the y-axes (risk) are different on the 3 graphs due to different mortality risks between the estimation categories.

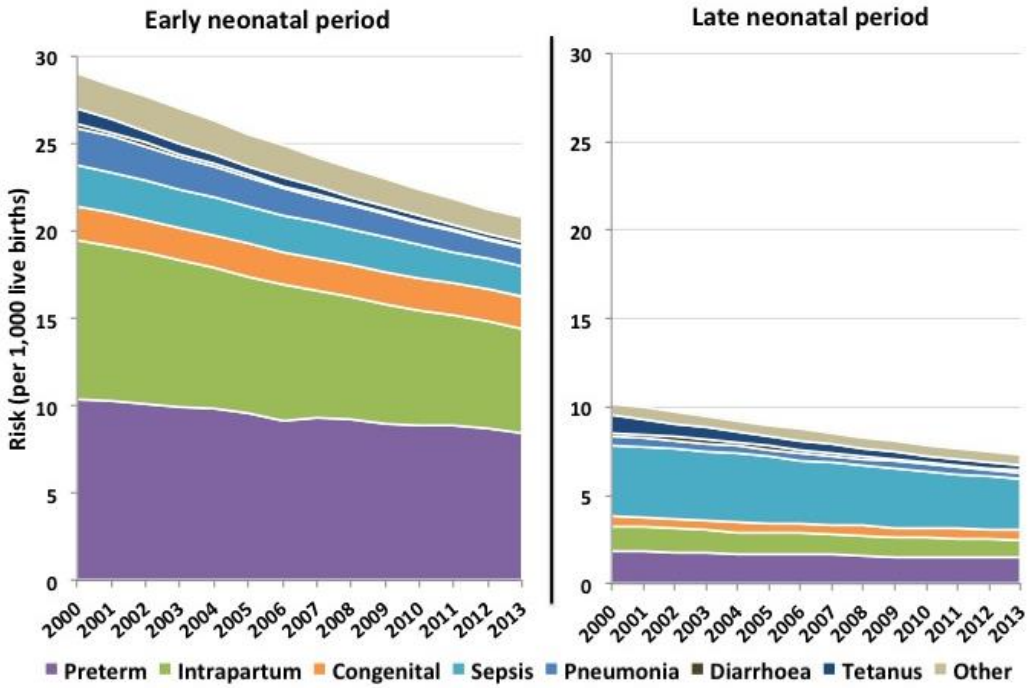
a)



b)



c)



4.4 Sensitivity analyses

In this section, I describe sensitivity and validation exercises we performed for this work.

4.4.1 Analysis assuming 65% and 85% for the proportion of early neonatal deaths

Based on previous work [12], we assumed that 74% and 26% of neonatal deaths occurred in the early and late periods, respectively, for countries without adequate VR data. The three-quarters/one-quarter split is quite consistent across countries, including ones with widely varying contexts [12]. However, to test how this affected our results, we estimated the global cause-of-death distribution for the overall neonatal period if the proportion of early neonatal deaths was 65% or 85% instead of 74%. The results are shown in Table 4.9. There was little difference in the estimates when the proportion of early deaths was assumed to be 65% or 85% instead of 74% for modelled countries, and these changes made no difference to the ranking of causes.

Table 4.9: Global cause-specific proportions and numbers of neonatal deaths in 2013 (with uncertainty) assuming different proportions of deaths in the early neonatal period

	Assuming 74%[1] deaths in early neonatal period (current method)		Assuming 65%[1] deaths in early neonatal period*		Assuming 85%[1] deaths in early neonatal period*	
	%	# of deaths in 1000s[2]	%	# of deaths in 1000s	%	# of deaths in 1000s
Preterm	35.7	986.9 (699.1-1312.5)	34.0	939.8	37.8	1044.4
Intrapartum	23.4	644.8 (462.4-844.7)	22.1	610.4	24.9	686.9
Congenital	10.5	289.8 (183.3-450.4)	10.4	288.1	10.6	291.8
Sepsis	15.6	430.4 (218.9-664.8)	18.2	501.5	12.4	343.6
Pneumonia	4.9	136.4 (70.3-259.0)	5.0	137.0	4.9	135.8
Diarrhoea[3]	0.6	16.6 (3.2-83.0)	0.7	19.3	0.5	13.4
Tetanus[3]	1.7	48.2 (15.5-120.4)	2.0	55.0	1.4	39.8
Other[4]	7.5	207.8 (99.0-367.4)	7.6	209.8	7.4	205.4

[1] this split is only applied to countries for which modelled estimates are needed – the split for countries with high-quality VR data is taken from the data as is; [2] uncertainty ranges are in parentheses; ; [3] estimated only for the 80 high mortality model countries; [4] injuries are included within the "other" category

4.4.2 Uncapped prediction covariate values

As noted in section 4.2.2, we capped the predication covariate values to the minimum and maximum of the input covariate values in both the low and high mortality models to avoid predicting on covariates outside of the input data range. Capping the prediction data will tend to "shrink" the predicted cause distribution towards that seen in the input data. Here, I present the results of a sensitivity analysis in which we ran the low and high mortality models without the caps on the covariate values (Table 4.10).

Table 4.10: Estimates with capped versus uncapped prediction covariate values for 2013

	Early period			Late period		
	% of deaths (capped)[1]	% of deaths (uncapped)[2]	Absolute difference	% of deaths (capped)	% of deaths (uncapped)	Absolute difference
Low mortality model						
Preterm	43.0	44.2	+1.2	27.9	27.7	-0.2
Intrapartum	15.8	16.5	+0.7	9.0	9.2	+0.2
Congenital	21.8	19.1	-2.7	27.6	27.4	-0.2
Sepsis	5.8	6.1	+0.3	17.6	17.9	+0.3
Pneumonia	2.9	3.0	+0.1	9.9	9.8	-0.1
Injuries	0.7	0.7	0	1.5	1.5	0
Other	10.1	10.4	+0.3	6.7	6.6	-0.1
High mortality model						
Preterm	40.6	40.6	0	20.3	20.3	0
Intrapartum	28.8	28.7	-0.1	13.5	13.5	0
Congenital	9.0	9.0	0	7.6	7.6	0
Sepsis	8.3	8.3	0	40.0	40.1	+0.1
Pneumonia	5.2	5.2	0	4.7	4.7	0
Tetanus	1.2	1.2	0	4.4	4.3	-0.1
Diarrhoea	0.4	0.4	0	1.6	1.6	0
Other	6.7	6.7	0	7.9	7.9	0

[1] "capped" analysis is the main analysis; [2] "uncapped" is the sensitivity analysis.

In the low mortality model, excluding countries with few deaths (<50), the key differences (>5 percentage point difference) in countries were:

- Egypt (both periods, 2000-2006) – intrapartum was 5-9 percentage points higher with the uncapped covariates
- Syria (both periods, 2012-2013) – sepsis was 9-11 percentage points higher and congenital was 6 percentage points lower with the uncapped covariates
- Honduras (early period, 2000) – injuries was 8 percentage points higher with the uncapped covariates
- Jordan (early period, 2000) – injuries was 5 percentage points higher with the uncapped covariates

In the high mortality model, both Niger (2000, other) and Kazakhstan (2013, congenital) were 2 percentage points higher in the early period with uncapped covariates. No other countries had more than a one percentage point difference between the sensitivity versus main analysis.

Overall, the decision to cap or not appears to have had minimal influence on the results.

4.4.3 Comparison of different neonatal cause-of-death estimates for China

We used the low mortality model to estimate the proportional cause distribution in China due to its relatively low NMR (18.8 in 2000; 7.7 in 2013), obtaining around 14.4% of neonatal deaths attributable to intrapartum-related complications (Appendix B.4.4). While we chose to retain the low mortality model results for China, we have included two different estimates in Appendix B.4.4 for comparison. The first are the WHO estimates for China (Table B.8a; Appendix B.4.4) [152], based on single-cause models for the overall neonatal period, and the second are estimates for China using the high mortality model instead of low mortality model in our analysis (Tables B.8b and B.8c; Appendix B.4.4).

4.5 Discussion

We developed systematic, nationally comparable estimates of programmatically relevant neonatal causes of death. These estimates go beyond CHERG's previous estimates in providing separate decompositions for the early and late neonatal periods. The proportional neonatal cause distribution varies with a number of factors, including the early versus late neonatal periods, NMR level, and over time. To reduce neonatal deaths, such variations must be understood, and this knowledge must be incorporated into decisions regarding the selection of appropriate interventions. With the launch of the Sustainable Development Goals for the period 2015 to 2030, this is a particularly opportune time for affecting such change within countries.

The three leading categories of causes of neonatal death (preterm, intrapartum, and infections) are the same for the early and late neonatal periods, but their distribution is substantially different between the two periods. Globally, in the early period, preterm and intrapartum account for nearly 68% of deaths while infections (pneumonia, sepsis, tetanus, and diarrhoea) account for around 14%. In the late period, around 34% of deaths are due to intrapartum or preterm while roughly 48% are from infections. Intrapartum-related complications are expected to occur in the minutes or hours after birth, and hence cause more deaths during the early period.

Even within each neonatal period, considerable differences exist in the proportional cause distribution by NMR level. Generally, NMR is closely linked to the level of care available to neonates. Settings with very low NMR tend to have readily available intensive care for newborns, while areas with high NMRs often lack even simple interventions like clean delivery

kits and resuscitation equipment. In this analysis, low mortality countries had higher proportions of deaths from congenital disorders and lower proportions from intrapartum and infections, while the opposite was true in high mortality settings. While infection deaths accounted for 51% of deaths in the late period in high mortality countries, they caused less than 30% of late neonatal deaths in low mortality countries. This is likely because of better access to and availability of treatment and infection control in low mortality countries.

We used our model to predict trends in individual causes of death. Our model predicts that deaths due to intrapartum-related complications had the largest absolute risk reduction between 2000 and 2013, from 7.2 (UR: 4.8-9.5) to 4.7 (UR: 3.4-6.1) per 1,000 live births, possibly because of increased coverage of skilled obstetric care. The largest relative decrease in risk was predicted for neonatal tetanus, which dropped by 73% (from 1.1 [UR: 0.3-2.8] to 0.3 [UR: 0.1-0.9]) between 2000 and 2013. This may be due to increases in clean deliveries, facility birth, cord care, and tetanus toxoid vaccination (a predictor in the model; PAB), as well as contextual changes in maternal education and social norms. Additionally, a few countries in the low mortality model eliminated neonatal tetanus after 2000. Since tetanus is not estimated in the low mortality model, we may thus be underestimating the relative decline in risk.

The smallest predicted relative decrease in risk was for congenital (13% drop; from 2.4 [UR: 1.4-4.1] to 2.1 [UR: 1.3-3.3]). While the predicted risk of death due to preterm complications fell by 2.4 per 1,000 live births between 2000 and 2013 (from 9.6 [UR: 6.7-12.8] to 7.2 [UR: 5.1-9.5]), this represented the second smallest relative decrease (25%), despite the existence of simple and cost-effective interventions such Kangaroo Mother Care [12, 153]. In addition to prematurity, there is also evidence that babies that are small-for-gestational age are at higher risk of death [154].

Broadly, our results are similar to those from 2010 [15], with the exception that we estimated substantially fewer pneumonia deaths than before. This is likely due to improvements to the estimation approach, namely the inclusion of additional studies that split pneumonia from sepsis and the inclusion of pneumonia directly in the multinomial model.

We used the low mortality model to estimate the proportional cause distribution in China due to its low NMR (18.8 in 2000; 7.7 in 2013), obtaining around 14.4% of neonatal deaths attributable to intrapartum-related complications. Others have also estimated the neonatal COD distribution in China and their results differ somewhat from ours, most notably with higher

proportions of intrapartum-related deaths [155, 156]. Our approach differs in three important ways: 1) we used multi-cause instead of single-cause models, 2) we estimated results for the early and late neonatal periods separately, and 3) we included input data from outside of China. We believe the first two are advantages in our work, while using China-only data is an advantage of the other models. The global COD distribution changes little when we apply the WHO COD proportions [152] for China instead of ours, with the biggest differences in 2013 being intrapartum-related and other deaths increasing by about 2 percentage points.

Although the quantity and quality of data has improved in recent years, enormous data gaps still exist. We now have nearly 100,000 deaths and over 90 studies in the high mortality model compared to <14,000 deaths and <60 studies when the estimates were produced in 2005 [13]. However, while we used the high mortality model for 80 countries, the inputs only included data from 36 countries. Many of these studies were relatively small, and few were nationally representative. We could include data from only 13 Sub-Saharan African countries, the region with the highest risk of neonatal death. Excluding a large South African dataset, the studies from these countries contribute only 4,000 deaths to our database. It is an unfortunate reality that we know the least about the areas with the highest burdens.

As with all such modelling exercises, our estimates should be viewed as an interim measure to help policymakers, particularly in settings with little or no data currently. It is important to distinguish *estimates* from *data* and to recognize that not all estimates are "equal". We used UN-IGME estimates of the NMR and total number of neonatal deaths in each country. The UN-IGME estimates of all-cause mortality in each country are derived from data for that country and therefore it can be argued that they "track" mortality in each country, though not in real time. For most countries, our cause-specific estimates are not based on data from that country, but from a model bringing together data from many countries. The model then predicts the cause-of-death distribution, and changes in the cause-of-death distribution, in individual countries based on covariate values for the individual country. Some countries contribute little or no input data to the modelling process. For example, only 24 deaths in our input data come from Nigeria, one of the most populous countries. Our estimates should not therefore be interpreted as "tracking" changes in causes of death for the majority of countries, but rather as predictions of what might be occurring in countries. To track changes in burden due to specific causes of death requires each country to collect representative and consistent data on causes of death on a continuing basis. We emphasise that our estimates are not a panacea for actual data collection and should not be an excuse for a failure to collect data.

Fortunately, rapid but sensible improvements in data collection are possible. Recent examples of countries like South Africa improving their data systems to the point where their VR data is considered high quality are encouraging. As data systems improve, regression-based models should be replaced by reliable local COD data.

The validity of our estimates relies on the quality of the input/prediction data, and our modelling techniques. Quality is of particular concern for verbal autopsy studies, in which the reported cause distributions depend heavily on the case definitions and causal hierarchies used to attribute deaths [66]. Accurate cause attribution using VA will always be problematic for causes that are difficult to distinguish, such as sepsis and pneumonia, or difficult to identify, such as internal congenital abnormalities. The potential lack of comparability between different VA studies can affect the ability of our model to predict variation between settings. By following standardized methods when conducting VA studies, some of these problems can be partially alleviated [68]. Additionally, regression-based models inherently depend on the relationship between outcome variables and covariates, which should ideally come from the same population and time period. While we sought to include as much local covariate information as possible for the input studies, 52% of the total covariate information came from national data instead of from local/regional data. Finally, when re-categorizing reported VA causes of death (Table 4.1), we had to make choices, for example placing deaths reported as being due to "very low birth weight" into the preterm complication category. This may introduce a degree of misclassification as some "very low birth weight" deaths may be attributable to congenital abnormalities. We made those choices that we believed would introduce the least misclassification, but until VA methods improve, this will continue to be a challenge. Similar issues exist in ICD coding, but are more common in VA studies because of the limited and lower quality information collected.

Even high-quality VR data can have problems. Unfortunately, ICD-10 codes are not ideal for neonatal causes, particularly because several programmatically relevant causes are relegated to the often-unused fourth digit in the codes. Codes for the upcoming ICD-11 revision are being drafted, providing the opportunity to develop more appropriate, clinically relevant coding for neonatal causes. Additionally, ICD coding practices can vary between and even within countries and over time [60, 61]. Such variations reduce our model's ability to predict true variation in causes of death. Other issues in VR coding include changes during the transition between ICD revisions, differences in relegating certain causes to non-specific/ill-defined cause categories,

and the assumption inherent in our exclusion of such codes that the deaths attributed to them are a random sample of all deaths. Finally, the availability and quality of VR data in a country may change over time, especially in countries with newly emerging surveillance systems. Developing consistent time trend estimates given such changes remains a challenge.

We used multinomial models because they naturally ensure that proportional cause distributions sum to one, and thus the sum of cause-specific deaths equals total deaths. Single-cause models require post-hoc adjustments to retain this property, and there may be limited information on which to base such adjustments. An important concern for both types of models, however, is the attribution of death to a single cause. This does not allow for co-morbidities, which are a frequent occurrence in neonatal deaths, and may thus underestimate the impact of a given cause.

Given the considerable variations in health systems and contextual factors within individual countries, subnational neonatal COD estimates are needed and should be a target for future estimation exercises and data collection. National-level estimates like ours aim to ascertain the average causal distribution for a country, which can help guide national priorities, but may mask subnational variation. Some countries are beginning to collect the necessary information for subnational estimates. For example, our national India estimates were produced by aggregating state-level estimates. In the future, we also hope to further differentiate causes within the current broad categories like congenital disorders. However, such differentiation may only be possible for VR-based models, as VA-based data generally lack the detail needed to do this. Finally, we strongly believe that the production of such estimates should be transparent. In accordance with this, the datasets and Stata code we used for this analysis are available on the WHO Global Health Observatory website.

Neonates constitute an important component of the unfinished agenda of the fourth MDG, and reducing preventable neonatal deaths will be essential to achieve the “grand convergence” to which the global community now aspires [157]. The SDGs include a target of 12 or fewer neonatal deaths per 1,000 live births for all countries by 2030 [138], which provides fresh impetus for ending preventable newborn deaths and a future in which every baby has an equal chance of survival.

Theme 2: Improving modelling techniques

5 Issues with the current neonatal multinomial cause-of-death models

While revising the neonatal cause-of-death (COD) models to produce separate estimates by neonatal period (chapter 4), we identified modelling issues that we wanted to investigate further. In this chapter, I focus on three key issues that we identified. The first two, predictive accuracy and model stability, are separate but interlinked aspects of model performance. The third issue is whether and how to give more weight to country-specific empirical data when producing a country's modelled COD estimates. The work presented here was not an exhaustive investigation into these issues, but instead an attempt to understand how they may affect our estimates and identify possible solutions as we determined model improvement priorities. As such, the focus of this chapter is on potential weaknesses in our existing modelling approach rather than strengths; I provide a description of the strengths in sections 4.5 and 8.3.

5.1 Model performance

5.1.1 Introduction

Evaluating model performance is always important, but especially so for prediction models like ours where the main priority is to produce estimates that are as reliable as possible. This includes accuracy (i.e. unbiasedness), precision, and repeatability. In this section, I investigate two aspects of model performance that we identified for further investigation with our models: model stability and predictive accuracy.

Model instability, or large changes in the model due to small changes in the input data, can lead to unreliable results. In our case, changes in the model could include different covariates selected for inclusion in the regressions and/or differences in the estimated coefficient values. Substantial changes in the predictions derived from the model results would mean unreliable results. Mitigating this instability as much as possible is important for developing robust estimates. To evaluate predictive accuracy, we would ideally have separate datasets for building a model and then validating it [158]. These are commonly referred to as training and test datasets. Predictive accuracy of a model can be assessed by applying the coefficients estimated from the training dataset to the separate validation (i.e. test) dataset. Comparing the model predictions to the data in the test dataset thus results in true out-of-sample validation. However, in data sparse situations such as ours, predictive accuracy is typically gauged using the input dataset (or variations of it, e.g. using k-fold cross-validation) for both building and

validating the model [159]. Although not ideal, this type of internal validation is often the only possible solution when working with small datasets.

In the rest of this subsection, I present preliminary results of predictive accuracy and stability tests in our models, investigate factors that may affect model performance, and identify some potential solutions.

Note: the following metrics are used in this subsection.

- Chi-squared goodness-of-fit statistic (χ^2): this metric is the sum of the squared differences between observed and estimated deaths divided by estimated deaths for each cause. The equation is as follows:

  $$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \qquad \text{(Eq. 5.1)}$$

  where O is the number of cause-specific observed deaths, E is the number of cause-specific estimated deaths, and i is the observation (i.e. country-year or study).

- % reduction in residuals from the null (herein "% reduction from null"): this goodness-of-fit metric is used to select the best fitting equations during the covariate selection process. It is calculated by comparing the chi-squared statistic (Eq. 5.1) of the selected equation with that of the null model with no covariates. A positive value indicates an improvement in out-of-sample performance over the null model, while a negative value indicates worse performance. The equation is as follows:

  $$\%\ reduction\ in\ residuals = \frac{\chi^2_{null} - \chi^2_{equation}}{\chi^2_{null}} * 100 \qquad \text{(Eq. 5.2)}$$

- Average chi-squared statistic ($\overline{\chi^2}$): this variation of the χ^2 metric is used to compare models with different numbers of observations. The equation is as follows:

  $$\overline{\chi^2} = \frac{\chi^2}{n} \qquad \text{(Eq. 5.3)}$$

  where χ^2 is the chi-squared statistic (Eq. 5.1) and n is the total number of observations (i.e. country-years or studies).

- Normalized root mean square error (NRMSE): this metric is conventionally used for certain model averaging methods (e.g. bagging) to estimate the error rate between estimated and observed deaths. The equation is as follows:

  $$\sqrt{\frac{\sum_i (O_i - E_i)^2}{n}} / (O_{max} - O_{min}) \qquad \text{(Eq. 5.4)}$$

  where O is the number of cause-specific observed deaths, E is the number of cause-specific estimated deaths, i is the observation, n is the total number of observations, and O_{max} and O_{min} are the maximum and minimum number of deaths for the given cause

in the observed data. The denominator (O_{max}-O_{min}), used for standardization, measures the total range of variation of the observed quantity. The NRMSE thus tells us what fraction of this variation corresponds to the residual root mean square error.

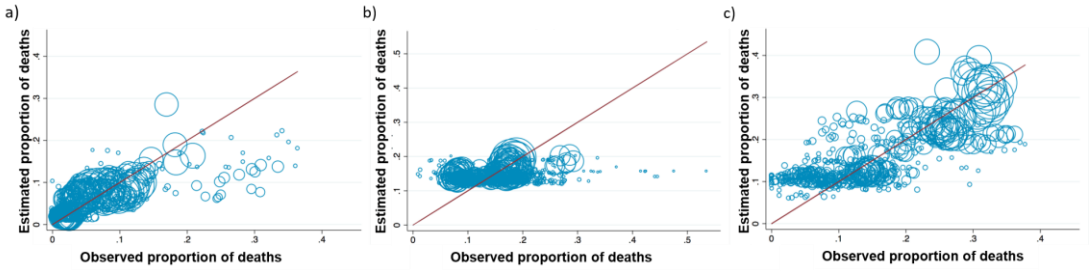
5.1.2 Do we have model performance issues?

Predictive accuracy

To evaluate predictive accuracy, we used our standard two-step modelling approach for neonatal COD estimates (section 4.2.1) with a slight modification (italicized): 1) covariate selection based on minimizing the jackknife out-of-sample prediction error for each non-baseline cause, and 2) running the multinomial regressions using the selected covariates and applying the resulting coefficients to covariates in the *input* dataset. By applying the coefficients to the input dataset, we can compare the predictions from our model to the observed data. We expect this to be an overestimate of our model performance since we used the input dataset for both training and testing the model, and thus overfitting is likely. However, this should provide an initial indication of how well our models are performing as a best-case scenario. We compared these modelled estimates with the observed data to evaluate the model performance. We did this for all four models: low mortality (early and late neonatal periods) and high mortality (early and late neonatal periods). We used the same input and prediction datasets as described in section 4.2.2, with the following updates: 1) the prediction datasets now had data through 2015 instead of 2013, 2) recent VR data (through 2014) was added to the low mortality model input dataset, and 3) 8 new studies with a total of 2,102 new deaths were added to the high mortality model input dataset. See Appendix C.1 for updated input datapoints.

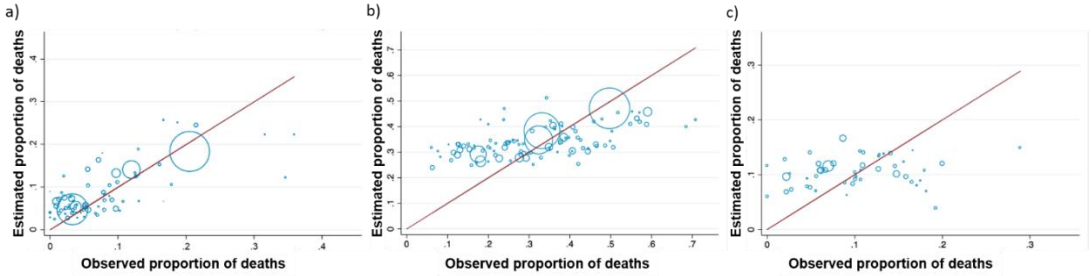
The predictive accuracy of the low mortality model was variable (Figure 5.1). Some causes had overall reasonable agreement between the observed and estimated values in the validation (e.g. Figure 5.1a) while others had poorer agreement (e.g. Figure 5.1b). For the latter (intrapartum; early period), the model predicted little variation in the proportion of deaths while the observed data suggested substantial variation. A few causes had good agreement on average but with a wide spread, suggesting poorer agreement for some observations (e.g. Figure 5.1c).

Figure 5.1: Low mortality model validation examples: observed versus estimated cause-specific estimates. a) pneumonia/late; b) intrapartum/early; c) sepsis/late. Note: circles are weighted by total neonatal deaths for given country-year compared to all other country-years.



A similar pattern was seen for the high mortality validation exercise (Figure 5.2). Some causes had reasonable agreement between the observed and predicted estimates (e.g. Figure 5.2a), some had a more mixed performance (e.g. Figure 5.2b), and others showed poorer agreement (e.g. Figure 5.2c).

Figure 5.2: High mortality model validation examples: observed versus estimated cause-specific estimates. a) congenital/late; b) preterm/early; c) pneumonia/early. Note: circles are weighted by total neonatal deaths for given study compared to all other studies.



This exercise indicates that our models tend to "squeeze" the cause-specific proportions into a narrower range than the observed proportions. (i.e. the observed data typically have wider ranges than the predicted) (Table 5.1). Some of this "squeezing" of the predicted estimates appears to be severe. For instance, in both models there were causes with a predicted upper bound that was at least 30 percentage points lower than the observed upper bound (Table 5.1).

Table 5.1: Observed versus predicted cause-specific proportions by model and period

	Low mortality model				High mortality model			
	Early period		Late period		Early period		Late period	
	Obs[1] range (%)	Pred[2] range (%)	Obs range (%)	Pred range (%)	Obs range (%)	Pred range (%)	Obs range (%)	Pred range (%)
Intrapartum	1-54	12-20	0-25	6-11	4-51	17-36	0-51	11-35
Congenital	3-65	7-33	5-66	6-42	0-35	3-20	0-36	2-26
Preterm	8-79	39-54	0-65	20-35	6-71	24-51	0-69	19-47
Sepsis	0-26	1-12	0-38	6-41	0-58	1-23	0-87	2-31
Pneumonia	0-31	0-10	0-36	2-28	0-29	4-17	0-44	6-17
Injuries	0-9	0-3	0-10	0-2	---	---	---	---
Diarrhoea	---	---	---	---	0-21	0-8	0-33	0-8
Tetanus	---	---	---	---	0-50	0-29	0-50	0-32
Other	0-41	0-16	0-29	5-16	0-35	1-13	0-35	1-13

[1] obs = observed; [2] pred = predicted

Some level of "squeezing" is expected because random stochastic noise will tend to widen the range of the observed data compared with the range in the underlying proportions. The squeezing seen here brings the estimates closer to the average, thereby removing some of the variability seen in the observed data. The wide range in data quality and sample sizes of our input data increases the likelihood of high levels of stochastic noise in the observed data. Thus, pulling of the modelled estimates towards the mean is arguably reasonable. However, the severity of some of the squeezing may limit the models' ability to predict variations in the proportions of certain causes. I discuss potential reasons for suboptimal prediction accuracy in section 5.1.3.

Model stability

While running our COD models, we noticed a potential model instability issue. Our choice of covariates for each equation in the multinomial model is based on out-of-sample goodness-of-fit (GOF) as described in section 4.2.3. A potential source of instability is if there are large differences in the choice of covariates as a result of small changes in the GOF metric. We sometimes noticed the presence of such instability in this work. For example, the difference between the first- and second-best equations based on GOF for intrapartum (early; low mortality model) was only 3.7% (χ^2 of 17025 versus 17649). But there was no overlap in covariates between these top two equations: female literacy for the top equation; DPT, GNI, and U5MR for the second-best equation. A similar example in the high mortality model was tetanus (early period). For this cause, the two best equations based on GOF both had female

literacy and the early/late period covariates. But the first equation also had BCG and NMR while the second had ANC and DPT instead. The GOF difference between these equations was only 1.4% (χ^2 of 1619 versus 1642).

The key concern here is that slightly different GOF metrics may yield different covariate equations which could potentially lead to substantially different final COD estimates. Differences in which covariates are selected for an equation are not necessarily a problem if 1) the primary goal is predictive accuracy rather than model interpretability and 2) the predicted estimates themselves are stable. The latter is possible because the modelled covariate coefficients in different equations can still yield similar final estimates. However, such stability is more likely for the global average rather than across all predicted country-years, and ultimately our models need to be stable at the country-year level. Furthermore, there is no guarantee that there will be similar final estimates when there are large differences in which covariates are included in a cause equation. For instance, despite only a 3% difference in GOF, the modelled numbers of deaths for sepsis (early; high mortality model) were 97,000 (5.6%) versus 245,000 (14.3%) between the first- and second-best models. While large differences in number of deaths between top models were uncommon in our results, such instability is impossible to predict in advance. Therefore, even this simple demonstration of model instability is enough to warrant further investigation because it is an indication that such differences could occur again, unpredictably, in future models. In the next subsection, I discuss our work on ascertaining some potential causes of model instability.

5.1.3 Factors affecting model performance

Model performance can be affected by several factors. Here, I discuss quality and quantity of the COD input data, covariate availability and quality, the strength of the covariate-cause relationships, and the covariate selection method. The choice of modelling strategy, which is also an important factor, is discussed in more detail in chapters 6 and 7.

Quality and quantity of the cause-of-death input data

As noted in section 2.3, the COD input data in our models are of variable quality. This is particularly true for the high mortality model, which is largely based on VA studies. Some concerns with VA data include misidentification of causes, unclassified deaths, and use of different case definitions and causal hierarchies across studies (section 2.3.2). Some of these can manifest as systematic biases in the data (e.g. inevitably underreported causes in VAs like

"hidden" congenital abnormalities). Others add "noise" to our input data, which can obscure true covariate-cause relationships. The lack of standardization for both VA methods and reporting of results for publication can lead to noisy input data in our models even when data may otherwise have been comparable. Additionally, the COD input dataset is moderate in size, with about 100,000 deaths across 124 observations from 95 studies in the most recent dataset version. Only 37 of 81 countries in the high mortality model have COD input data. Issues like noise, overfitting, and outliers can be more pronounced in smaller datasets.

The low mortality model has fewer COD input quality issues because it uses high-quality VR data. Although there are some quality concerns with VR data (section 2.3.1), these are less extreme than those found in the VA data. There are also more input data, with over 2.2 million neonatal deaths across 4,000 observations from 73 countries. Thus, the data issues noted for the high mortality model may be less of a concern (though model instability is not precluded, as demonstrated by the intrapartum example in section 5.1.2). One potential issue for the low mortality model, however, is that the input versus prediction countries do not overlap (i.e. prediction countries do not have input data). Both the input and prediction countries have many similarities, including low mortality levels (less than an NMR of about 10). But the input countries are on average wealthier and have somewhat better health indicators (e.g. median NMR of 4.7 vs 9.1 for input versus prediction countries; Table 4.5a). This has the potential to affect predictive accuracy (even if the model had good validation performance) if the covariates have different relationships with the COD distribution in these different country groups. Including lower-quality VR data from modelled countries into the input dataset, and the trade-off this involves between data quality/noise versus inclusivity, is discussed later (section 8.4).

The data quality and quantity issues discussed here have the potential to affect predictive accuracy and stability of the models. For example, different causal hierarchies being used across studies can mask true covariate-cause relationships. Such relationships are the cornerstone of regression analyses, and thus such data issues can weaken our ability to produce robust, stable estimates. Unfortunately, we have no control over these issues without being directly involved in original data collection. I provide recommendations for improved COD data collection and reporting in section 8.5.

Covariate availability and quality

The underlying assumption of our modelling strategy is that a set of covariates, in an appropriate modelling framework, can be used to predict the outcome. The accuracy and stability of the predictions thus depends partly on the covariates themselves. The two main covariate-related issues that may affect the performance of our models are availability of covariates and noise.

Availability of covariates

In our models, we included those covariates which had annual, nationally comparable estimates available across 194+ countries from at least 2000 onward (section 2.4.1). The covariates we were able to include cover a wide range of areas, from indicators directly linked to neonates (e.g. NMR, SBA, ANC) and various health system factors (e.g. vaccination coverage) to broader socioeconomic metrics (e.g. GNI, GFR). The restriction on being able to include only covariates with nationally comparable time series is necessary for the type of model we use and the aim of our work (i.e. country-year predictions). Ideally, however, we would include in our models all covariates that best predict the COD distribution. This means including the "right" covariates (i.e. those that have the strongest predictive power).

The covariates we were unable to include are heterogenous. Some are expected to be directly linked to specific causes (e.g. access to antibiotics for infection) while others likely affect a number of causes (e.g. exclusive breastfeeding). Some are directly associated with the health system (e.g. facility births) while others capture broader socioeconomic factors (e.g. occupation or neighbourhood characteristics). There are also factors that likely have some association with the COD distribution but are difficult or nearly impossible to capture as covariates. For such factors, a variable (or some combination of variables) could be used as a proxy, but this may not fully capture the intended information. For example, we included vaccination coverage covariates as proxies for access to and availability of basic healthcare interventions. These proxies may be reasonable given the available covariates, but they do not necessarily capture the full extent of information we would have liked to include.

Several of the covariates in our models are coverage rather than quality indicators. For example, our models include ANC and SBA coverage, which indicate the percentage of women attending antenatal visits or giving birth with a skilled birth attendant present. But the quality of these interactions may be particularly important for the COD distribution. Measuring quality instead of only coverage is a complicated issue, but increasingly acknowledged as crucial [160]. Such quality-based covariates are not yet reliably available as nationally comparable time series

[161], so we were unable to include them in our models. But these would be important to consider for inclusion if and when such time series do become available.

Noise from covariates

Another important covariate issue is "noise". Two main sources of noise in our covariate dataset are measurement error (including through incomplete data) and stochastic noise due to finite sample sizes.

First, covariates can be difficult to measure. Some, like the percentage of women attending ANC visits, typically come from self-reported responses in cross-sectional household surveys and thus can suffer from biases common to such surveys. Even a core annual economic measure like GNI can be underestimated in countries with large informal sectors [162], which are common in low-income countries. Other covariates, like SBA, are either difficult to define precisely or it is difficult to apply their definition precisely when collecting data.

There can also be differences in measurement certainty within a covariate. The covariate data in our models are a mixture of data reported by countries, determined from household surveys, modelled, and/or imputed (section 2.4). A particular covariate time series may have a combination of these sources across countries or over time for the same country. For example, the uncertainty in NMR estimates varies widely across countries, largely because of the mortality and birth data sources available in a country (section 2.2). Such a mix in quality within a covariate series can add noise to the dataset. Some of these noise issues are mitigated in our models since we run the low and high mortality models separately. The covariate data for the low mortality models tend to have less uncertainty (e.g. NMR calculated from CRVS data), while those for the high mortality models have more uncertainty (e.g. NMR modelled based on survey data). Other causes of variation within a covariate series are also important. The nationally comparable covariate time series data we use involve many datapoints. For example, a time series for 194 countries from 1980 to 2015 consists of 6,984 country-year observations. Ensuring that the data within these time series are truly comparable is a difficult task, especially when covariate definitions or data collection methods change over time.

Within the high mortality input data, an additional source of noise is due to the mismatch of local-level COD distributions against regional/national-level covariates. Although we use the local-level COD distribution from a given study, we do not always have corresponding local-level

covariate information (section 2.4). We use regional or national-level covariates when such local-level data are missing. This noise can further mask true covariate-cause relationships.

There is substantial heterogeneity in data quality for the covariates in our models, both between and within covariates. This is because of a combination of what is being measured, how it is measured, and how the time series data are put together. Such variation may not always be known or fully understood, and is usually impossible to specifically identify within national covariate time series data. We have worked to reduce some of these noise issues by verifying that the covariate data are reasonable (e.g. do not have unrealistic spikes in the time series), and working directly with the World Health Organization (WHO) to correct such noise issues in their covariate time series data through smoothing algorithms and searching for additional data sources.

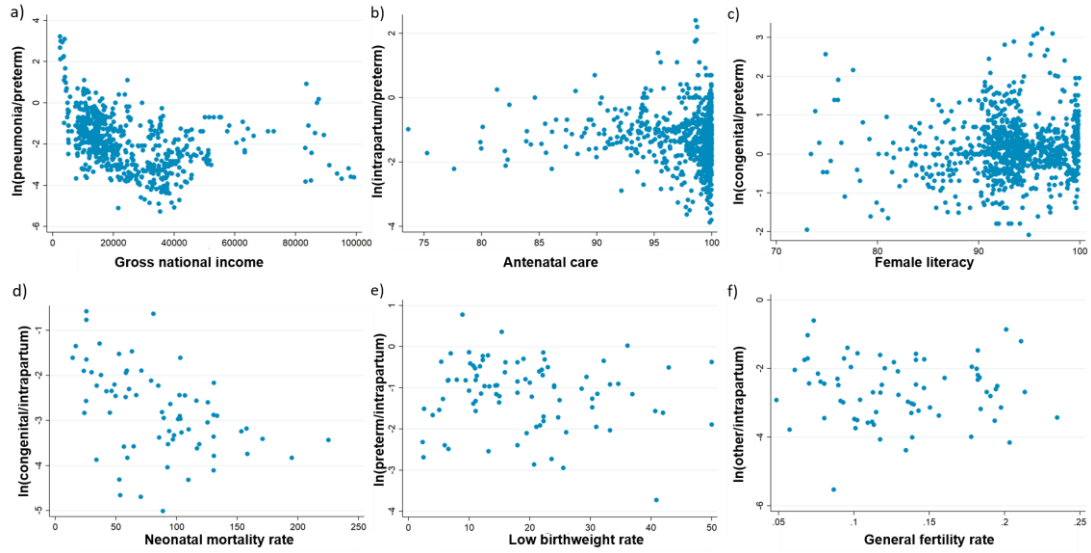
Covariate and cause relationships

The cornerstone of a regression analysis is the relationship between the covariates (i.e. independent variables) and the outcomes (i.e. dependent variables). The stronger the relationship between a covariate (or set of covariates) and the COD distribution, the better our ability to predict this distribution. Without a strong relationship, both the predictive accuracy and stability of the model will be more challenging to achieve. More minor but still important relationships include those between covariates and those between causes.

Relationships between causes and covariates

Evaluating the relationship between a covariate and outcome is an important step in both understanding the strength of their relationship and whether the right functional relationship is being used (e.g. linear versus quadratic). To assess the relationship between the covariates and causes of death in our models, we graphed each covariate against the respective outcome for each cause (e.g. female literacy versus log(other/intrapartum) in the high mortality model). Figure 5.3 includes representative examples of the covariate-cause relationships in our data in the low mortality model (top row) and high mortality model (bottom row). There were some covariate-cause pairs which seemed to visually have a relatively strong association. For example, both pneumonia/GNI (Figure 5.3a) and congenital/NMR (Figure 5.3d) appear to have decreasing proportional burden as the covariate value increases, which is to be expected. However, the majority of covariate-cause relationships appeared to be cloud-like without clear patterns (e.g. Figures 5.3b-c and 5.3e-f). This included instances for which we would expect to have seen a relationship, such as intrapartum/ANC (Figure 5.3b) and preterm/LBW (Figure 5.3e).

Figure 5.3: Examples of covariate-cause relationships in the low (top row) and high (bottom row) mortality models for the early neonatal period. a) pneumonia and GNI; b) intrapartum and ANC; c) congenital and female literacy; d) congenital and NMR; e) preterm and LBW; and f) other and GFR.



For simplicity, we graphed these associations for single covariate-cause pairs instead of the multidimensional combinations that are relevant for multinomial or multivariate models. Multinomial models have additional complexity because the multiple equations are fit simultaneously, and thus the covariates for one cause could affect the estimates of another cause. However, the single covariate-cause graphs still give an indication of the strength of these relationships.

Overall, these analyses indicate relatively weak relationships between many of the cause-covariate pairs. This is an important issue, but it lacks a clear immediate solution. The strength of these relationships is likely due (at least partially) to some of the COD and covariate data issues described earlier in this section, including data quality/noise and covariate availability.

Relationships between covariates

Multicollinearity, which occurs when covariates are strongly correlated, can be a source of instability and thus affect predictive accuracy. We evaluated multicollinearity in our input data and found that IMR and U5MR were, unsurprisingly, highly correlated with NMR in both models (>0.8 correlation coefficient in the high mortality model; >0.9 in the low mortality model). Retaining just one of a set of highly correlated variables in the model is a reasonable solution to this problem.

Relationships between causes

In the COD distribution, some causes are more closely linked than others. This can contribute to misclassification of deaths. For example, sepsis and pneumonia are difficult to distinguish in the neonatal period because of similarities in symptoms and challenges with diagnostics. In other cases, it may be difficult to distinguish an underlying cause from the immediate cause. For example, a death from the underlying cause of respiratory distress syndrome in a preterm baby may be coded as birth asphyxia (i.e. intrapartum) instead of "preterm", especially if the gestational age is miscalculated (e.g. the baby is truly preterm but is misidentified as full term). Finally, some causes may be systematically miscoded. For example, hidden congenital abnormalities (e.g. cardiac abnormalities) are likely to be systematically under-reported, potentially into the "other" category.

Medium- to long-term solutions to help mitigate some of these problems are discussed in chapter 8.

Covariate selection method

In work such as ours, the covariate selection approach is important to consider when investigating model performance because it dictates which covariates are included in the final models. Covariate selection methods can contribute to model stability or instability depending on the modelling approach and data involved. Some general considerations when picking a covariate selection method include balancing appropriateness, completeness (e.g. how many covariate combinations to test), and computational intensity/time.

There are two issues with our current method which could affect model performance and therefore warrant further investigation: 1) our approach does not evaluate all possible covariate combinations and 2) we use binomial covariate selection for multinomial models. Additionally, it is useful to look closely at the metric used to calculate GOF since this is also linked to which covariates are chosen for a model. In this subsection, I discuss each of these and whether they affect model performance.

Summary of current covariate selection method

As described in section 4.2.3, our COD modelling method involves a two-step procedure: 1) covariate selection to choose a set of covariates for each COD equation and 2) running the multinomial regressions using the selected covariates to produce country-specific COD distributions. The covariate selection approach uses a version of forward stepwise selection

where we sequentially add covariates depending on their improvement to out-of-sample GOF under a jackknife process. The out-of-sample GOF is calculated using the chi-squared GOF metric (Equation 5.1). This covariate selection process (using all binomial models) has the following steps:

1) Perform univariate regressions on linear, quadratic, and spline forms of each covariate and select the form with the best GOF to include in step 2
2) Run multivariable regressions
    a. rank covariates based on GOF from step 1
    b. use best ranked covariate to begin building the multivariable regression
    c. add to the multivariable regression the covariate with next best GOF which is not already in the regression, and run the regression
        i. If overall GOF worse/same, drop covariate and repeat step c
        ii. If overall GOF better, retain covariate and repeat step c

Once the set of covariates with the best out-of-sample GOF for each cause/baseline equation is identified, we include them in the multinomial model, and re-estimate all of the model coefficients, from which we derive the country-year predictions. At present, we do not include a threshold of GOF improvement for inclusion in the model; any covariate which improves the GOF in step c is added to the equation. A threshold could be added, however, as is done with some conventional covariate selection methods.

Searching over all possible covariate combinations

For each cause equation, there are 2^n possible covariate combinations. Assuming n=13 covariates (i.e. 8,192 covariate combinations), 7 non-baseline causes, and separate models for the early and late neonatal periods, this means evaluating 57,344 regressions for GOF. Using jackknife to perform out-of-sample GOF, assuming 100 study observations (e.g. similar to our high-mortality model), therefore means 5,734,400 regressions. If each regression took one second (as it did in our test), this would be about 9.5 weeks of computation on a single core computer. Although the low mortality model has one less cause category and fewer covariates included, it has far more observations and deaths in the database, which make fitting each regression slower. Overall, performing covariate selection using out-of-sample GOF on all covariate combinations can be a very computationally intensive task. It is for this reason that we have been using the approach described above, which is a partial instead of full search over the multidimensional covariate space and therefore substantially less computationally intensive.

However, the only way to evaluate whether our existing "approximation" is working well is to compare its results to those obtained through searching the full space of covariate combinations. We performed this comparison for each of the four models. We included 13 covariates in the high mortality model, excluding U5MR and IMR for feasibility (thereby reducing the possible covariate combinations from 32,768 to 8,192). For the low mortality model, we included 10 covariates (i.e. 1,024 possible covariate combinations). For full comparability, we re-ran the high mortality covariate selection partial search algorithm excluding U5MR and IMR as well.

The partial search algorithm was highly efficient. The average number of covariate combinations tested by the partial search algorithm for the high mortality model ranged from 26 to 38 across cause and neonatal period, which is less than 0.5% of the 8,192 combinations tested with the full search (Table 5.2). This algorithm was similarly efficient for the low mortality model; it searched less than 2.5% of the 1,024 combinations for each equation. For most causes, the partial search performed well. The median percentage point differences in the % reduction from null metric were 2.9 (high mortality model) and 1.3 (low mortality model) between the best fit full versus partial search models (Table 5.2). The partial search algorithm performed less well for a few causes, with up to a 39.5 percentage point difference in % reduction from null for congenital in the early period high mortality model (Table 5.2).

Table 5.2: Comparison of best fit cause equation results using a full versus partial search of covariate combinations for the high and low mortality models by neonatal period

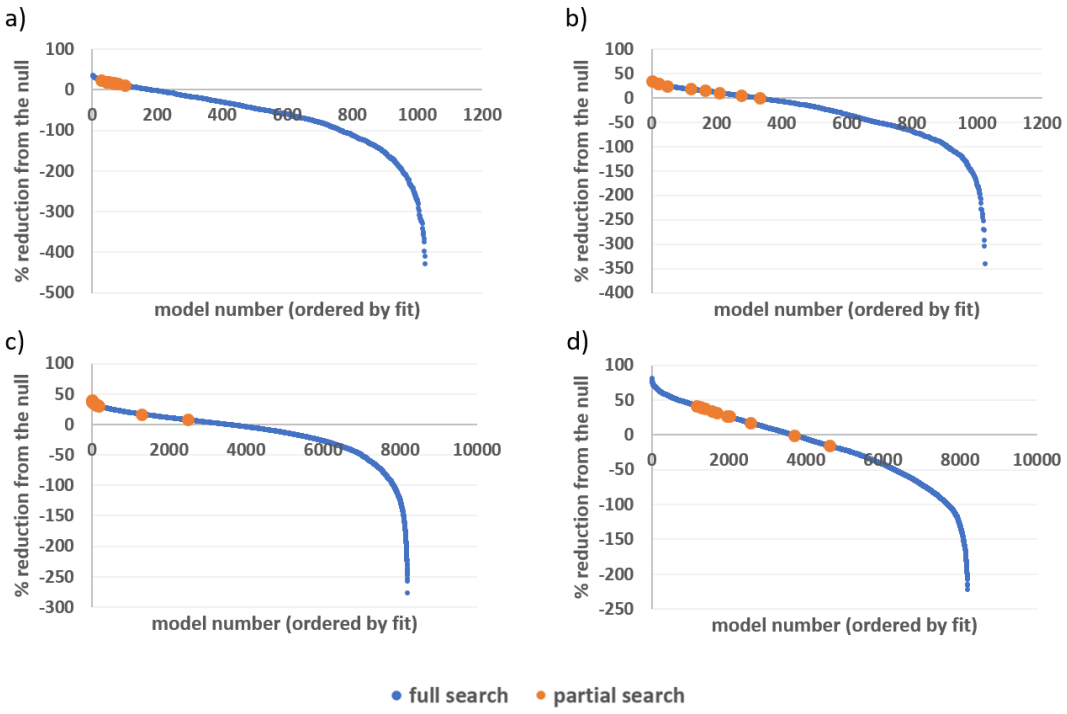
	# of covariate combinations tested in "partial search"[1]	Rank of best "partial search" model in "full search"	% reduction from null of best "full search" model	% reduction from null of best "partial search" model	Absolute difference between % reduction from null for best "full" and "partial" models
High mortality model; early period					
Preterm	26	30	53.2	44.1	9.1
Congenital	26	1154	81.4	41.9	39.5
Sepsis	28	3	81.1	80.7	0.4
Pneumonia	26	1	13.5	13.5	0.0
Diarrhoea	35	2	81.1	79.1	2.0
Tetanus	34	12	87.8	87.0	0.8
Other	26	38	85.5	80.2	5.3
High mortality model; late period					
Preterm	36	589	79.1	62.1	17.0
Congenital	34	60	91.4	87.7	3.7
Sepsis	33	1	39.5	39.5	0.0
Pneumonia	26	1	27.3	27.3	0.0
Diarrhoea	26	1	30.5	30.5	0.0
Tetanus	27	94	70.3	66.6	3.7
Other	38	17	64.7	57.8	6.9
Low mortality model; early period					
Intrapartum	20	1	13.9	13.9	0.0
Congenital	21	2	52.7	52.2	0.5
Sepsis	24	16	59.5	57.3	2.2
Pneumonia	20	27	36.4	23.4	13.0
Injuries	20	8	17.7	13.9	3.8
Other	24	23	26.7	18.3	8.4
Low mortality model; late period					
Intrapartum	20	1	20.1	20.1	0.0
Congenital	25	2	8.9	8.7	0.2
Sepsis	25	20	66.5	63.2	3.3
Pneumonia	19	2	34.2	33.9	0.3
Injuries	20	2	0.2	0.0	0.2
Other	25	9	29.7	27.4	2.3

[1] out of 8,192 possible combinations in high mortality model and 1,024 in low mortality model

Figure 5.4 shows examples of results for the % reduction from null calculations for the full and partial search algorithms. A positive % reduction from null indicates that a model has better out-of-sample prediction performance than the null model; conversely a negative value indicates worse performance than the null. As can be seen in Figure 5.4, only a small fraction of all possible models performed better than the null model. Some of the partial search results were clustered in terms of their GOF across the full space of models (e.g. Figure 5.4a), others were more dispersed (e.g. Figure 5.4b), and some were partially clustered and partially dispersed (e.g. Figure 5.4c). For most causes in each model, the partial search algorithm results

were located in the top third best fit full search results (i.e. left third of the graph). Figure 5.4d shows an example of a partial search result which performed less well (congenital/early; high mortality model).

Figure 5.4: Example graphs of the % reduction from null for the full versus partial search algorithm results. a) pneumonia/early, low mortality model; b) pneumonia/late, low mortality model; c) sepsis/late, high mortality model; d) congenital/early, high mortality model.



Overall, the results suggest that the partial search is highly efficient (i.e. about 1 minute per cause) and effective in most cases. However, there are times when it performs less well (Figure 5.4d), and it is impossible to know in advance how well the algorithm will perform. The full search algorithm is the only way to guarantee that the selected covariate combination is the best fitting of all possible combinations, but it is computationally intensive to run. For these models, the full search with jackknife out-of-sample and multiple computer cores took 1-2 days per cause (i.e. about one month for the four models). This also restricts the number of covariates that can be feasibly included (e.g. 13 versus 15 covariates is a difference of nearly 25,000 covariate combinations to test before out-of-sample considerations). Based on this analysis, it is debatable whether the benefits of the full search (i.e. guarantee to find best fitting model) outweigh the benefits of the partial search (i.e. highly efficient, mostly good performance, can include more covariates).

Binomial covariate selection for multinomial models

Our current covariate selection process involves choosing covariates that produce the best out-of-sample predictions for single cause/baseline equations. The covariates chosen through this binomial approach are then used in a multinomial model. We chose to use binomial equations for the covariate selection because a similar multinomial covariate selection approach would result in a combinatorial explosion that would be computationally prohibitive. However, it is possible that the equations selected based on binomial covariate selection have a worse GOF once placed in the multinomial environment where all of the equations are fitted simultaneously.

We conducted an analysis to compare how well-aligned the out-of-sample GOF metrics were for each cause between its binomial covariate selection regression and the relevant multinomial model. To evaluate out-of-sample GOF in the multinomial, we did the following for each of the four models: 1) included the best fit equations from the binomial covariate selection in the multinomial (as usual) and 2) performed jackknife out-of-sample validation using the multinomial model on the input dataset. The second step involved refitting the model parameters for each jackknife sample (i.e. n-1 observations) and applying the coefficient values to the excluded study or country-year observation. To compare the results, we calculated the average chi-squared ($\overline{\chi^2}$) metric for each cause in the binomial and multinomial models.

Table 5.3 shows the results of this out-of-sample GOF comparison for the binomial and multinomial results. The binomial and multinomial GOFs appear to be relatively similar for most of the causes, with the multinomial on average having slightly worse GOF. This is unsurprising since the strategy is 'optimized' for the binomial, and not multinomial, setting. However, the multinomial GOF was substantially worse for a few causes in the high mortality model. This is starkly evident for sepsis in the early period, with a $\overline{\chi^2}$ of 286.3 for the multinomial and 7.7 for the binomial (Table 5.3), and present but less severe for a few other causes in the high mortality model (e.g. other/early; sepsis/late).

Table 5.3: Comparison of the average chi-squared out-of-sample goodness-of-fit statistic for binomial covariate selection regressions versus multinomial models by cause and period for the low and high mortality models

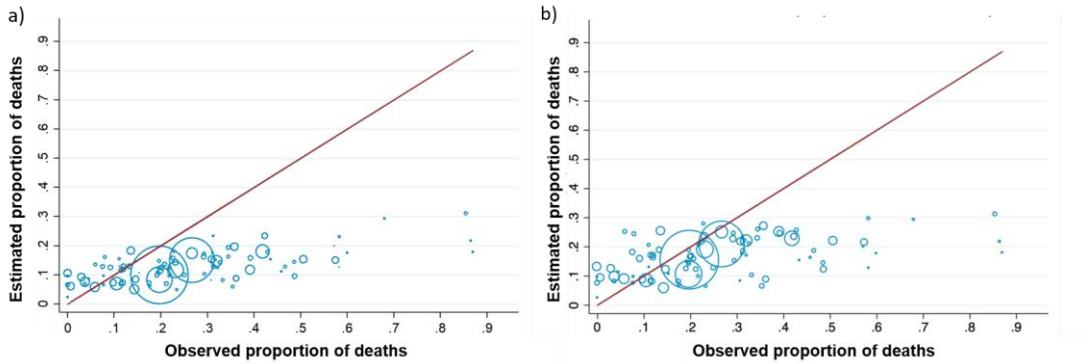
	Low mortality model				High mortality model			
	Early period		Late period		Early period		Late period	
	Binom[1]	Multi[2]	Binom	Multi	Binom	Multi	Binom	Multi
Intrapartum	13	11	4	5	baseline	23	baseline	11
Congenital	10	13	3	4	7	9	4	3
Preterm	baseline	15	baseline	8	6	19	4	10
Sepsis	7	8	5	7	8	286	7	25
Pneumonia	11	15	7	9	5	8	5	7
Injuries	4	4	3	3	---	---	---	---
Diarrhoea	---	---	---	---	7	4	6	5
Tetanus	---	---	---	---	17	35	9	10
Other	20	24	4	6	11	74	9	16

[1] Binom = binomial; [2] Multi = multinomial

We investigated sepsis (early period; high mortality model) further with an aim to better understand which covariate(s) may have been responsible for the substantially worse performance upon inclusion in the multinomial model. To do this, we used a manual non-exhaustive algorithm we developed to help identify potentially problematic covariates when COD estimates had unusual behaviour. The steps were as follows: 1) identify countries with unusual or poorly performing estimates (e.g. substantial spikes in cause-specific proportions; unrealistic proportions); 2) search their covariate values in the prediction dataset to select potentially problematic covariate(s) for further investigation, 3) re-run the covariate selection and model without the potentially problematic covariate (removing one covariate at a time if more than one), 4) compare the original estimates with those using the "tweaked" equations. If no covariate from the relevant cause equation appeared to be causing issues (or if the problematic estimates were for the baseline cause), we looked at the other cause equations as well.

We found that the period covariate appeared to add volatility when we applied this method to the sepsis equation discussed above. We re-ran the covariate selection without this covariate. Figure 5.5 shows validation graphs of the original and the "tweaked" versions for this cause. The tweaked version does appear to be performing better, as is also indicated by the change in its $\overline{\chi^2}$ value from 286 to 49.

Figure 5.5: Multinomial validation for sepsis (early period; high mortality model) comparing estimated versus observed proportions a) before and b) after "tweaking" the covariate equation. Note: circles are weighted by total neonatal deaths for given study compared to all other studies.



A consequence of modifying a cause equation in the multinomial model, however, is that other causes are likely to change as well because of the push and pull that occurs during the simultaneous fitting of the multiple equations. In this case, for example, the performance of congenital worsened from a $\overline{\chi^2}$ of 9 to 28. Although this is worse, the sepsis improvement was substantially greater, and thus such a tweak would be justifiable.

Ideally, the covariate selection strategy should align with the modelling process such that the best fit equations found through covariate selection are the optimal equations to use in whichever model is used for fitting the parameters. The caveat is that ideal strategies are not always computationally feasible. Based on the results of this preliminary investigation, our existing approach seems to be acceptable in most cases. While it is difficult to predict whether the best fit equation selected from the binomial regressions will fit more poorly once in the multinomial model, the multinomial fit can be examined empirically as shown above. Tweaking the covariate selection process as we did above is suboptimal as part of a formal modelling strategy because the approach is manual, not systematic, and not exhaustive. However, an algorithm could be developed to make this process less subjective. For example, our manual process could be coded and automated such that covariates are investigated if the GOF statistic (or the covariate parameter estimates) change a certain amount between the binomial covariate selection and multinomial models. I also discuss some alternative covariate selection strategies in section 5.1.4.

Choice of the χ^2 metric for goodness of fit

We use the χ^2 metric (Equation 5.1), a standard GOF metric, to estimate the out-of-sample GOF in the covariate selection process. The % reduction from null calculation (Equation 5.2) is also

based on the χ^2 metric. Given that covariates are ultimately selected or rejected from the model equations based on their out-of-sample GOF, the choice of the χ^2 metric plays an important role in our covariate selection.

Table 5.4 shows the % reduction from null for the best fit equations of each cause and model for our current models. This provides an indication of how well the selected set of covariates is predicting out-of-sample data compared to a null model (i.e. intercept-only model with no covariates).

Table 5.4: % reduction from null for the best fitting equations by cause and period in the low and high mortality models.

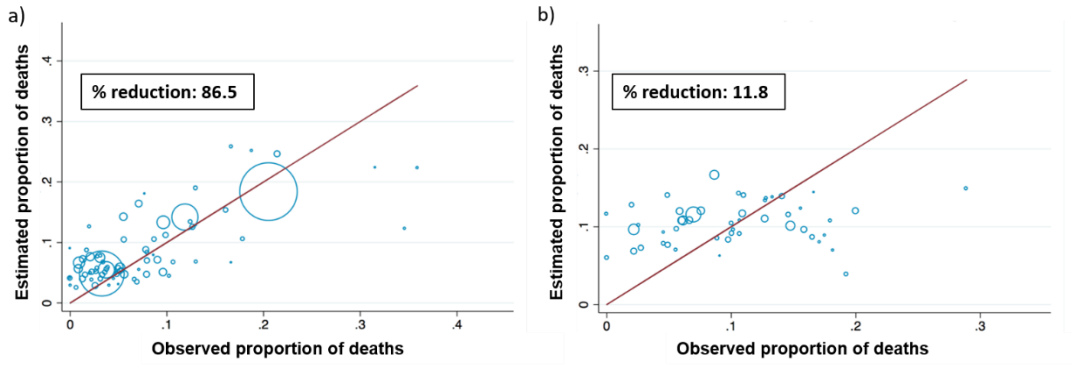
	Low mortality model		High mortality model	
	Early period	**Late period**	**Early period**	**Late period**
Intrapartum	13.9	20.1	(baseline)	(baseline)
Congenital	52.2	8.7	72.7	86.5
Preterm	(baseline)	(baseline)	44.4	63.2
Sepsis	57.3	63.2	82.0	40.1
Pneumonia	23.4	33.9	11.8	31.5
Injuries	13.9	0.0	---	---
Diarrhoea	---	---	83.6	38.9
Tetanus	---	---	88.5	80.2
Other	18.3	27.4	83.0	56.0

Note: High mortality results here differ slightly from Table 5.2 because the partial search algorithm for that analysis was adjusted to exclude some covariates for comparability to the full search algorithm.

There is a relatively wide range in the percent reduction from null, from 0% for injuries (late period; low mortality model) because the null model had the best fit to 88.5% for tetanus (early period; high mortality model). Overall, the high mortality model had higher % reductions from null than the low mortality model.

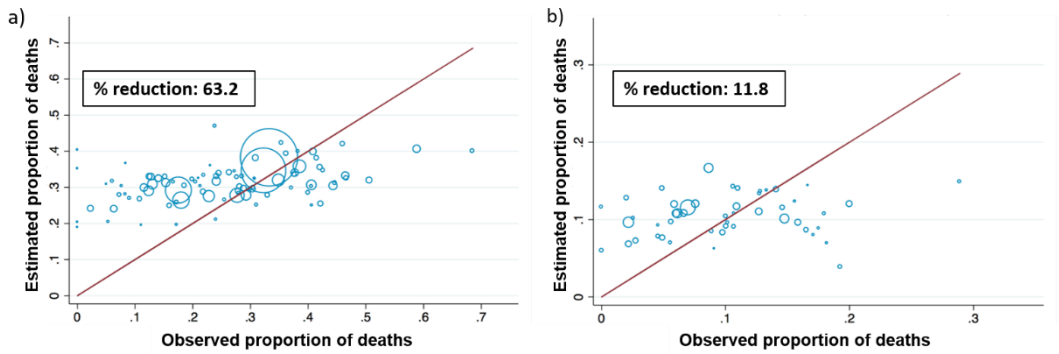
In general, the impression obtained from visual inspection of the observed versus estimated proportions is consistent with the GOF metric. For example, the model appears to perform well for congenital (late period; high mortality model) which had an 86.5% reduction from null (Figure 5.6a). In contrast, the model both appears to perform poorly and has a low 11.8% reduction from null for pneumonia (early period; high mortality model) (Figure 5.6b).

Figure 5.6: Estimated versus observed death proportions for a well versus poorly performing cause in the high mortality model. a) congenital/late; b) pneumonia/early. Note: circles are weighted by total neonatal deaths for given study compared to all other studies.



A few causes, however, appear to have poor GOF but still a relatively high % reduction from null based on visual inspection. This is because the GOF calculation, based on the χ^2 metric, prioritizes the model fit for observations with larger numbers of deaths. An example of the consequence of this can be seen in Figure 5.7. Here, both graphs appear to have poor GOF (i.e. most points are not along the red diagonal line). Yet, the % reduction from null for preterm (late period, high mortality model; Figure 5.7a) is 63.2%, which is substantially higher than the 11.8% for pneumonia (included again for comparison; Figure 5.7b). The key difference appears to be that the model is performing well for the largest studies (indicated by circle size) for preterm, which influences the χ^2 GOF calculation.

Figure 5.7: Model validation for causes with similar performance but divergent % reduction from null in the high mortality model. a) preterm/late); b) pneumonia/early. Note: circles are weighted by total neonatal deaths for given study compared to all other studies.



This may also help explain why the high mortality model on average has higher % reduction from the null than the low mortality model (Table 5.4). The low mortality model has far more input observations than the high mortality model, and there are many datapoints across the range of numbers of deaths (see Figure 5.1 for examples). Thus, even if the model fit attempts

to prioritize larger observations, it is difficult to fit all of them well unless there is a strong relationship between causes and covariates that truly explains most of the variation. In contrast, the high mortality model has only a few large studies, and thus it is easier for the model to prioritize fitting these well. This suggests that a higher GOF may be achievable if a few data points are much larger than the rest.

Such a metric is typically appropriate since large studies tend to be more reliable because they suffer less from stochastic noise effects. In our case, however, there are only a few large studies in the input dataset, and we generally believe that the input data quality is variable, including for the larger studies. Thus, prioritizing the fit towards a few studies, including ones that are not necessarily high quality, may not be ideal. Alternative strategies could include adding weights to average out the emphasis between study points and number of deaths in a study or choosing a different GOF metric. Ultimately, however, this is a relatively minor issue. As noted above, the majority of causes had decent alignment between the visual GOF and the % reduction from null statistic. Additionally, it is justifiable to use a metric which gives higher priority to larger studies (thereby reducing stochastic noise concerns), especially when the actual data quality of each study is unknown.

5.1.4 Potential solutions to improve model performance

Some of the issues highlighted in the previous section are ones which we cannot directly influence (e.g. data quality of input studies) or appear to have relatively minor impact on model performance (e.g. partial search algorithm for covariate selection). Others can be solved simply (e.g. removing covariates that are collinear). In this subsection, I focus on two major categories of methods that stand out as having potential to more broadly improve the robustness of our models. Ensemble methods aim to reduce instability by accounting for model uncertainty. Regularization methods aim to penalize model complexity, which can improve both stability and predictive accuracy.

Here, I briefly describe some of these potential solutions in the context of our models. The goal of this work was to identify some plausible solutions to consider for future implementation. Where possible, we performed basic analyses to assess feasibility; full implementation and validation of each technique was outside the scope of this objective.

Ensemble methods

Ensemble methods are a family of methods used to improve the stability and predictive accuracy of statistical models by combining results from multiple models instead of selecting only one [163, 164]. This approach focuses on the fact that there is usually uncertainty in selecting a model (as demonstrated in section 5.1.2), and therefore selecting a single model may be inappropriate. These methods, increasingly popular in disciplines like machine learning, come in many forms but can be divided into two key approaches: parallel versus serial model building. Serial methods, such as boosting, have been shown to perform more poorly with noisy data [165]. Given the noisiness of our input data, we focus on parallel ensemble methods in the rest of this subsection.

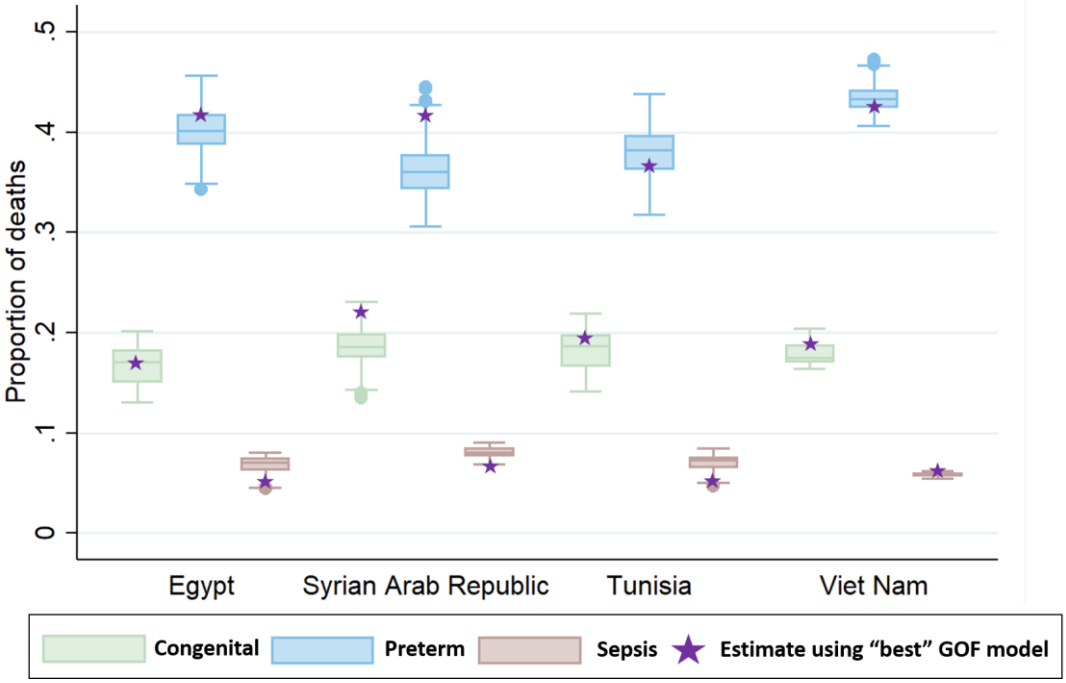
Parallel ensemble methods involve running several different iterations of a model and averaging the results together. The "different iterations" can imply a range of options, including averaging across models that use variations of the input data or ones that use different covariate selection outputs. A consequence of this averaging approach is that there is no single model to output as a result of the analysis, which can make model interpretability difficult. After reviewing the literature on ensemble methods, we decided to test two approaches that we believed were potentially suitable for our models: 1) simple model averaging with random selection from among the top cause-specific equations and 2) bagging. Here we describe each in more detail and describe preliminary work we did to test these with our models.

Simple model averaging

In model averaging, instead of choosing one model through the covariate selection process, several models are built and run in parallel, with their results averaged together (different weighting schemes have been proposed) to account for uncertainty in model selection [163, 166]. We applied simple model averaging as follows: 1) identify the top n covariate equations for each cause based on the out-of-sample GOF using the full search covariate selection algorithm, 2) randomly select an equation from these top n equations for each cause to form the multinomial equation set for the model, run the multinomial model, and produce COD estimates; 3) perform m iterations of step 2; and 4) produce final prediction estimates by averaging together the proportions by cause across all m iterations. This method aims to address instability due to the uncertainty in covariate selection at the individual cause level, but also when included in the multinomial model.

We performed the above analysis for n = 5 and m = 500. Figure 5.8 shows variation in the proportion of deaths from congenital, preterm, and sepsis for four example countries. These results demonstrate that there is potential for variation in COD estimates at the country-level depending on choice of cause equations. For example, the proportional estimate for congenital deaths in Syria had a median value of 0.2, but the IQR was 0.1 to 0.28. The mean values (not shown), which would be used to ensure that the proportions add up to 1, were nearly identical to the median values; only one had a difference of more than 0.5 percentage points (Syria/congenital: mean was 1.2 percentage points lower than the median). These results also indicate that the "best" GOF model estimates are often within the middle 50% of the model averaging results, but can at times be outside that range (Figure 5.8). This model averaging approach could thus provide some stability to our country-level estimates in situations where model uncertainty affects the final estimates.

Figure 5.8: Predicted proportion of deaths in 2013 using model averaging over 500 iterations for three causes across four example countries.



While such an approach decreases some instability, it is unable to deal with a covariate that may be heavily skewing the results but shows up in all or most of the top equations. Potential solutions like manual tweaking to test for and remove problematic covariates are inefficient and not systematic (though algorithms could be programmed as described in section 5.1.3)

Bagging

Bagging, short for bootstrap aggregation, relies on building different models based on random resampling of the input data, and then averaging the predictions from the models [164, 167]. In doing so, this method can help address model uncertainty arising from the input data [168]. Bagging involves first generating n bootstrap samples with replacement from the input dataset. The modelling process is then run on each of the samples (including the covariate selection step), and the average predicted values are used for the final estimates. Bagging has been shown to decrease variance [169] and is considered to be useful for noisy data [170]. The latter is particularly important for our models given the noisiness of our input data. An additional benefit for our models is that this method reduces the risk of a "dominant" covariate adding instability since it is performed on many bootstrapped samples.

We performed a preliminary bagging analysis with 130 iterations for the early period low mortality model. We left out approximately one-third of the observations for each bootstrap sample and then performed predictions on this leftover "out-of-bag" sample with the fitted model. To assess model accuracy, we calculated the normalized root mean squared error (NRMSE) metric (Equation 5.4) on the out-of-bag samples. Figure 5.9 shows an example of the bagging results for one country-year (UK, 2005) in the low mortality model. The decrease and then stabilization in the NRSME in this graph are expected. Here, the NRSME appears to mostly stabilize by about 40 iterations, with almost no further decrease by 100 iterations. This suggests that between 40 and 100 iterations would be sufficient in this case to have improved model stability through reduced variance. These results suggest that bagging may help to increase the stability of our models.

Figure 5.9: A bagging analysis example: the normalized root mean square error by number of iterations for the UK in 2005.

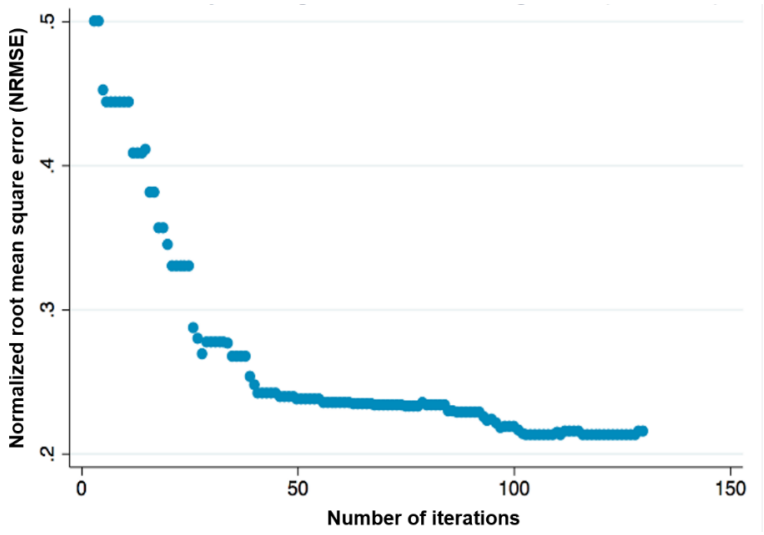
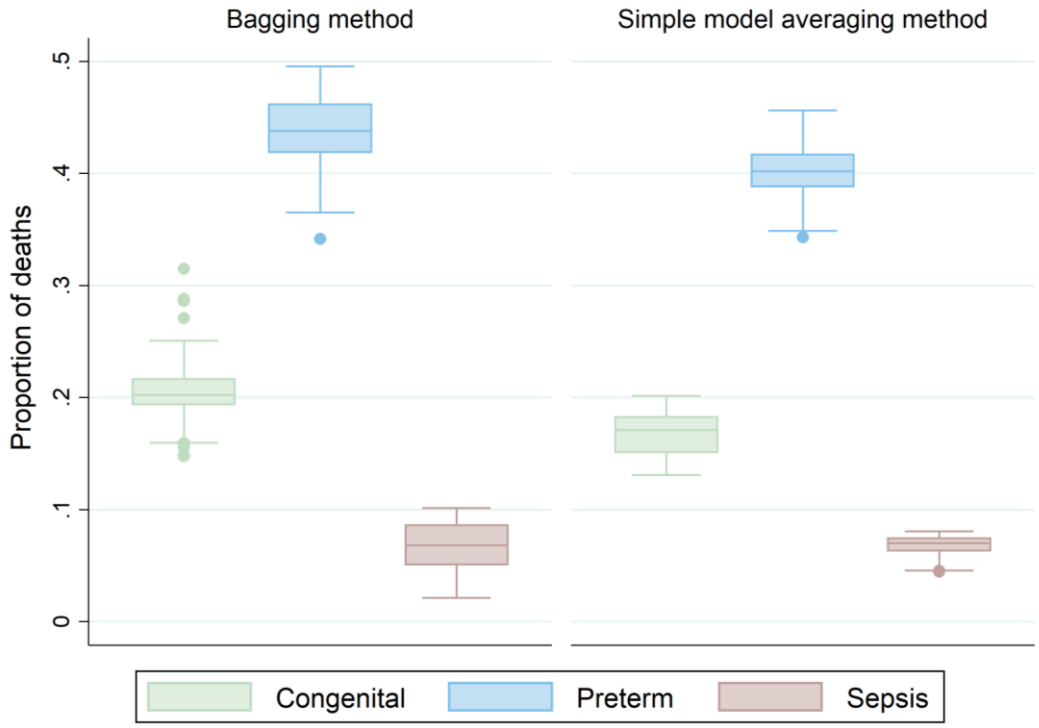


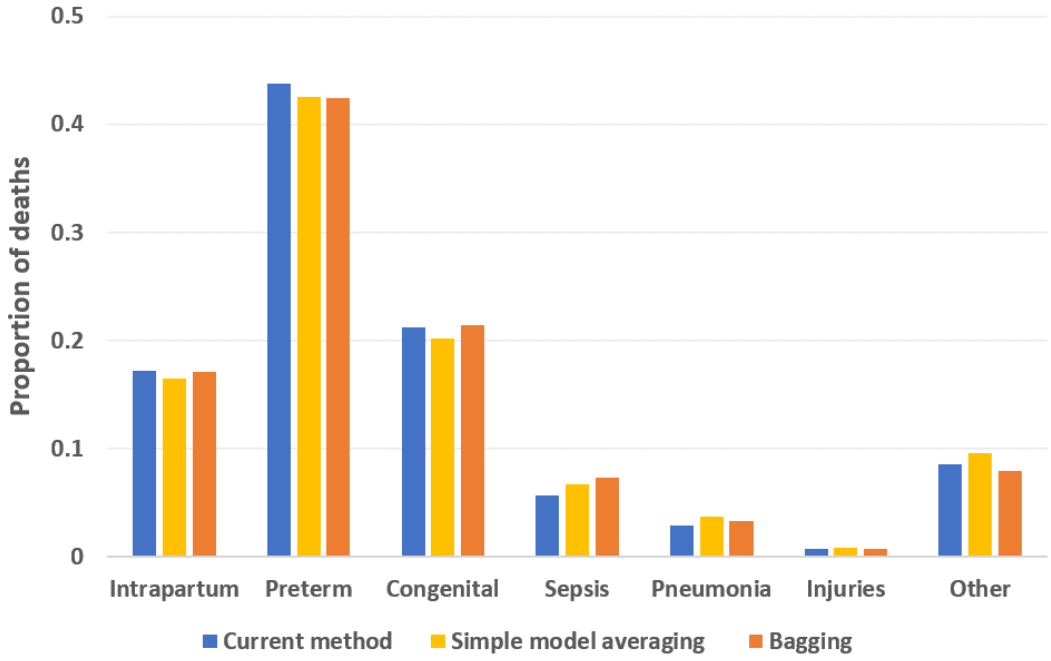
Figure 5.10 shows the bagging results alongside the simple model averaging results for Egypt (as shown in Figure 5.8) as a demonstration of how the two approaches may compare. In general, the median estimates are similar but not the same, and bagging appears to have wider ranges. This latter result is not surprising because bagging effectively creates different input data and re-runs the covariate selection process for each. If the input data are noisy, as we expect ours are, this would mean that the bagging iterations could have a relatively wide spread.

Figure 5.10: Example of bagging versus simple model averaging for three causes (Egypt, 2013).



These examples have demonstrated that country-specific differences can arise when using the different methods (current, simple model averaging, bagging). The aggregate results, however, appear to have less variation between these methods. Figure 5.11 shows the aggregated results across the low mortality model countries (n=47) for the early neonatal period. The results are remarkably similar across the three compared methods.

Figure 5.11: Comparison of the cause-of-death distribution by current, simple model averaging, and bagging methods for low mortality model countries (n=47)



These aggregated results are reassuring, and suggest that the current method is working well at the global level when compared with methods that better incorporate uncertainty in the input data or covariate selection method. However, our estimates need to be reliable at the country level. Since there are demonstrated differences at the country level (Figures 5.8, 5.10), a model averaging method could help improve robustness of the estimates. While simple model averaging and bagging appear to be promising, we need to assess their utility across the range of model improvement methods we are considering. This will also entail evaluating the computational feasibility of these model averaging approaches in the context of other modelling improvement choices (e.g. the modelling framework, covariate selection method). Decisions around model improvements are discussed in section 5.3.3.

Regularization methods

The bias-variance trade-off is at the heart of regularization methods. Increasing the complexity of a model (e.g. by adding covariates) tends to reduce bias because the model will be able to better fit more of the data points. But this increases the risk of overfitting, and typically results in higher variance. Simpler models usually have less variance (and thus more stability) since they cannot "move" around as much because they have fewer degrees of freedom. But such models have greater risk of underfitting. Thus, choosing a model involves carefully thinking through the interrelated issues of bias versus variance, overfitting versus underfitting, and

complexity versus simplicity in terms of model performance. Regularization methods try to balance GOF and complexity through consideration of these trade-offs.

Numerous approaches have been developed in recent decades that aim to balance model fit and complexity [171-173]. Regularization methods generally place a penalty on model complexity. Most of these methods focus on two ways in which instability is associated with covariates: 1) increasing the number of covariates increases model complexity and therefore the risk of model instability and 2) large coefficient values can increase instability. Implementing the penalty, and whether and how it can be tuned, depends on the specific method used. Some analytical methods, like the Akaike information criterion (AIC) and risk inflation criterion (RIC), use penalization mechanisms based on calculated scores to choose between models [171]. Several regression methods, on the other hand, balance fit and complexity through coefficient shrinkage [169, 174-176].

A subset of regularization methods exist that may also be able to help address the issue of using binomial covariate selection for multinomial models. These methods are efficient alternatives to the full *multinomial* covariate search that is computationally prohibitive using our current approach. Lasso and ridge regressions are examples of such regularization methods [169]. The general approach is as follows: 1) include all covariates in the model itself and 2) maximize the log likelihood minus a penalization/regularization term. This results in coefficient values shrinking. The lasso regression seems well suited to our modelling approach. Increasing the penalization term in this regression will push some covariates to zero, hence performing a form of covariate selection within the multinomial model itself.

We attempted to implement the lasso regression within our multinomial modelling framework. We did this by adding a penalty term to the likelihood function. The implementation appeared to work in terms of reduction of covariate coefficients towards zero as the lasso penalty (lambda) increased in value. However, we consistently ran into convergence problems, which we were unable to solve within Stata.

Some of these issues may be related to Stata's ability to handle maximum likelihood estimation for complex models, especially with user-written code. Lasso implementation is relatively new in Stata, with an official lasso package only released as part of Stata 16 in June 2019 [177]. The convergence issues we encountered may be solvable through more extensive re-writing of our multinomial modelling code, or by using a different software package (e.g. R). Alternatively, we

could attempt an implementation of a different regularization method. Since any of these changes would entail a more substantial effort, we decided that it would be prudent to do this after further decisions around model changes were made. This was especially the case given the possibility of moving to different modelling frameworks (e.g. binomial models, chapter 6; Bayesian models, chapter 7) which would change the lasso implementation approach.

In general, regularization methods appear to be a promising approach for adding stability to the models while potentially also solving the problem of multinomial covariate selection. They also have the advantage of generally being less computationally intensive than ensemble methods. However, the decision on whether to implement regularization methods (and the choice of which one) partly depends on which modelling changes are chosen as priorities to improve the COD models. These decisions are discussed in section 5.3.3, as well as chapters six and seven.

5.2 Weighting of empirical data for country-specific estimates

5.2.1 Introduction

Our current modelling strategy does not give additional weight to input data from a given country for its modelled estimates. Therefore, country-specific empirical data do not influence a country's modelled COD proportions any more than data from other countries. The reasoning for this was that the majority of studies in our input database were small and not nationally representative. Thus, it was not obvious that the national-level estimates for a country should be moved towards the data points from small non-national studies. However, an increasing number of countries now have data from nationally representative VA studies, including for neonatal deaths. How to use these nationally representative VA studies appropriately in our modelling process is an important question as we consider model improvements.

These nationally representative studies are not homogenous in their characteristics; they vary in size, study design, and various other factors that affect their quality and comparability. For example, many appear to use different hierarchies to assign causes of death, even across VA studies done as follow-ups to otherwise standardized DHS. Such differences in the nationally representative studies are not surprising given that they are conducted by a wide range of implementers and researchers, just like the smaller non-national studies. Although a few of these are ongoing (e.g. Million Deaths Study in India), most are one-off studies (e.g. a VA study follow-up to a DHS survey). Cross-sectional studies report COD distributions for only one or a few years instead of a longer time trend. For these reasons, and the inherent concerns with VA

data (section 2.3.2), we do not believe the COD distributions from these studies should be used "as-is" for the country's estimates in the way that we use high-quality VR data directly.

Currently, we include these nationally representative studies in the high mortality input database and treat them the same as all other studies. There is no mechanism within the model for such data to be given more weight. In the following subsections, I discuss the implications of these studies for our models and country-level random effects as a potential modelling solution.

5.2.2 How important is this for our models?

In some ways, the answer to how important this issue is for our models is a matter of perspective. For example, this issue looks less important in our current models than if we are forward thinking and consider contemporary data collection trends. There is also the (perhaps more philosophical) question about the role of modelling and how to balance modelling with imperfect empirical data. Here, I discuss each of these perspectives briefly.

On the practical side, inclusion of country-level random effects would not affect the current low mortality model estimates because there are no input data from the prediction countries in this model. The lack of such input data is because the prediction countries are those with VR data deemed to be of inadequate quality (see section 4.2.2 for more details). Lower-quality VR data from modelled countries could be considered for inclusion in the input dataset, ideally with a quality threshold to balance the desire for inclusion of relevant data with concerns about quality/reliability (this is discussed further in section 8.4).

For the high mortality model, 37 of 81 countries have studies included in the input database. Of these, eight have nationally representative data, accounting for 9% (11/124) of the data inputs in the high mortality model. These nationally representative studies make up 15% of deaths in the high mortality input database and 44% of global neonatal deaths occur in these countries. Thus, the nationally representative data points make up a small fraction of our dataset, but come from countries that make up a disproportionately large percentage of global neonatal deaths. Nationally representative VA studies are increasing, and thus it is likely that the fraction of our input data that will consist of such studies will increase in the future.

On a more philosophical level, it is important to keep in mind the ultimate goal of this work. Ideally, there would be no COD modelling because each country would have its own high-quality

CRVS system that continuously collects data at the local level. Cause-of-death modelling is a suboptimal alternative, but one of the few viable current options for many countries (see section 2.3.4 for more discussion on this topic). I have explained in section 5.2.1 why we do not use data from VA studies as-is. Yet, if nationally representative COD data are available, giving them more weight in the modelling process for that country's estimate is arguably the correct choice given our viewpoint on modelling. While we retain concerns about the quality of the VA data, we also have (different but parallel) concerns about our modelled estimates (as described in section 5.1). Thus, an approach that gives some additional weight to the nationally representative data from a country when deriving modelled estimates appears to be a reasonable solution.

5.2.3 Country-level random effects as a potential modelling solution

A mixed effects (ME) model with random country intercepts is a promising solution to address the issue of giving more weight to country-specific empirical data. An ME model has a hierarchical modelling structure that allows some parameters to not vary (i.e. the fixed effects [FE] component) while others are treated as random variables (i.e. the random effects [RE] component) [178]. An example of one equation of a multi-cause multinomial logistic regression with no random effect is shown in Equation 5.5.

$$\ln\left(\frac{P_{C|ij}}{P_{0|ij}}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kij} \qquad \text{(Eq. 5.5)}$$

The left-hand side is the log of the odds of the probability for cause C ($P_{C|ij}$) over the probability of the baseline cause ($P_{0|ij}$) for the ith prediction observation unit (e.g. year) of the jth group (e.g. country). On the right-hand side, β_k is the regression coefficient for the kth covariate (with β_0 as the intercept and K as the total number of covariates) and X_{kij} is the kth covariate for the ith prediction observation unit of the jth group.

The same regression with a random intercept is shown in Equation 5.6.

$$\ln\left(\frac{P_{C|ij}}{P_{0|ij}}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kij} + u_j \qquad \text{(Eq. 5.6)}$$

where u_j is the random effect representing the effect of the jth group.

In our case, the random effect would be at the country level, which is based on the assumption that there is variation between countries that is not picked up by covariates in the fixed effect part of the model. Introducing country-level random effects into our COD modelling strategy will result in a country's estimates being "pulled" towards that country's nationally representative VA data.

We attempted to incorporate random effects into our multinomial models in Stata using various methods but were largely unsuccessful. We investigated GLLAMM [179] and mixlogit [180] commands in Stata, but these had a number of issues, ranging from convergence problems to being time intensive to incompatibility with our models. The GSEM function in Stata 13 [181] appeared promising, and we implemented it using a simplified version of the low mortality model. We did this with no covariate selection and included only one covariate (NMR) in each equation. Over 24 hours were needed to run just one model with one covariate, making this computationally infeasible given our need for covariate selection and multiple (different) covariates in each equation of the multinomial. This method also does not account for redistribution of unreported causes, which we need for the high mortality model unless we drop studies with any unreported causes. This latter option is non-ideal, especially given the general dearth of input data in the high mortality model.

Mixed effects multinomial logistic models are relatively new for statistical software. It was only in 2017, after we had completed these analyses, that Stata released a mixed effects multinomial logit function [182]. The GSEM function in Stata which we described using above was introduced in 2013 [183], less than two years before we conducted this analysis. At the time we did this work, few packages existed across different statistical programmes for the type of analysis we were trying to conduct, and even fewer seemed to work well for complicated models like ours. For example, concerns were documented in forums and elsewhere about convergence and time intensiveness issues for complicated models with GSEM [181, 184, 185]. Thus, incorporating country-level random effects into our multinomial logistic regressions given our existing modelling strategy was not straightforward.

There are two plausible alternatives to consider for further investigating country-level random effects as part of our COD modelling strategy. The first is to replace the multinomial model with a set of binomial models and apply the random effects to the latter. The programming for mixed effects in a binomial model is more straightforward since it is a simpler model and is a better-established method in statistical software, including Stata [186]. We performed this analysis and have presented the results in the next chapter. The second option is to shift the COD models to a Bayesian framework, which is well-suited for random effects. In chapter 7, I present our results for a proof-of-concept Bayesian COD model with random effects.

5.3 Discussion

In this chapter, I have detailed three main issues that required further investigation as we consider ways to improve the existing neonatal COD models: predictive accuracy, model instability, and weighting a country's modelled estimates towards its empirical data. In this section, I discuss our findings and describe the decision-making process that took place to determine which model improvements we would prioritize for implementation.

5.3.1 Model performance issues

In this section, I described some model performance issues we wanted to further investigate, along with factors which may be causing them. I also introduced a few possible solutions which may improve model stability and predictive accuracy. Two of the key factors appeared to be noisy input data (both for cause-of-death distributions and covariates) and the strength of the relationships between causes and covariates. These are likely the greatest challenge for our models, and ones which we have little to no control over directly. If these problems are severe enough, then even models which technically can perform well would be unable to accurately estimate the outputs. But there are several results we have presented which are promising, and suggest that these issues are not severely impacting our models: 1) the country-specific results obtained by the existing models are within reasonable ranges (section 4.3); 2) the results from the existing approach are relatively similar to the model averaging methods, which are designed to account for uncertainty in modelling processes and input data – including due to noise; 3) building stability-enhancing processes (e.g. model averaging, regularization) into the code are ways to help mitigate some of these issues even at the country level; and 4) input data are generally increasing in quantity and quality, which can help overcome some of these issues over time as the input datasets continue to be updated.

I also presented various potential solutions which may help to improve model performance. In particular, ensemble methods like bagging and/or regularization methods like lasso regression are major modelling changes that could help to address some key concerns, which include uncertainty in our model selection, instability from model complexity, and the use of binomial covariate selection for multinomial models. We have also demonstrated proof-of-concept implementation of these to gauge feasibility. Overall, these preliminary analyses suggest that there is scope for such methods to be useful in our work to reduce instability. Selecting an appropriate and feasible method will depend on other model changes (e.g. covariate selection method). Thus, a more thorough investigation of these methods should occur after further

decisions are made on model improvement choices. Full implementation of these would be done after decisions around model changes are finalized, as the implementation approaches will be different depending on various choices (e.g. modelling framework).

Another potential solution to address the concern of using binomial covariate selection is to use a set of binomial models for the modelling as well. This would mean the covariate selection and modelling approaches are consistent. However, important drawbacks include cause-specific deaths not adding up to total deaths without scaling and the inability to adjust for unreported causes of death (as can be done in the multinomial model by re-writing the likelihood function). We implemented this alternative binomial model strategy and present the results for it in the following chapter.

Additional smaller considerations mentioned in section 5.1.3 are also important to catalogue and review when discussing model performance improvements. For covariates, testing for multicollinearity and regularly seeking updates on newly available (or updated) covariate time series are important. For COD data, lower-quality VR data from countries modelled in the low mortality model could be considered for inclusion in the input data. This would be particularly useful if the mixed effects models are implemented. For the GOF metric, weights could be added such that the weighting is intermediate between giving equal weight to each death and equal weight to each study or country-year in the input data (similar to what is done during the multinomial regression step). Alternate GOF metrics can also be identified and tested to see how they affect model performance, particularly through predictive accuracy.

In general, there are two broad threads of a modelling process which impact model performance: the input data and the statistical modelling processes. No matter how high-quality or appropriate one of these is, if the other is not, then the model risks having poor "performance". In our case, I would argue that there can be improvements made to both. On the modelling side, we can add components to our models which help mitigate some of the issues discussed in this section. These can include practical changes like stability-enhancing modelling strategies (e.g. ensemble methods and/or regularization), as well as some which also have philosophical underpinnings for our modelling strategy (e.g. random effects). It is also important to integrate internal validation and other relevant checks into the formal modelling strategy, instead of conducting them on an ad-hoc basis. I present recommendations related to designing modelling strategies in section 8.6.

While the input data are difficult for us to improve directly, there are positive signs that such data are generally becoming more plentiful and higher in quality. For example, an updated literature review conducted in Summer 2018 added 51 data points across 14 studies published between 2015 and 2018 to the high mortality dataset. This is a substantial increase over the number of studies added through similar literature review updates previously, suggesting that more studies are being conducted and published with neonatal COD outcomes. The size and quality of these studies is still heterogenous, but appear to be generally trending in a positive direction (potentially due to recent pushes like those for standardizing VA methods). Similarly, there are major efforts in place related to covariate measurement, including improving consistency of definitions and introducing more nuanced covariate measures (e.g. quality instead of just coverage indicators). These potential improvements have the real possibility of improving some of the issues we highlighted as being factors in our model performance.

In the last two decades, the field of metrics and burden estimation has grown dramatically. As noted in other parts of this thesis, in an ideal world we would not use models to produce estimates for neonatal COD distributions; instead, countries would extract such information from their own high-quality CRVS systems. But many countries still lack the infrastructure and systems required to produce such internal data. Until this gap is filled (through increasing implementation of CRVS systems or other innovative strategies), there remains value to modelling burden, especially when and where empirical data have been lacking.

However, it is critical that such estimates are produced using careful and thoughtful modelling approaches, and that the caveats around the input data and modelling are well understood. The very reason models are still needed (i.e. inadequate high-quality empirical data) is also part of the reason why such modelling faces issues around instability and accuracy. Thus, understanding the sources of potential model performance issues, and building in robustness to mitigate these issues should be an essential part of such estimation work. I attempted to highlight some of the issues which may affect our models in this section. I have also included a broader discussion on modelling topics in chapter 8, including recommendations on the uses and limits of such models.

5.3.2 Weighting of empirical data for country-specific estimates

In this section, I summarized the practical and philosophical reasons for weighting the modelled estimates towards a country's own empirical data. There is a compelling case to be made for including country-level random effects in our models, both because of the increasing number

of nationally representative studies and because it honours the goal of moving towards using empirical instead of modelled data. The addition of such random effects is not trivial, as discussed in section 5.2.3 and as will be discussed in section 6.4 for binomial models. Thus, if it is deemed a priority in the near to medium future, it is likely wiser to implement it alongside other improvements, especially since its inclusion will partly dictate which other modelling changes are feasible.

An important point to consider about the inclusion of such random effects is what they truly represent given the quality issues with the real-world data which are included in our input datasets. How much, for example, are the random effects capturing real COD distribution variation between countries versus measurement error/methodology differences between input studies. Understanding the nuances around this point is important for interpreting the mixed effect models. A more detailed discussion of what random effects mean for models such as ours is included in section 7.4.

5.3.3 Decision-making process for model improvements

In general, many of the model improvement strategies discussed above are interlinked or dependent on other modelling decisions. For example, implementation of the lasso regression is different in the existing multinomial framework versus the Bayesian framework. Thus, we chose to investigate these modelling issues to the point of allowing us to understand the problems and potential solutions, but did not pursue an exhaustive search of each option. In this section, I describe the larger decision-making process around choosing specific model improvement priorities.

The work described in this and the following chapter (on binomial COD models) was presented as a series of five talks at the 2016 annual Maternal Child Epidemiology Estimation (MCEE) meeting (see section 2.3.4 for more information about MCEE). This meeting was attended by experts on issues related to child causes of death, both for neonates and older ages. The goal of the presentations was to obtain feedback on our work, and to choose a set of model improvement priorities to implement and finalize for the neonatal and 1-59 month COD models. Individuals at the meeting included academic researchers from Johns Hopkins, the London School of Hygiene and Tropical Medicine, and the University of Edinburgh. Representatives from the WHO (which officially publishes our neonatal COD estimates) and the Bill and Melinda Gates Foundation (which funds MCEE) were also represented. As such, the decisions around

model improvements were not purely academic but were also driven by policy and donor priorities.

Based on our presentations and the ensuing discussions, the following decisions were made: 1) inclusion of country-level random effects was a key priority, and one that should be implemented early as it would have an effect on which other additional strategies were feasible; 2) a multinomial Bayesian modelling framework should be tested given the suboptimal results of the binomial models (discussed in the next chapter) and the difficulty of implementing a mixed effects model in the current frequentist framework, and 3) a regularization approach using multinomial covariate selection (e.g. lasso) looked like a promising solution to investigate in the Bayesian framework.

In the following chapters, I present results for alternative COD modelling strategies: a set of binomial models instead of a multinomial model (chapter 6) and multinomial models with mixed effects and lasso in the Bayesian framework (chapter 7).

6 Alternative modelling approach: set of binomial models

Although we use multinomial models for our current neonatal cause-of-death (COD) estimation approach, the challenges described in chapter 5 led us to investigate using a set of binomial models instead. In particular, we believed that 1) predictive accuracy may improve since our current covariate selection approach relies on binomial models and 2) adding country-level random effects may be easier in the binomial framework. In this chapter, I describe our work on developing a modelling framework using a set of binomial regressions for the neonatal COD estimates, including a comparison with the multinomial results and the addition of country-level random effects. While uncertainty intervals are an essential component of final estimates, I have not included them in this chapter since the focus was not to develop publishable estimates, but rather to determine whether the binomial framework could help to resolve some of the aforementioned challenges.

6.1 Introduction

Choosing an appropriate statistical model is important for theoretical and practical reasons. Factors that contribute to the choice of model range from understanding which probability distribution best reflects the underlying data generation mechanism (theoretical) to the computational sophistication and intensity required to run different models (practical).

For our work, we originally chose to use the multinomial distribution – where the causes of death are fit simultaneously – for a number of reasons. First, the underlying data generating mechanism for a COD distribution is more accurately reflected by the multinomial distribution when there are multiple causes of death. Additionally, modelling the proportional cause distribution results in the sum of deaths across the causes to equal 100% of deaths. In contrast, a scaling factor (typically chosen somewhat arbitrarily) is required when using a set of binomial models so that the summed deaths across causes can equal the total number of deaths. Additionally, analysing a set of binomial models is less efficient and produces larger standard errors than a multinomial model [187]. Finally, if a specific cause is not reported in a study, the multinomial model can be used to incorporate uncertainty about whether any deaths for that cause occurred. This is not possible when using a set of binomial models.

However, there are important practical reasons for exploring the set of binomial models instead. First, our covariate selection step is based on binomial models (section 5.1.3). This is because multinomial covariate selection is computationally prohibitive with our current covariate

selection approach. As discussed in section 5.1.3, even just the full search algorithm using the binomial models is computationally intensive. However, covariates chosen based on out-of-sample goodness-of-fit (GOF) of binomial models may not result in the best out-of-sample GOF in a multinomial model. Given the challenges of multinomial covariate selection, we considered the option of using a set of binomial logistic regressions instead of the multinomial regression for the neonatal COD model. Another important reason for investigating the use of binomial models is that programming is simpler for binomial versus multinomial models, which in turn means that implementing some techniques like country-level random effects should be simpler in the binomial models. As discussed in section 5.3.3, incorporating country-level random effects was deemed a priority for future COD estimation, but we faced challenges doing this in the multinomial model (section 5.2.3). Finally, it is plausible that the covariate-cause relationships may be stronger when modelling the odds for a cause against all other causes (as we would do in a set of binomial models) instead of against a baseline cause (as is done in the multinomial models).

For these reasons, we chose to compare our multinomial logistic regression approach for estimating the neonatal COD distribution to one that used a set of binomial logistic models instead. In this chapter, I describe the methods we used to build the binomial estimation framework, how the resulting estimates compared to our previous estimation results and the incorporation of random effects into this binomial framework. I also include a discussion on the advantages and disadvantages of this new modelling framework.

6.2 Methods

6.2.1 Input and prediction data

We used the same input data and covariates for prediction as used for the multinomial models described in section 5.1.2. We assessed the relationship between the covariates and causes of death in this input data (similarly to what was done for the multinomial models; section 5.1.3). Here, we graphed each covariate against the respective outcome for each cause (e.g. female literacy versus log(other/(N-other)) where N is number of all-cause deaths.

6.2.2 Statistical modelling

All statistical analyses were undertaken using Stata version 12 (www.stata.com). We used the same modelling approach described in section 4.2, with a few changes. First, in the covariate

selection, the baseline cause for the logistic regressions was now all causes except the one being estimated. For example, if the cause being estimated was pneumonia, then the baseline "cause" was the sum of deaths across all non-pneumonia deaths. In contrast, the baseline cause in the multinomial model was set to preterm (low mortality model) and intrapartum (high mortality model). We re-ran the covariate selection process using this approach, and selected the covariates based on best out-of-sample GOF. Next, the multinomial model was replaced by a set of binomial logistic regression models using the blogit command in Stata. We used the modelled coefficients from the cause-specific regressions to predict the number of deaths for each cause. Unlike the multinomial model, the causes are not all fit simultaneously with this method and it is not possible to re-write the COD distribution to account for unreported causes (see section 4.2.3 for how the multinomial model is adapted to account for unreported causes). Therefore, studies with one or more unreported causes were dropped when fitting the regression for that cause.

6.2.3 Assessing predictive accuracy

To assess predictive accuracy, we used the same method as described in section 5.1.2. Briefly, we applied the coefficients from the binomial model results to the input data to assess the differences between the observed and predicted data. We used the average chi-squared statistic ($\overline{\chi^2}$) as defined in Equation 5.3 (section 5.1.1) to assess the predictive accuracy, allowing us compare models with different numbers of observations. This is useful since studies with missing causes were dropped for the relevant binomial models (section 6.2.2) and therefore the binomial and multinomial models may have different numbers of observations.

6.2.4 Estimation of cause-specific numbers and proportions of deaths

We had to adjust the modelled death estimates because a set of binomial models does not automatically result in the deaths across causes to sum to the total number of all-cause deaths. We chose a proportional scaling factor for this. To do this, we calculated the ratio of the summed cause-specific deaths to the total number of all-cause deaths. We then divided each cause-specific death estimate by this ratio. We calculated scaled cause-specific proportions by dividing the scaled cause-specific deaths by the total number of all-cause deaths. We calculated un-scaled cause-specific proportions by dividing the un-scaled cause-specific deaths by the un-scaled total number of deaths estimated by the binomial models.

6.2.5 Incorporating country-level random effects

We implemented a mixed effects (ME) modelling structure within the binomial COD models using the xtmelogit command in Stata 12. We then performed both validation and estimation analyses. For the validation exercise, we assumed that the random effect was country-level for the low mortality model and study-level for the high mortality model. The analyses were similar to those described in sections 6.2.2-6.2.4, with the following exceptions. For the covariate selection process, we used only the fixed effects components of the ME models. For estimation (including validation), we produced two sets of estimates: 1) using the full ME model (i.e. fixed and random effects components) and 2) using only the fixed effects component of the ME model. This allows us to gauge how much the change in results is due to the random effects component versus different coefficient values for the covariates (i.e. the fixed effect component). For this exercise, we assumed that the random effects are applicable to all input data (i.e. not only to nationally representative studies).

I present the results for the binomial ME analysis as part of a comparison with the previous non-ME models. I will use the following terminology to refer to the models:

- Multinomial – the non-ME multinomial models where all causes are fit simultaneously (described in section 4.2.3)
- Binomial – the set of non-ME binomial models (described in section 6.2.2)
- Binomial FE – the set of binomial ME regressions, but using only the fixed effects component for predictions (described in this section)
- Binomial RE – the set of binomial ME regressions, using both the fixed and random effects components for predictions (described in this section)

6.3 Main results

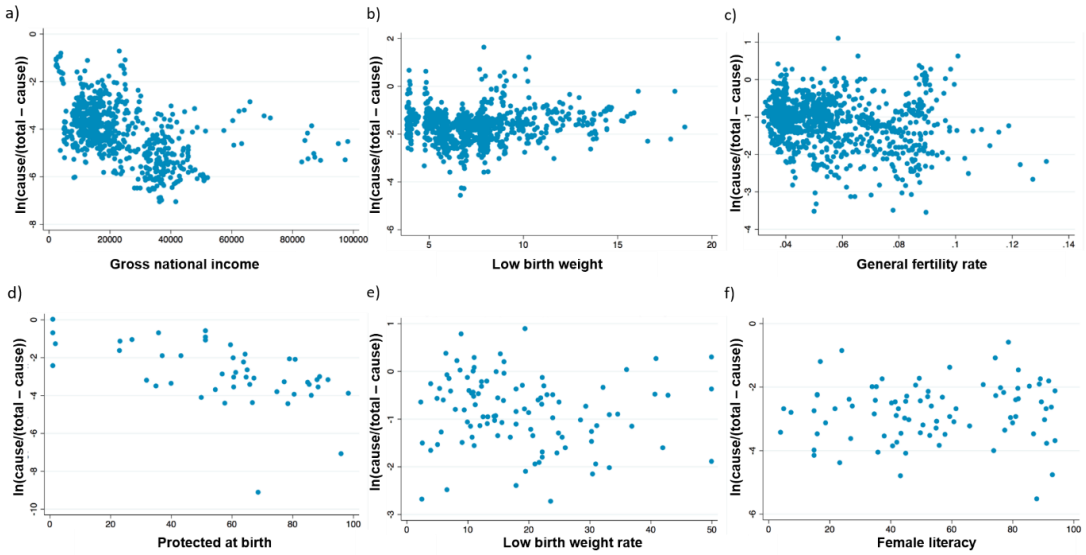
In this subsection, I present the key results from the binomial COD analyses, including for validation, COD estimation, and inclusion of country-level random effects. Since the main reason for conducting the binomial analyses is to consider using binomial instead of multinomial models for the COD estimation, I have included comparisons with the multinomial models where relevant.

6.3.1 Relationship between covariates and causes

Figure 6.1 includes representative examples of the covariate-cause relationships in our data in the low mortality model (top row) and high mortality model (bottom row). There were a few

covariate-cause pairs which seemed to visually have a strong association. For example, both pneumonia/GNI (Figure 6.1a) and tetanus/PAB (Figure 6.1d) appeared to have decreasing proportional burden as the covariate value increased, which is to be expected. However, most of covariate-cause relationships appeared to lack strong associations (e.g. Figures 6.1b-c and 6.1e-f). This included instances for which we would expect to have seen a relationship, such as intrapartum/LBW (Figure 6.1b) and preterm/LBW (Figure 6.1e).

Figure 6.1: Examples of covariate-cause relationships in the low (top row) and high (bottom row) mortality models for the early neonatal period. a) pneumonia and GNI; b) intrapartum and LBW; c) congenital and GFR; d) tetanus and PAB; e) preterm and LBW; and f) other and female literacy.



These results are similar to the multinomial results presented in section 5.1.3. Overall, these analyses indicate relatively weak relationships between many of the cause-covariate pairs.

6.3.2 Validation results

The average chi-squared GOF results in Table 6.1 suggest that the binomial and multinomial models performed similarly for the low mortality model. In contrast, the binomial models appeared to generally perform better for the high mortality model (i.e. lower $\overline{\chi^2}$). Note that the binomial results presented here are based on a baseline in the regression of all deaths other than those from the given cause (unlike the multinomial baselines of preterm and intrapartum), and these models are run for all causes (i.e. no missing "baseline" cause).

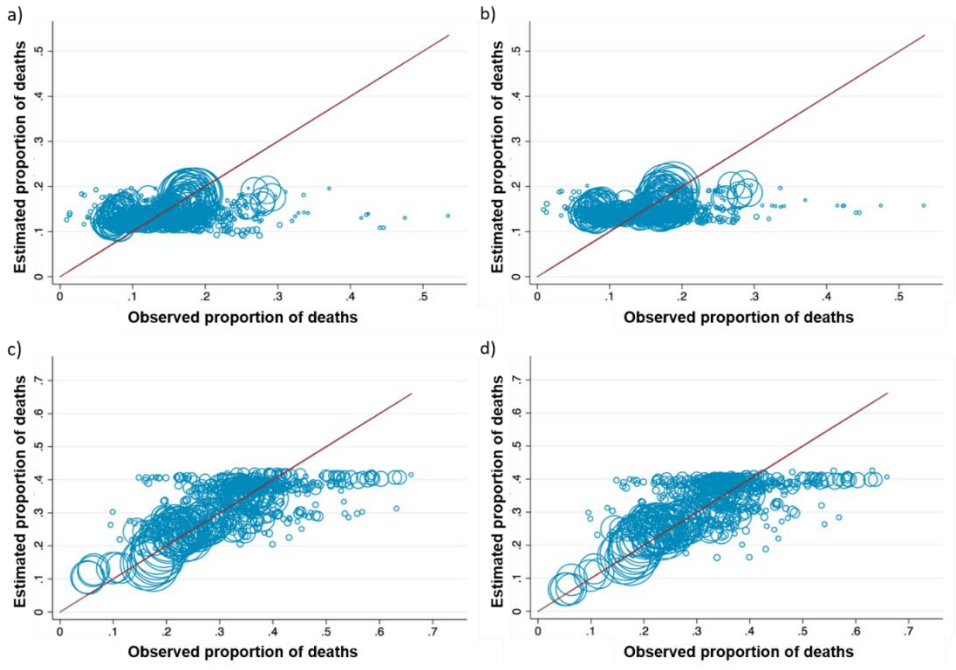
Table 6.1: Comparison of the average chi-squared out-of-sample goodness-of-fit statistic for the binomial versus multinomial models by cause and period for the low and high mortality models.

	Low mortality model				High mortality model			
	Early period		Late period		Early period		Late period	
	Binom[1]	Multi[2]	Binom	Multi	Binom	Multi[3]	Binom	Multi
Intrapartum	14	11	5	5	20	22	14	11
Congenital	14	13	4	4	5	28	3	3
Preterm	13	15	7	8	10	24	10	10
Sepsis	7	8	6	7	11	49	11	25
Pneumonia	11	15	11	9	4	11	7	7
Injuries	3	4	4	3	---	---	---	---
Diarrhoea	---	---	---	---	6	7	4	5
Tetanus	---	---	---	---	15	44	8	10
Other	19	24	6	6	11	67	9	16

[1] Binom = binomial; [2] Multi = multinomial; [3] based on the "tweaked" sepsis model described in section 5.1.3.

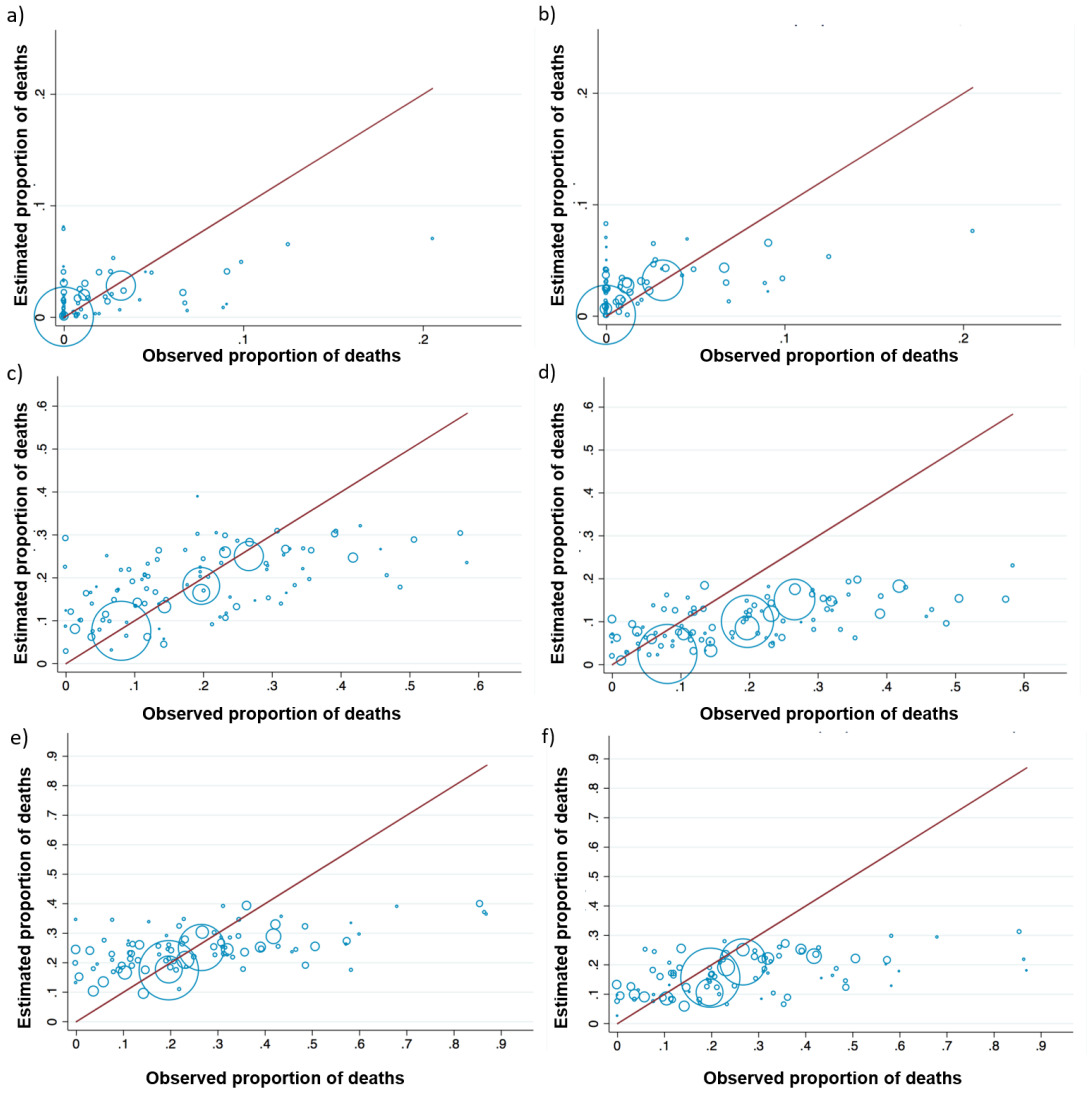
The similarity in predictive accuracy between the binomial and multinomial low mortality models can be seen visually as well (Figure 6.2). The two examples included in this figure (intrapartum/early [Figures 6.2a-b] and congenital/late [Figures 6.2c-d]) are strikingly similar between the binomial and multinomial models, and this type of similarity is generally true across the low mortality model causes.

Figure 6.2: Comparison of binomial (scaled) and multinomial validation results in the low mortality model. a) intrapartum/early (binomial); b) intrapartum/early (multinomial); c) congenital/late (binomial); d) congenital/late (multinomial). Note: circles are weighted by total neonatal deaths for given country-year compared to all other country-years.



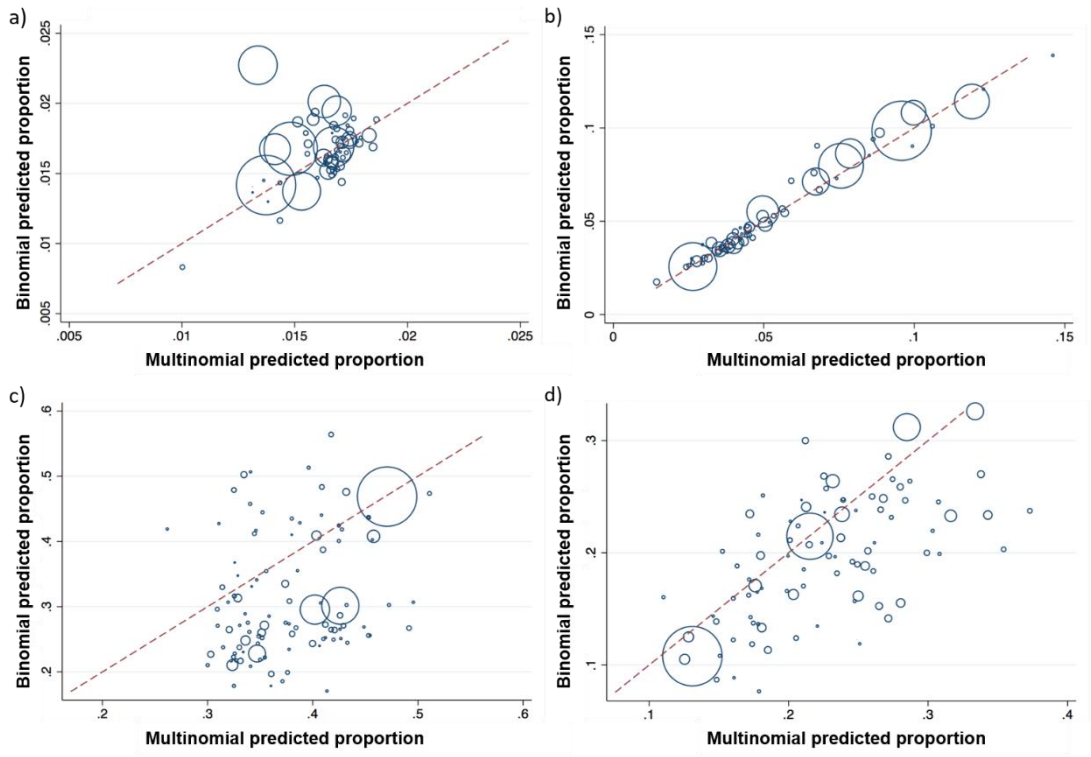
The validation graphs for the high mortality model, however, tell a more complicated story. For several causes, the binomial and multinomial models perform similarly (e.g. Figures 6.3a-b), while for some causes the binomial models appear to be performing better (e.g. Figure 6.3c-d). A few, however, do not appear to have substantial improvements with the binomial model, even though their $\overline{\chi^2}$ values in Table 6.1 are dissimilar. For example, sepsis (early period) has $\overline{\chi^2}$ values of 11 and 49 for the binomial and multinomial models, respectively. However, the binomial improvement appears less stark visually and in fact may mostly be limited to the larger studies (Figures 6.3e-f).

Figure 6.3: Comparison of binomial (scaled) and multinomial validation results in the high mortality model. a) diarrhoea/early (binomial); b) diarrhoea/early (multinomial); c) other/early (binomial); d) other/early (multinomial); e) sepsis/early (binomial); f) sepsis/early (multinomial). Note: circles are weighted by total neonatal deaths for given study compared to all other studies.



When we compared the predicted COD estimates based on the validation exercise (i.e. model used to predict on the input data) between the binomial and multinomial models, we found a wide range of agreement levels. The low mortality model generally had moderate (e.g. Figure 6.4a) to good agreement (e.g. Figure 6.4b). The high mortality model had lower levels of agreement, but also ranged from lower (e.g. Figure 6.4c) to higher (e.g. Figure 6.4d) agreement between the binomial and multinomial validation predictions. These results were consistent before and after scaling.

Figure 6.4: Direct comparison of binomial and multinomial validation estimates. a) injuries/late, low mortality model; b) sepsis/early, low mortality model; c) preterm/early, high mortality model; d) intrapartum/late, high mortality model. Note: circles are weighted by total neonatal deaths for given country-year (or study) compared to all other country-years (or studies).



In general, the binomial and multinomial estimates appeared to produce similar results for many of the causes, especially in the low mortality model. For some causes, the binomial models appeared to be performing better based on the $\overline{\chi^2}$ metric. However, the visual evidence was not as compelling for the binomial improvement. The apparent discrepancy between the $\overline{\chi^2}$ results and the visual comparison is due to the $\overline{\chi^2}$ statistic prioritizing the fit of large studies (as discussed in section 5.1.3). The high mortality model results also highlight that we should investigate multinomial covariate selection if we ultimately choose multinomial

models for the COD estimation. Overall, even when the binomial model appeared to have had better out-of-sample GOF than the multinomial, this GOF was not necessarily greatly improved.

6.3.3 Estimation results

Scaling to fit total number of deaths

The amount of scaling required to fit the binomial death estimates into the envelope of total deaths was relatively minor for the low mortality model (Table 6.2). The mean ratio of summed binomial deaths to the envelope was 1.01 across country-years, suggesting on average these binomial models overestimated the number of deaths by 1%. The high mortality model predictions, on the other hand, were closer to 10% over the total envelope. However, the amount of scaling needed was more severe for some country-years, as suggested by the scaling ratio range (Table 6.2).

Table 6.2: Amount of scaling needed to fit predicted binomial death estimates into envelope of total number of deaths using proportional scaling

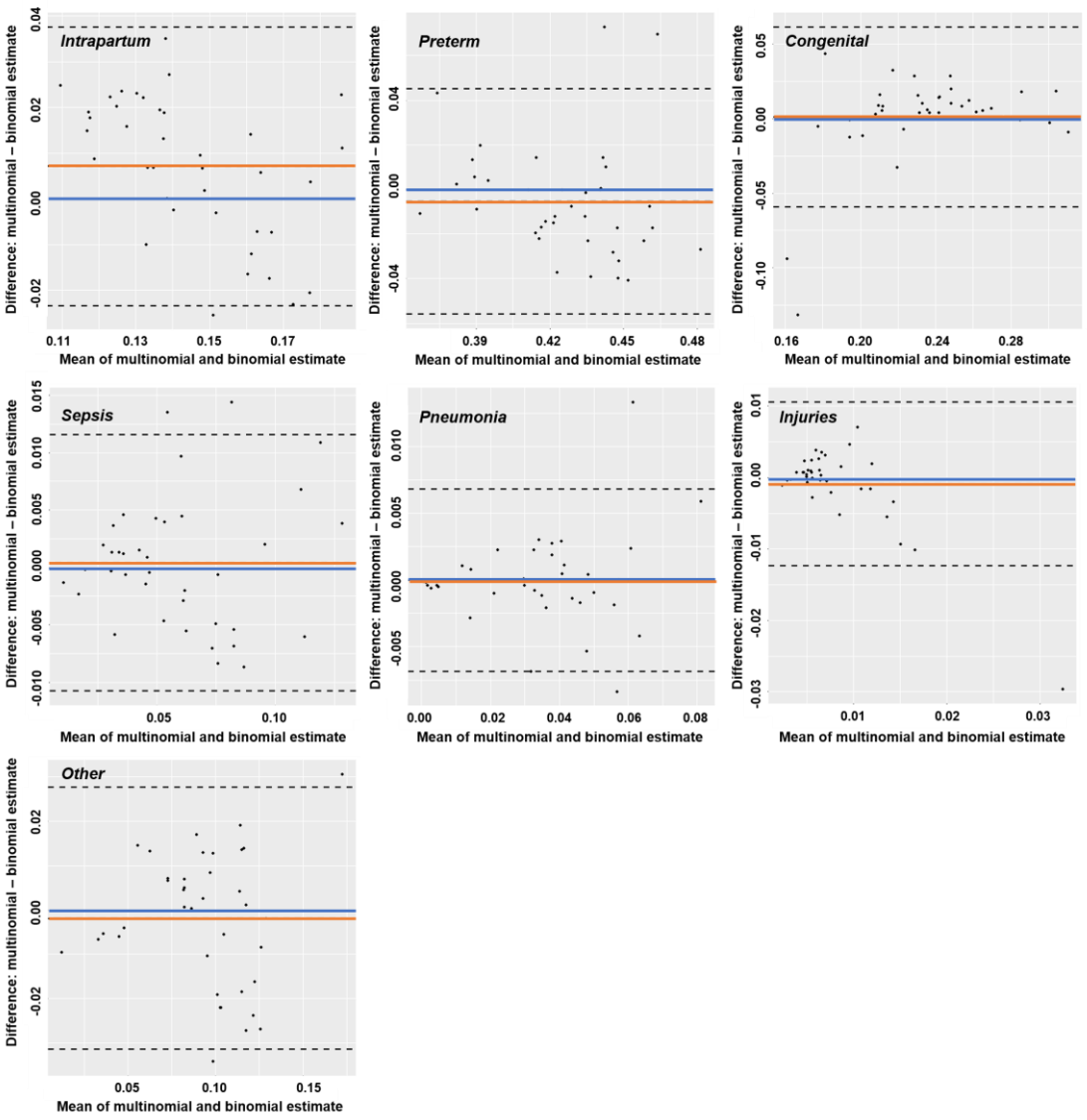
	Mean ratio[1] (SD)[2]	Median ratio (IQR)[3]	Ratio range
High mortality model			
Early period	1.08 (0.09)	1.06 (1.02-1.20)	0.84-1.32
Late period	1.11 (0.10)	1.09 (1.04-1.14)	0.90-1.44
Low mortality model			
Early period	1.01 (0.06)	1.01 (0.97-1.05)	0.84-1.22
Late period	1.01 (0.04)	1.01 (0.99-1.04)	0.78-1.10

[1] ratio refers to scaling ratio as described in section 6.2.3; [1] SD = standard deviation; [1] IQR = interquartile range

Binomial cause-of-death distribution estimates

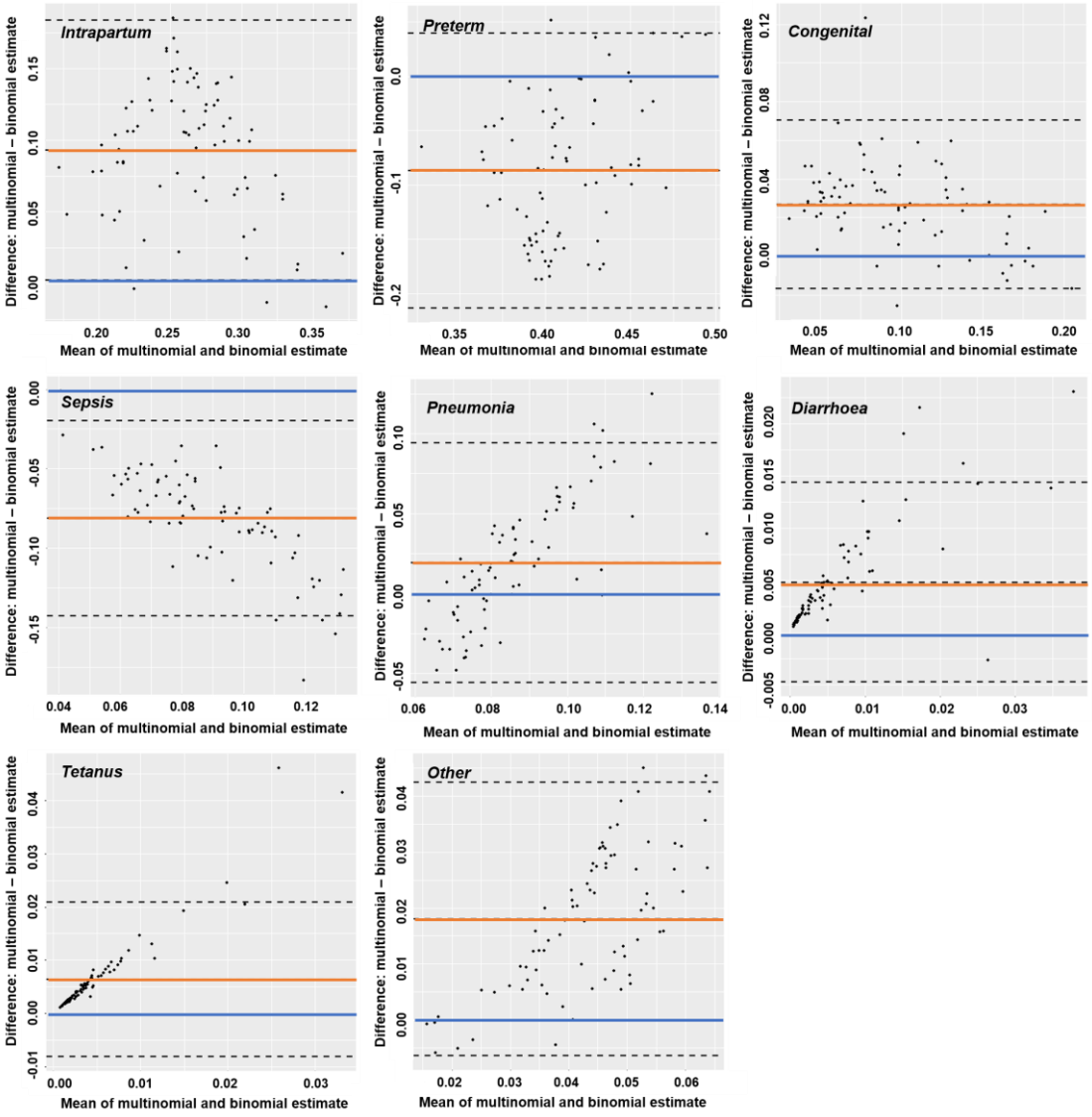
The predicted COD distributions obtained using the binomial and multinomial low mortality models were relatively similar (Figure 6.5). The largest differences were seen for intrapartum, but these were still relatively minor. The only cause that appeared to have some systematic bias was congenital, which tended to have higher estimates in the multinomial model. But even this was relatively minor, with most of the differences within three percentage points.

Figure 6.5: Comparison of binomial and multinomial estimates for all causes of the low mortality model (early period). Note: orange line = mean difference between multinomial and binomial models, blue line = no difference (i.e. 0); top and bottom black dashed lines are +/- 1.96 SDs from the orange line.



The high mortality model had larger differences (Figure 6.6). The sepsis and preterm proportions were generally estimated to be higher in the early period binomial models than the multinomial model. In contrast, the remaining causes tended to have higher estimates in the early period multinomial model. The proportions of some causes appeared less variable in the binomial compared to multinomial (e.g. pneumonia, tetanus, other), while sepsis had more variation in the binomial. Some systematic patterns also appeared to be present. The multinomial predictions had higher variation for pneumonia, diarrhoea, tetanus, and other as the mean between the two models increased. The opposite was true for sepsis; the binomial model predictions had more variation as the mean increased.

Figure 6.6: Comparison of binomial and multinomial estimates for all causes of the high mortality model (early period). Note: orange line = mean difference between multinomial and binomial models, blue line = no difference (i.e. 0); top and bottom black dashed lines are +/- 1.96 SDs from the orange line.



A comparison of the proportional COD estimates for 2015 are shown in Table 6.3. Similar to the analyses above, these suggest that the binomial and multinomial predictions are similar for the low mortality model while the high mortality model estimates have more substantial differences for some causes (e.g. sepsis).

Table 6.3: Comparison of estimated cause-of-death proportions in 2015 for modelled countries using the binomial versus multinomial models

	Low mortality model				High mortality model			
	Early period		Late period		Early period		Late period	
	Binom[1]	Multi[2]	Binom	Multi	Binom	Multi	Binom	Multi
Intrapartum	13.6	14.4	7.7	8.0	24.9	31.0	16.8	21.0
Congenital	22.6	23.3	29.6	31.2	6.4	9.3	6.2	6.0
Preterm	43.2	42.2	29.4	27.6	43.2	37.2	23.3	27.9
Sepsis	5.0	5.0	14.8	14.8	14.2	5.5	37.9	24.1
Pneumonia	3.9	3.8	7.5	7.4	7.4	10.8	6.7	7.9
Injuries	0.5	0.7	1.3	1.5	---	---	---	---
Diarrhoea	---	---	---	---	0.8	1.2	2.1	1.8
Tetanus	---	---	---	---	0.2	1.1	2.0	3.0
Other	11.2	10.4	9.6	9.4	3.0	4.1	5.0	8.3

[1] Binom = binomial; [2] Multi = multinomial

6.3.4 Country-level random effects results

Covariate selection

Table 6.4 includes a comparison of the covariates chosen by the covariate selection process in the binomial and binomial FE models (since the random effects component was not used for covariate selection). Covariates found to be common between the two are italicized in the table. The selected covariates vary quite a bit between the two model types, with several causes having no covariate overlap between the two. A likely contributing factor to this difference is that the non-ME binomial models have no weighting (i.e. each death is weighted equally) whereas the ME models, which take into account the non-independence of deaths within studies, essentially down-weight deaths in large studies.

The % reduction from null is not clearly better or worse for the binomial versus binomial FE model. For example, the binomial FE model had a higher % reduction from null for 56% (9/16) of causes (early and late period) in the high mortality model, while the binomial model without random effects had a higher % reduction for 64% (9/14) of causes in the low mortality model (Table 6.4). Some causes had substantial differences in the % reduction from null. For example, sepsis in the early period low mortality model had 68.3% reduction from null in the binomial model versus 0.8% in the binomial FE model. Pneumonia in the early period high mortality model had 7.3% reduction from null in the binomial model compared to 49.3% in the binomial FE model. Several causes had less severe, but still large, differences (Table 6.4). It is possible that part of this difference could be resolved using the full search algorithm (see section 5.1.3). For example, the sepsis binomial FE model with 0.8% reduction had only DPT in the model while the simple binomial model had five covariates, none of which were DPT. The full instead of partial search algorithm may have been able to find a better fitting set of covariates that did not

involve DPT. However, a full search algorithm would be very computationally intensive with the ME model, and at least some of the models are likely to encounter convergence difficulties. Thus, this may not be a plausible alternative.

Table 6.4: Selected covariates in cause equations and % reduction from null for the binomial versus binomial FE models (i.e. binomial ME model using only fixed effects component)

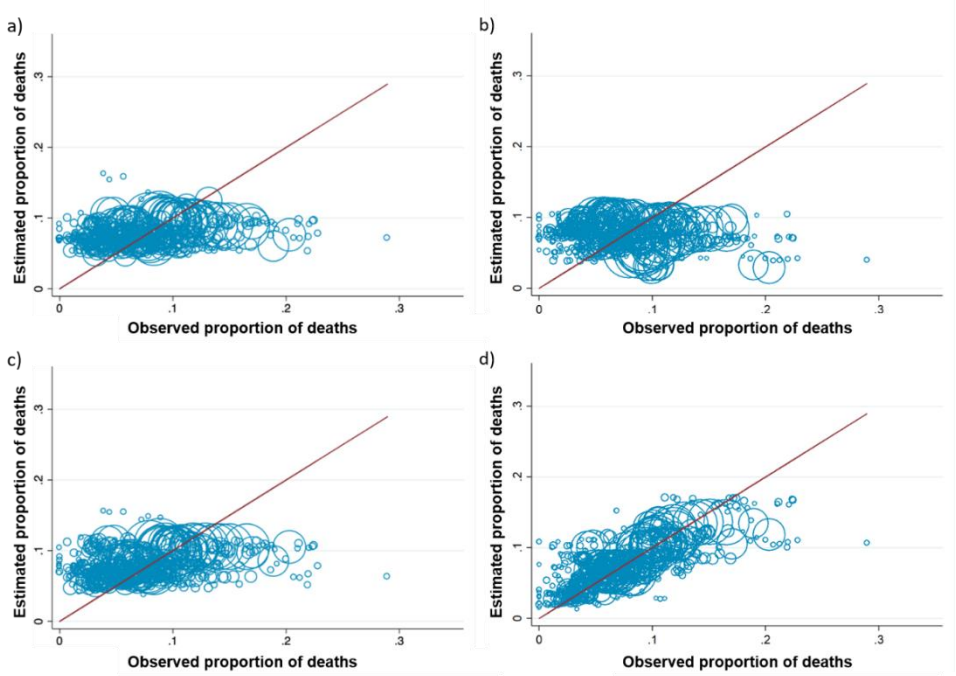
	Binomial model		Binomial FE model	
	Best model[1]	**% red.[2]**	**Best model**	**% red.**
High mortality model; early period				
Intrapartum	*PAB^s*	50.7	*PAB^s*, GFR, LBW, PvLBW, period, SBA^q, NMR^q	68.5
Preterm	*period*, ANC^s, LBW, *PvLBW*	81.1	SBA^q, *period*, PAB, IMR, SA, *PvLBW*	84.3
Congenital	U5MR, *NMR^q*, FLR^q, *PvLBW*, BCG^s, *SSA*, IMR	85.7	*NMR^s*, DPT^s, *PvLBW*, *SSA*	71.8
Sepsis	*period*, ANC^s, BCG, FLR, D*PT*	81.9	PAB, *DPT*, *period*, GFR^q, IMR^q	76.7
Pneumonia	*SBA^s, BCG*	7.3	*BCG^q, SBA*, GFR^s	49.3
Diarrhoea	NMR^q, SSA, U5MR, BCG	87.1	DPT^q, LBW^s, GFR, PAB	54.8
Tetanus	NMR, period, BCG	93.1	PAB	74.1
Other	SSA, ANC	81.7	PAB^q, period, LBW^q, NMR, SBA^q, DPT, BCG, U5MR	91.2
High mortality model; late period				
Intrapartum	*GFR^s, PAB^s*	52.7	*PAB^s, GFR^s,* SSA	51.5
Preterm	LBW^s, *BCG*, SA	18.1	DPT^s, PAB, SBA^q, *BCG*	29.6
Congenital	U5MR^q, SBA^s, *PvLBW*	94.9	NMR^q, DPT^s, *PvLBW*	95.1
Sepsis	NMR^q, FLR^q, period, SSA	37.5	IMR, SA	33.4
Pneumonia	DPT^q	5.2	GFR, SA	8.1
Diarrhoea	*LBW^s, U5MR^q, period*, IMR, GFR	24.0	ANC, DPT, *LBW^s, U5MR^q,* NMR, SSA, *period*	56.3
Tetanus	U5MR, *ANC*, BCG, LBW^q	91.0	*ANC*, PAB, PvLBW, IMR^q, period	90.8
Other	ANC, period, SSA	25.3	PAB^q, DPT, NMR, IMR, LBW^q, BCG^s, SBA^q, SA	47.2
Low mortality model; early period				
Intrapartum	*FLR^s*	38.3	*FLR^d*, U5MR, ANC^d	18.6
Preterm	*NMR^s*	4.8	*NMR^s*, U5MR, DPT, IMR	12.9
Congenital	*NMR*, GINI^d	65.7	*NMR*	61.9
Sepsis	GNI, GINI, IMR^s, ANC^s, NMR^d	68.3	DPT	0.8
Pneumonia	*GNI*, GINI, U5MR	36.4	*GNI*	26.1
Injuries	GFR^s, ANC, GNI	28.2	---	0.0
Other	GFR^s, *U5MR*, FLR	23.0	DPT, *U5MR*, GINI, ANC, IMR	29.7
High mortality model; late period				
Intrapartum	DPT^d	3.0	FLR^d	8.4
Preterm	FLR^d, GINI	25.7	DPT, GNI, NMR, U5MR, ANC	18.5
Congenital	*IMR*, GINI	67.5	*IMR*, GFR^d, FLR^s, NMR, DPT^s	63.0
Sepsis	*GINI*, IMR^d, FLR, *GNI*, *ANC^s*	73.8	*GINI, GNI*, NMR, GFR, *ANC*	80.0
Pneumonia	ANC^s	24.4	IMR, NMR, GFR^s	48.7
Injuries	FLR, DPT	3.1	LBW	0.2
Other	GFR^s, FLR	28.3	GINI^d, DPT, FLR, LBW	26.6

[1] covariates common between both models are italicized, see Table 4.2 for covariate acronym definitions; [2] % red. = % reduction from null; ^{s} spline instead of linear; ^{q} quadratic instead of linear

Predictive accuracy

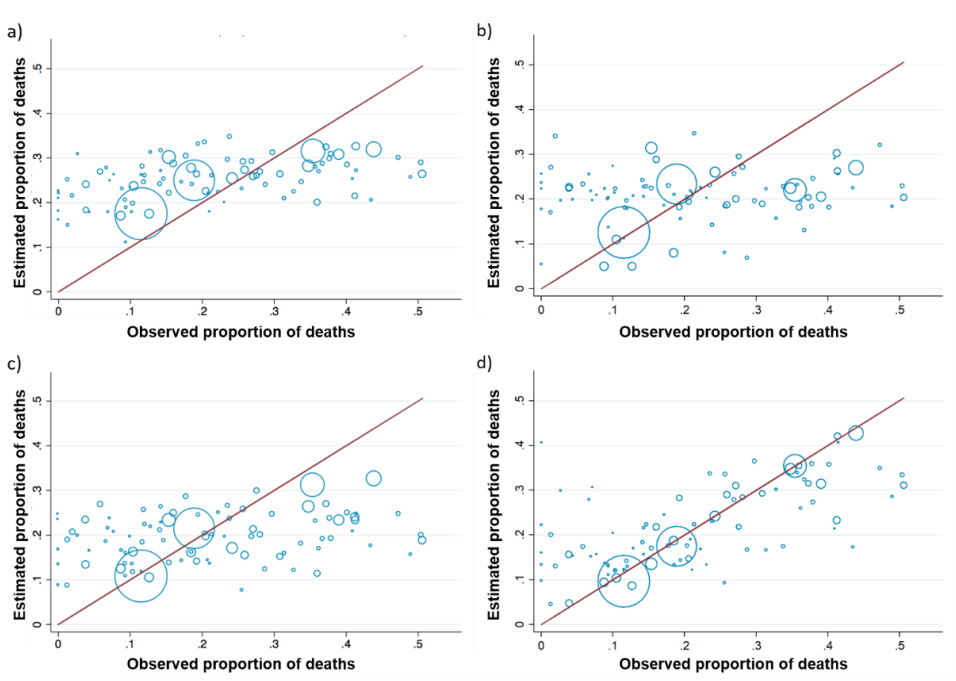
As expected, the random effects component appears to dramatically increase the predictive accuracy for the COD models. An example in the low mortality model (other/late) is included in Figure 6.7. This figure compares the observed versus predicted validation exercise for the multinomial (Figure 6.7a), binomial (Figure 6.7b), binomial FE (Figure 6.7c), and binomial RE (Figure 6.7d) models. The first three graphs look very similar (including the binomial FE model), with relatively poor GOF between the observed and predicted proportions. The binomial RE model, on the other hand, performs substantially better in this validation exercise.

Figure 6.7: Comparison of the multinomial, binomial, binomial FE, and binomial RE models in the low mortality model: example of "other" in the late period



A similar pattern is seen for the high mortality model. An example of intrapartum in the late period is shown in Figure 6.8. Similar to the previous figure, the multinomial (Figure 6.8a), binomial (Figure 6.8b), and binomial FE (Figure 6.8c) models are performing relatively poorly, while the binomial RE model (Figure 6.8d) appears to have better predictive accuracy in this validation exercise.

Figure 6.8: Comparison of the multinomial, binomial, binomial FE, and binomial RE models in the high mortality model: example of "intrapartum" in the late period



Two examples of how country-level COD predictions change when random effects are included are shown in Table 6.5. Afghanistan and Kenya each had one input study. The binomial RE model results show that the estimates are being pulled towards the study data (i.e. binomial RE results are between study data and binomial FE results). This is the expected result of including country-level random effects.

Table 6.5: Differences in predicted estimates from the binomial FE and binomial RE models for two example countries (Afghanistan and Kenya)

	Afghanistan			Kenya		
	% in study	**Average[1] binomial FE[2] prediction**	**Average binomial RE[3] prediction**	**% in study**	**Average binomial FE prediction**	**Average binomial RE prediction**
Intrapartum	10.5	35.2	18.7	21.3	42.6	24.6
Congenital	3.2	6.9	5.6	---	5.1	3.4
Preterm	12.5	22.5	15.0	20.0	26.0	21.2
Sepsis	39.1	15.1	27.2	46.7	13.1	40.6
Pneumonia	---	11.2	10.8	---	9.9	6.6
Diarrhoea	1.2	4.2	1.3	2.7	0.4	1.4
Tetanus	4.0	1.0	10.2	---	0.6	0.4
Other	29.6	3.9	11.1	9.3	2.3	1.8

[1] Average prediction is mean percentage for 2000-2015 prediction; [2] binomial FE = model using only fixed effect component of mixed effects binomial model, [3] binomial RE = model using fixed and random effects components of mixed effect binomial model

6.4 Discussion

In this chapter, I have presented some results of implementing a neonatal COD modelling strategy based on a set of binomial models instead of the existing multinomial approach. Compared to the multinomial models, the binomial models have theoretical advantages (e.g. alignment between our covariate selection and model regression approaches) and disadvantages (e.g. scaling factor required to fit deaths into the total envelope). Our findings suggest that the binomial models (without random effects) are not a clear improvement over the existing multinomial models. For several causes in the low and high mortality models, the binomial model performance was similar to that of the multinomial. When the binomial had better goodness of fit, much of this improvement appeared to be due to a better fit for a small number of observations with large numbers of deaths. For causes that had moderate or poor fit in the multinomial models, this generally continued to be true for the binomial models.

Prior to this analysis, we believed it was plausible that the covariate-cause relationships may be stronger for the binomial models since here we are modelling the odds versus all other causes combined (instead of against some baseline cause as is done in the multinomial). However, we found that these underlying cause/covariate relationships appeared to be similar for both the binomial multinomial models, often without clear patterns. These weak underlying relationships are potentially a leading cause of the predictive accuracy issues discussed in chapter 5.

I also presented results from including country-level random effects in the binomial models. This analysis was performed because we were unable to incorporate random effects into our multinomial models (section 5.2.3). We found that adding random effects to the binomial models was a non-trivial task. The approach we used, which relied on using Stata's xtmelogit command, appeared to be generally inefficient and sometimes finicky. For example, we encountered issues with the software crashing at initial value selection, which was potentially a software bug at the time we conducted this analysis. This may be indicative of how nascent some of the relevant programming is even within well-established statistical software. Given these issues, we are not confident that this approach would work well if we changed the current models (e.g. with updated input data) or included additional modelling components (e.g. model averaging).

These results did demonstrate, however, that the random effects performed as expected and were able to pull a given country's estimates towards the empirical data from that country. This

result was consistent across all causes and models. This is also the key reason that the binomial RE model appeared to perform better than the other three models (multinomial, binomial, binomial FE) in the validation exercise. We expect such an improvement since the validation uses the input dataset for prediction, and therefore each country (or study) is present in both the input and prediction datasets. The lack of improvement in predictive accuracy for the binomial FE model suggests that the only improvement appears to come from inclusion of the random effects component. Thus, the binomial RE model is unlikely to improve predictive accuracy for countries that do not have observations in the input dataset. Currently, that consists of all low mortality prediction countries (since their data are not included in the input data) and most of the high mortality prediction countries (since only eight countries have nationally representative VA studies so far).

As mentioned briefly in section 5.3.2, there are a few important issues to consider when implementing and interpreting random effects given the quality of data in our input dataset. First, most input studies are single data points and not time series, and most countries with a nationally representative study have only one such study (so far). Additionally, the input studies, including the nationally representative ones, have wide ranges in size and quality. Thus, one question is how much a single data point should affect a country's modelled estimates. In other words, what strength of random effect should be allowed? And given the methodological issues in the input data, how much of the difference between the modelled fixed effect estimates and empirical data represents a real difference in the COD distribution compared to random or systematic errors in the empirical data. I discuss each of these issues and other related topics in more detail in section 7.4.

If we were to make the switch to binomial instead of multinomial models, there are several issues that would still require investigation. First, we would need to test for model instability as was done in chapter 5 for the multinomial models. This would help inform decisions on modelling strategies such as the need for model averaging, and help us understand if the differences in predictions between the binomial and multinomial models are arising at least partly through instability. Additionally, we would investigate alternative mixed effects programming, including potentially switching to another programming language (e.g. R). Finally, we would need to understand which model performance solutions are feasible with this mixed effects model. For example, the full search algorithm presented in section 5.1.3 for covariate selection is computationally infeasible with this approach, and several model averaging techniques are likely to be difficult to perform.

As mentioned in section 5.3.3, the modelling issues from chapter 5 were presented alongside the binomial analyses in this chapter during the 2016 annual MCEE meeting. During this meeting of an expert group on child COD issues, it was decided that we would investigate a shift to the Bayesian framework for the COD models. This decision was made for several reasons, including the lack of definite predictive accuracy improvements with the binomial models, the challenges we encountered while implementing the binomial mixed effects models, and because of advantages we saw with the multinomial approach (e.g. not having to drop studies with missing input data for certain causes). A more detailed discussion is included in section 5.3, and results from our Bayesian proof-of-concept COD models are presented in chapter 7.

7 Alternative modelling approach: Bayesian framework with random effects

The frequentist versus Bayesian framework choice is sometimes categorized as a philosophical debate about the meaning of probability. In this work, we sidestep this debate: we chose to convert our cause-of-death (COD) models into the Bayesian framework for practical reasons. Namely, we had difficulty in incorporating random effects into our existing COD models (section 5.2.3), and the Bayesian framework is well suited to this problem. In this section, I describe our work on shifting the neonatal COD models from our previous "classical" approach into the Bayesian framework, including the incorporation of country-level random effects and implementation of covariate selection within this framework. The work presented here is a proof-of-concept demonstration, with ongoing efforts to finalize these Bayesian models in order to produce publishable national-level neonatal COD time series estimates. As such, various analyses including out-of-sample validation and calculating credible intervals (i.e. Bayesian uncertainty) are important but beyond the scope of this current exercise.

7.1 Introduction

7.1.1 Overview of Bayesian statistics

Thomas Bayes first laid out the theorem which is the foundation of Bayesian theory and methods in the mid-18th century, at a time when the mathematical field of probability was nascent but rapidly expanding [188]. The idea Bayes put forth was simple: one's belief should be updated based on new evidence. Pierre-Simon Laplace soon furthered and generalized this foundation into the Bayesian theory we are familiar with today [188]. But over time, Bayesian methods fell out of favour for two key reasons: they were considered more 'subjective' than other statistical methods and they were less tractable prior to the advent and widespread availability of high-speed computers. Instead, 'frequentist' methods, popularized by statisticians like R.A. Fisher, J. Neyman, and E. Pearson, dominated statistical thinking for most of the 20th century. But the use of Bayesian methods has become increasingly popular over the last few decades in a wide range of disciplines, from computer science and hard sciences like physics to social sciences like psychology and economics [189, 190].

The cornerstone of Bayesian methodology is a specific view of probability. For Bayesians, probabilities are a way to describe uncertainty about our knowledge of the world. Probability

distributions for parameters are a way of expressing this uncertainty; we do not have complete information about parameters, so we have some uncertainty (formalized through a probability distribution) about what we consider to be reasonable values for these parameters. This probability changes as we are presented with new data and as we update our prior beliefs. This view of probability is embodied by the simple but profound Bayes theorem (Equation 7.1):

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \qquad \text{(Eq. 7.1)}$$

where P(H) is the prior probability of the hypothesis; P(D) is the probability of observing the data; P(D|H), or the likelihood, is the probability of observing the data D given our hypothesis H; and P(H|D), or the posterior distribution, is the probability of the hypothesis given observing the data D. Thus, the posterior belief is based on a combination of the probabilities associated with the hypothesis being true, the data being observed, and the data being observed given the hypothesis being true.

The prior P(H) is based on previous beliefs and is thus the part of Bayes theorem that adds some level of 'subjectivity'. Priors can range from strongly informative to uninformative depending on the strength of belief in the hypothesis. Careful selection of the priors is essential as they can have a substantial effect on the posterior distribution, especially when the data (i.e. D above) are limited. The prior is also an important part of the iterative nature of Bayesian analyses. If new evidence is introduced, then the previous posterior would become the new prior.

The ultimate objective of Bayesian analysis is to find the posterior distribution. This provides information on the credibility of the parameter values (e.g. based on the posterior probability density), as well as the associated uncertainty. A simplification of Equation 7.1 is to say that the posterior distribution is proportional to the likelihood times the prior. While some posterior distributions can be calculated analytically, the majority consist of intractable integrals. This was a major challenge for Bayesian statistics, and was not overcome until the Markov chain Monte Carlo (MCMC) class of algorithms became more widely applied to Bayesian computation in the last few decades [191, 192]. I discuss the details of Bayesian analysis, including MCMC methods, in the next section.

7.1.2 Relevant elements of Bayesian modelling

Just as with frequentist statistics, there are many aspects of Bayesian modelling. Here, I introduce key elements of Bayesian modelling which are relevant for our work.

Model building

The core elements for performing a Bayesian analysis are to: 1) determine the parameter(s) of interest; 2) formulate a probability model that is suitable for the data; 3) choose prior distributions based on existing evidence and/or expert opinion; 4) gather the data; 5) construct a likelihood function based on the data and probability model; 6) calculate the posterior distribution by combining the prior distribution with the likelihood function using Bayes theorem; and 7) summarize features of interest from the posterior distribution [193].

The complexity of this last step depends on the nature of the posterior distribution. When the prior and posterior are conjugate distributions (i.e. in the same family of distributions), the posterior distribution is simple to solve for analytically. This requires the right combination of distributions for the prior and likelihood (e.g. both Gaussians; beta distribution for prior and binomial distribution for likelihood). However, such conjugate distributions are uncommon in real-world Bayesian analyses, and most posterior distributions will instead be intractable integrals that must be solved for numerically. For these, methods have been developed which sample from the posterior distribution (next subsection).

Sampling from the posterior distribution

MCMC (Markov chain Monte Carlo) methods can approximate the posterior distribution numerically through many repeat random samples from a distribution, thereby circumventing the analytically impossible task of solving intractable integrals. This methodological development, alongside improvements in computational power, revolutionized Bayesian statistics by the 1980s [194]. While there are some other methods for determining the posterior distribution, MCMC algorithms are the overwhelmingly popular choice.

Monte Carlo refers to a class of algorithms that draws repeat random samples from a distribution, thereby allowing estimation of various distribution properties without knowing the distribution itself. Markov chains are stochastic memoryless processes, meaning that a random sample depends only on the previous random sample and none of those before this previous one. MCMC methods, which combine these two concepts, are a powerful tool in Bayesian analysis because they allow for a sequence of memoryless random samples of the distribution

of interest (i.e. the posterior distribution in Bayesian analysis). In general, MCMC algorithms work as follows [195]: 1) initiate the chain with a plausible starting value (or allow the algorithm to select a starting value), 2) move to a new random location and calculate the posterior probability there; remain at the new location if the posterior probability is higher or go back to the previous value if not, and 3) repeat step 2 for as many iterations as are specified for the process. With enough iterations, this will ideally result in a good approximation of the posterior distribution. Multiple chains can be run for the same analysis, with convergence occurring once the different chains stably reach the same value. There are various MCMC algorithms, including the popular Gibbs sampling method [196] which we use in this work. Gibbs sampling draws random samples from the conditional probability distribution for a parameter (i.e. the distribution when other parameters are held at a fixed value).

Covariate selection

There are numerous covariate selection approaches when using Bayesian methods [197-199]. Covariate selection within a Bayesian framework is a quickly growing area of research, with more sophisticated methods emerging each year. The types of methods vary substantially, and a thorough investigation of Bayesian covariate methods is beyond the scope of this thesis. However, some criteria that can help narrow our search include needing a method that can be used for regressions and works with MCMC, as well as ideally improving model stability and allowing for multinomial covariate selection. For these reasons, we chose to investigate the implementation of the Bayesian lasso given the potential suitability of the lasso method for our models as described in section 5.1.4.

The Bayesian version of the lasso regression [175] works similarly to the classical one. Namely, the lasso involves adding all potential covariates to the model (thus permitting multinomial covariate selection) and including a term that penalizes coefficient size. This is a shrinkage method, and thereby aims to improve stability by reducing the covariate coefficient values. In the frequentist version, the penalty term is added to the likelihood function while in the Bayesian lasso it is implemented by applying a Laplacian prior (i.e. a double exponential) to the parameters. In the former, the lasso regression will reduce some coefficients to zero and thus perform covariate selection. The coefficient values can never be exactly zero in the Bayesian lasso because of the double exponential. However, with a strong enough penalty many of the coefficient values will be very close to zero. The strength of the Laplacian prior, and thus the amount of coefficient shrinkage, is determined by a term (lambda) that can be adjusted to tune the penalty.

7.1.3 Frequentist versus Bayesian viewpoints

The key difference between the frequentist and Bayesian frameworks is the lens through which each views the concept of probability. In contrast to the Bayesian view of probability described in section 7.1.1, frequentists view probability as a fixed measure that is equal to the frequency at which an event occurs. Uncertainty is only present because we are unable to sample to infinity; if we were, we would arrive at the "true" probability for a given event. Thus, in this viewpoint, there is no underlying randomness aside from that in the data generating process; parameters and probabilities are fixed entities. Since there is a "true" probability, the frequentist approach views subjectivity as unnecessary or even unscientific.

This difference in what probability means affects how statistical problems are approached. Table 7.1 presents a brief comparison of how frequentist and Bayesian methodologies approach the various components of hypothesis testing, including data, parameters, and uncertainty.

Table 7.1: Comparison of Bayesian versus frequentist viewpoints

	Frequentist viewpoint	Bayesian viewpoint
View of probability	Fixed measure of frequency as sampling goes to infinity	The degree of belief in a statement about an unknown variable
View of data	Repeatable through random sampling	Fixed
View of parameters	Fixed (but unknown)	Probabilistically described
Prior information	Considered subjective; intentionally not included	Fundamental to theory
Output from calculations	$P(D \mid H_0)$	$P(H \mid D)$
Uncertainty	Variability of data for fixed parameter value	Variability of parameter for fixed data
Computation	Fairly simple	Can be intensive (hence algorithms)

There are circumstances in which these different probability viewpoints and analytical approaches yield strikingly different results for the same problem [200, 201]. However, Bayesian and frequentist approaches produce the same or similar results for many problems, particularly those with large samples and uninformative priors [201]. Statisticians do not always fall cleanly into one of the two camps, and may choose either Bayesian or frequentist approaches for pragmatic rather than philosophical reasons.

7.1.4 Our work

In the rest of this chapter, I describe our work in transitioning our neonatal COD models from the classical to Bayesian framework. The three main goals, all within the Bayesian framework, were to: 1) reproduce the classical neonatal COD results, 2) include country-level random effects, and 3) implement a covariate selection method. The work presented here is a proof-of-concept demonstration that such a transition to the Bayesian framework is feasible for our models. For this reason, I have not produced a full set of neonatal COD estimates as I did in chapter 4. Instead, in this chapter I describe our shift to the Bayesian framework, and compare the results to the equivalent version of our classical models.

7.2 Methods

7.2.1 Overview of differences between the classical and Bayesian cause-of-death modelling methods

We tried to keep the Bayesian modelling strategy as similar as possible to the classical approach (section 4.2) for this proof-of-concept exercise. The key changes we made were those which were needed to shift the multinomial models into the Bayesian framework. Many of the other features of the modelling approach remained the same between our classical and Bayesian models, including dividing countries into three estimation groups and having four separate models (early and late neonatal periods each for low and high mortality countries) to estimate the COD distributions. Table 7.2 includes the list of changes for the Bayesian modelling strategy, with explanations for why the changes were made. For comparability with the Bayesian results, we re-ran the classical model with the relevant changes described in Table 7.2. These and the Bayesian changes are described in more detail in the following sections.

Table 7.2: Differences between the classical and Bayesian modelling strategies

	Change in Bayesian model	Reason for change
Country groupings	No change (i.e. high mortality model, low mortality model, high-quality VR countries)	---
Model groupings	No change (i.e. 4 models: early and late period models for high and low mortality country groups)	---
Cause categories	Sepsis, pneumonia, diarrhoea, and tetanus grouped together as "infections" in high mortality model	To reduce the number of input studies which need to be dropped due to unreported causes
Input data	Dropped studies with unreported causes for modelled cause categories	To run the multinomial model without re-writing the likelihood to account for unreported causes
Covariate selection to recreate classical results	Used same covariates selected by classical approach (i.e. did not perform new covariate selection)	To ensure direct comparability with classical results
Bayesian covariate selection	Implemented lasso regression covariate selection as proof-of-concept	To incorporate Bayesian covariate selection into multinomial model
Regression approach	No change for regression type (i.e. multinomial logistic regressions); did not re-write the likelihood to account for unreported causes in input studies	Did not re-write likelihood to simplify Bayesian proof-of-concept model
Weighting of input observations[1]	Did not include weighting	Simpler to remove from classical model than to add to Bayesian model for proof-of-concept test
Statistical package(s)	R (v3.2.2) and JAGS within R (R2jags, v0.5-7)	R is well-tested for the types of Bayesian analyses we needed
Random effects	Implemented at country level	Unable to implement in classical multinomial model
Outputs	No change (i.e. proportional cause distribution and number of deaths)	---

[1] To give intermediate weighting between equal weight to each death and equal weight to each study or country-year in the input data (as described in section 4.2.3).

7.2.2 Bayesian modelling to recreate 'classical' results

The Bayesian statistical analyses were done using R (version 3.2.2) and JAGS within R (package R2jags, version 0.5-7). In this section, I describe the processes we used for building the proof-of-concept Bayesian multinomial models and comparing them to the classical results.

Data inputs

We used the same data inputs as those described in section 5.1.2. Modifications made to the input data, as described in Table 7.2, included the following:

- Collapsing sepsis, pneumonia, diarrhoea, and tetanus into one broad "infection" category for modelling, resulting in a total of five modelled categories (intrapartum, preterm, infection, congenital, and other)
- Dropping input observations with missing data for any of the five above-mentioned modelled causes

We made these changes to deal with the "unreported causes" issue for high mortality model input studies. In the classical model, we made assumptions about the cause category into which deaths from an unreported cause would have been assigned and re-wrote the likelihood function accordingly (section 4.2.3). For example, if pneumonia, diarrhoea, or tetanus were unreported, these were assumed to have been in the "sepsis" category. To simplify the modelling, we chose to not include this step for the proof-of-concept Bayesian exercise. We therefore needed to drop high mortality model input studies which had unreported causes (the input vital registration [VR] data for the low mortality model had no unreported causes)[1].

The non-sepsis infection categories (pneumonia, diarrhoea, and tetanus) had the largest number of unreported causes in studies (e.g. pneumonia was missing for 55% of input observations, Table 4.3). Thus, we grouped these causes alongside sepsis into a broader "infection" category for the modelling, which reduced the number of studies we needed to drop. This resulted in a modified high mortality input dataset of 87 observations (17 dropped) for the early period model and 76 (15 dropped) for the late period model. For comparability, we also used this modified input dataset when we re-ran the classical model.

Covariate selection for models

We used the same covariates selected through the classical covariate selection method (section 4.2.3) for the classical and Bayesian models in order to facilitate a direct comparison between the results. Because we used a collapsed set of causes for the Bayesian high mortality model (i.e. combining infectious causes into one category), we re-ran the classical covariate selection method for the high mortality model on these same collapsed causes. We then used the results from this covariate selection for both the revised classical and Bayesian models. Our later work on Bayesian covariate selection is described in section 7.2.5.

[1] Since the completion of the work reported here, Dr. David Prieto-Merino (a member of our team) has developed a method to account for unreported causes in this Bayesian framework.

Multinomial models

Like the classical model, we used multi-cause multinomial logistic regression models to fit the data for all causes simultaneously. For these regressions, we estimated the log of the ratio of each of the other causes to the baseline cause (the "log-cause ratio") as a function of the selected covariates. For simplicity, we did not include weights on the input observations (used as an intermediate between giving equal weighting to each death versus each study or country-year) as we did in the classical model (section 4.2.3). For comparability, we re-ran the classical models without this weighting.

Parameter priors

We used weakly informative Gaussian priors (mean: 0; standard deviation: 100) for the parameter priors of the multinomial regression equations. As a sensitivity test, we also evaluated Gaussian priors with standard deviations ranging from 5 to 50.

Sampling from the posterior distribution

To sample from the posterior distribution, we used the Gibbs sampling method for MCMC. For each of the models, we used 4 MCMC chains and a tenth of the iterations (i.e. thinning) for inference. The other options, including number of iterations (both for burn-in and total) varied based on the model. We chose these based on convergence of the MCMC chains, which we determined visually using parameter trace plots and density plots of the posterior distributions.

Predicted proportions and number of neonatal deaths

To estimate the predicted proportional COD distribution, we applied the posterior distribution parameter estimates from each thinned MCMC step to the country-year prediction covariate values. We then averaged these predictions to obtain the final estimated proportional COD distribution for each modelled country-year. We applied these proportions to the envelope of neonatal deaths for that country-year to obtain cause-specific numbers of deaths. Finally, we compared the results of these outputs from the Bayesian analysis with the corresponding classical results.

7.2.3 Incorporating country-level random effects

We implemented a mixed effects (ME) modelling structure by adding a country-level random intercept to each non-baseline equation of the Bayesian multinomial model. We assumed a Gaussian distribution for the random effects, and "tuned" the strength of these random effects

by adjusting the uniform prior on the standard deviation of the Gaussian distribution. We began with a weak prior (which allows potentially large random effects) by allowing the random effect to be drawn from a Gaussian distribution with a standard deviation (SD) up to 1.6. This corresponds to the odds ratios (ORs) between countries varying between 1/5 to 5. Such wide variation in the OR is very unlikely to be due to true between-country effects, and thus is an overly generous allowance for the variation. We also applied a very strong prior (i.e. allowing small random effects: SD up to 0.01; OR between 1/1.01 and 1.01) to test that the random effect implementation was working as expected. Finally, we tested various SDs to find a medium prior (i.e. an estimate partly between the non-ME and observed data). By applying these different SDs to the random effects prior, we were able to "tune" the strength of the country-level random effect (i.e. how much the modelled estimates would be pulled towards the country's empirical data).

Similarly to the inclusion of random effects in the binomial models (section 6.2.5), we produced two sets of estimates: 1) using the full ME model (i.e. fixed and random effects components) and 2) using only the fixed effects component of the ME model. This allowed us to gauge how much the change in results was due to the random effects component versus different coefficient values for the covariates (i.e. the fixed effect component). For this exercise, we assumed that the random effects were applicable to all input data (i.e. not only to nationally representative studies). All other aspects of the modelling remained the same as described in section 7.2.2.

We also compared the cause-specific goodness-of-fit (GOF) for the non-ME versus ME models by calculating their χ^2 goodness-of-fit metric (Equation 5.1) based on the observed data versus estimated predictions. We did this for the ME model using only the fixed effect components, as well as the full ME model using fixed and random effects.

I present the results for the Bayesian ME analysis as part of a comparison with the classical results and the Bayesian non-ME models. I will use the following terminology to refer to the models:

- Classical – the non-ME classical multinomial models (described in section 4.2.3)
- MCMC – the non-ME Bayesian models (described in section 7.2.2)
- MCMC FE – the Bayesian ME regressions, but using only the fixed effects component for predictions (described in this section)

- MCMC RE – the Bayesian ME regressions, using both the fixed and random effects components for predictions (described in this section)

For some aggregate results (e.g. regional), I include MCMC FE/RE results. This category means MCMC FE results were used for countries with no input data and MCMC RE results were used for countries with input data.

7.2.4 Implementing the lasso for Bayesian covariate selection

We implemented the Bayesian lasso by using Laplacian (i.e. double exponential) priors for the parameter estimates. First, we centered and normalized the covariates in our model to have means of 0 and standard deviations of 1. This was done because the lasso penalty is on the sum of the absolute value of the covariate coefficients, and thus depends on the scale of the coefficients. We included all potential covariates in each cause equation of the multinomial and gave each one a Laplacian prior with a penalty term (lambda). For this exercise, we used only the linear form of covariates and re-ran the classical model covariate selection accordingly. We tested a wide range of lambdas (1, 10, 50, 100, 500, 1000, 1500, 2500, 3000), evaluated their corresponding coefficient values, and generated country-specific proportional COD distributions based on each lambda.

7.3 Main results

7.3.1 Bayesian modelling to recreate classical results

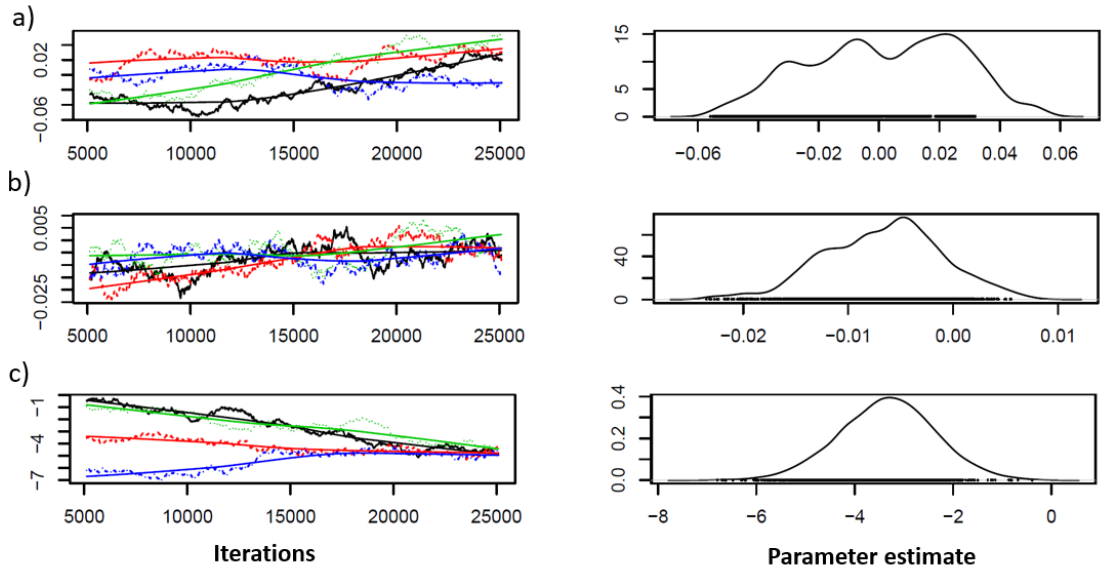
Here, I describe the results of building the Bayesian model to recreate the classical model results, with a focus on convergence of the MCMC chains, parameter estimates obtained, and a comparison of results from the two models. The results presented here are for the early neonatal period.

<u>Convergence</u>

We evaluated convergence visually using trace and density plots. Once convergence appeared to have been reached, we ran several thousand further iterations to verify stability of the chains. An example of these at different levels of convergence for the low mortality model with 25,000 iterations is shown in Figure 7.1. For the low mortality model, the majority of parameters reached convergence by approximately 50,000 iterations, but some needed more than 150,000 iterations and a few required nearly 250,000. Running 50,000 iterations took about 11 hours to run on a high-performance laptop. The high mortality model required fewer iterations to reach convergence, with most parameters reaching convergence by about 20,000 iterations, nearly

all by 50,000, and a final few by around 200,000 iterations. This model also ran much faster, with 50,000 iterations requiring 16 minutes on the same high-performance laptop.

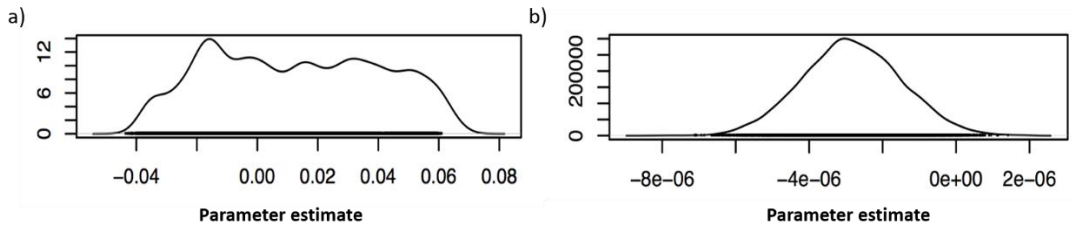
Figure 7.1: Examples of trace (left) and posterior distribution density (right) graphs for varying levels of convergence at 25,000 iterations (low mortality model). a) 3rd IMR spline, sepsis – not yet converged; b) DPT, intrapartum – nearly converged; constant; c) injuries – converged (verified by running further iterations – not shown).



Note: y-axes are "Parameter estimate" for trace plots (left) and "Density" for posterior distribution density graphs (right)

We tested whether convergence was reached more quickly for simpler models by removing quadratic and spline relationships for the covariates (i.e. keeping only linear relationships for each covariate-cause). The results indicated that reaching convergence with fewer iterations was more likely for simpler models. Figure 7.2 shows an example of this, with the density function not showing convergence at 100,000 iterations for a parameter with a spline function (7.2a) while showing convergence by 50,000 iterations for a linear parameter function (7.2b) in the low mortality model.

Figure 7.2: Comparison of the posterior distribution density for a) "complex" model parameter (IMR, 1st spline; sepsis) at 100,000 iterations versus b) "simple" model parameter (IMR, linear; sepsis) at 50,000 iterations in the low mortality model.



Note: y-axis is "Density" for both graphs

Parameter estimates

We compared the parameter estimates (i.e. coefficient values) from the classical versus Bayesian models. The parameter estimates from the Bayesian models were nearly the same as those of the classical model, even before there was visual evidence of convergence for all of the parameters. Table 7.3 shows an example of this for the low mortality model. The equation for pneumonia/preterm (top) has nearly all of the same coefficient values, and there was visual evidence to support the view that convergence had been reached. The covariate coefficients for the injuries/preterm equation were very similar, even though some of those parameters did not appear to have reached convergence at 20,000 iterations (e.g. 2nd and 3rd splines of GFR). By 200,000 iterations the coefficient values were the same as the classical model.

Table 7.3: Comparison of parameter estimates between the classical and Bayesian models: examples from the low mortality model

	Classical model	Bayesian model[1]
Parameter estimates for ln(pneumonia/preterm)		
GNI	6.59×10^{-5}	6.59×10^{-5}
GINI	-0.04	-0.04
U5MR	0.02	0.02
constant	-0.55	-0.54
Parameters estimates for ln(injuries/preterm)		
GFR (S1)	-41	-40
GFR (S2)	22	14
GFR (S3)	109	120
LBW	-0.15	-0.16
LBW^2	0.01	0.01
GNI	9.2×10^{-6}	8.6×10^{-6}
ANC	0.08	0.07
constant	-9.01	-8.19

[1] for 20,000 iterations

We further investigated the difference in parameter estimates over increasing iterations in the Bayesian model. As noted above, we found that the parameter estimates appeared to stabilize even before the parameters appeared visually to reach convergence. For example, some of the parameters in Table 7.4 did not appear to have reached convergence until after 50,000 steps. However, the parameter estimates did not change substantially between 50,000 and 200,000 iterations, and were even fairly close to the converged values by 20,000 iterations. But it is not possible to know in advance whether a parameter estimate has stabilized before the MCMC chain reaches convergence, and thus such a convergence check is still necessary.

Table 7.4: Example of parameter estimates for the classical model versus different numbers of iterations of the Bayesian model: ln(preterm/intrapartum) in the high mortality model

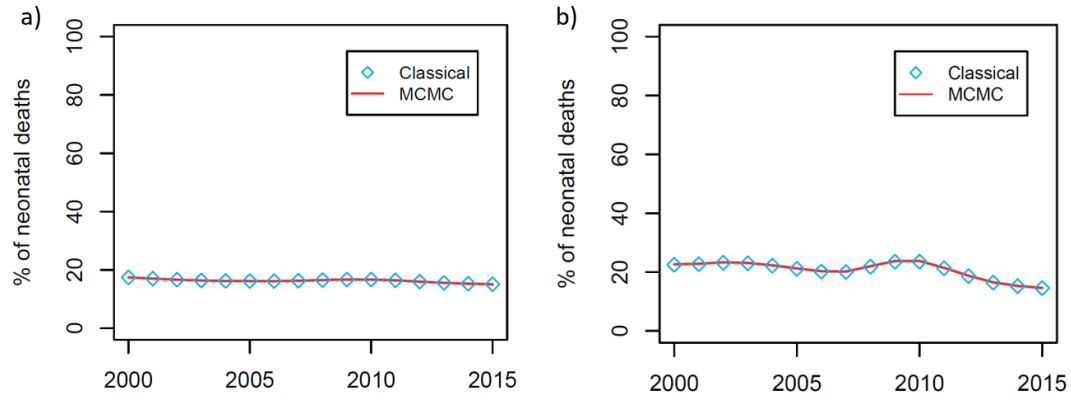
	GFR	GFR^2	LBW (S1)[1]	LBW (S2)[2]	constant
Classical	-44.5	142.8	$11.5x10^{-3}$	$9.0x10^{-3}$	3.0
Bayes: 2,000 steps	-33.8	101.9	$18.6x10^{-3}$	$3.7x10^{-3}$	2.3
Bayes: 20,000 steps	-44.0	140.7	$11.9x10^{-3}$	$8.3x10^{-3}$	3.0
Bayes: 50,000 steps	-44.8	144.3	$11.3x10^{-3}$	$9.3\ x10^{-3}$	3.0
Bayes: 200,000 steps	-44.5	142.9	$11.5x10^{-3}$	$9.0\ x10^{-3}$	3.0
[1] 1st spline, [2] 2nd spline					

We also tested a range of Gaussian priors (section 7.2.2) as a sensitivity analysis and found no difference in the results with these different priors.

Comparison of Bayesian versus classical cause-of-death estimates

As evidenced by the similar coefficient values for the equations, we were able to recreate the classical results using the Bayesian model with weakly informative priors for both the low and high mortality models. Examples of the Bayesian versus classical results by country and cause for the low and high mortality models are included in Figure 7.3.

Figure 7.3: "Recreating" the classical model results: examples of the % of neonatal deaths with the classical versus Bayesian models for the a) low mortality model (intrapartum, Peru) and b) high mortality model (sepsis, Central African Republic)



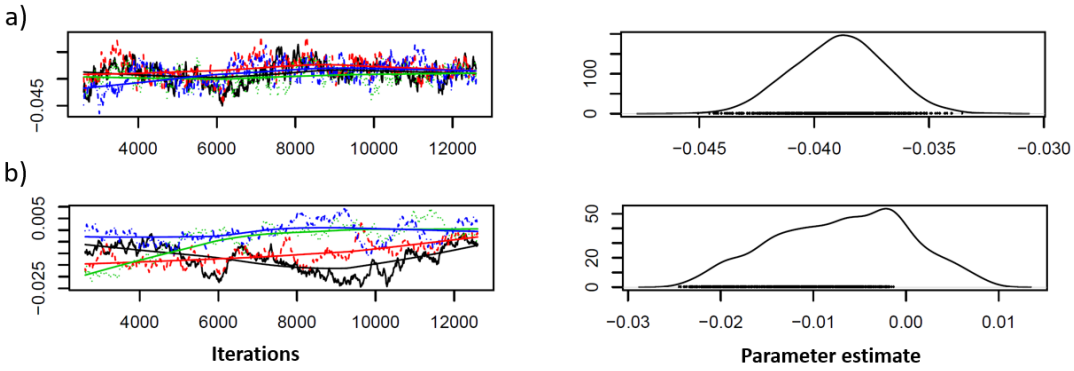
7.3.2 Incorporating country-level random effects

In this section, I present results from implementing the Bayesian ME models, including an examination of the extent to which parameter estimates changed, a comparison of COD predictions, and an assessment of how tuning the random effects prior affects the predictions. The first parts of the subsection are results based on allowing a strong random effect.

Convergence

The ME models required more iterations to reach convergence for the fixed components than did the non-ME models. An example of this is shown in Figure 7.4, where the U5MR coefficient parameter in the non-ME congenital equation (Figure 7.4a) appears to have reached convergence by 15,000 iterations (verified by running 20,000 more iterations – not shown) whereas this is not the case in the ME version (Figure 7.4b). The model run times also increased with the Bayesian ME model compared to the non-ME model. To run 50,000 iterations, the high mortality model required 21 minutes instead of 16 while the low mortality model required 15 hours instead of 11. Together, this meant that the Bayesian ME models typically required about 1.5x as long to run as did the non-ME models.

Figure 7.4: Examples of trace and posterior distribution density graphs for the Bayesian a) non-ME and b) ME models: U5MR in the high mortality congenital equation at 15,000 iterations



Note: y-axes are "Parameter estimate" for trace plots (left) and "Density" for posterior distribution density graphs (right)

Parameter estimates

We found that the parameter estimates between the MCMC and MCMC FE models ranged from similar to substantially different when the random effects were allowed to be strong. Two examples from the high mortality model are shown in Table 7.5. In the top example (preterm), the first spline of LBW had similar parameter estimates in the MCMC and MCMC FE models, while several of the other parameter estimates were rather different. The second example (congenital) is a demonstration of how sometimes even the sign changed between the two models (e.g. U5MR and SSA).

Table 7.5: Two examples of parameter estimates between the classical, MCMC, and MCMC FE models

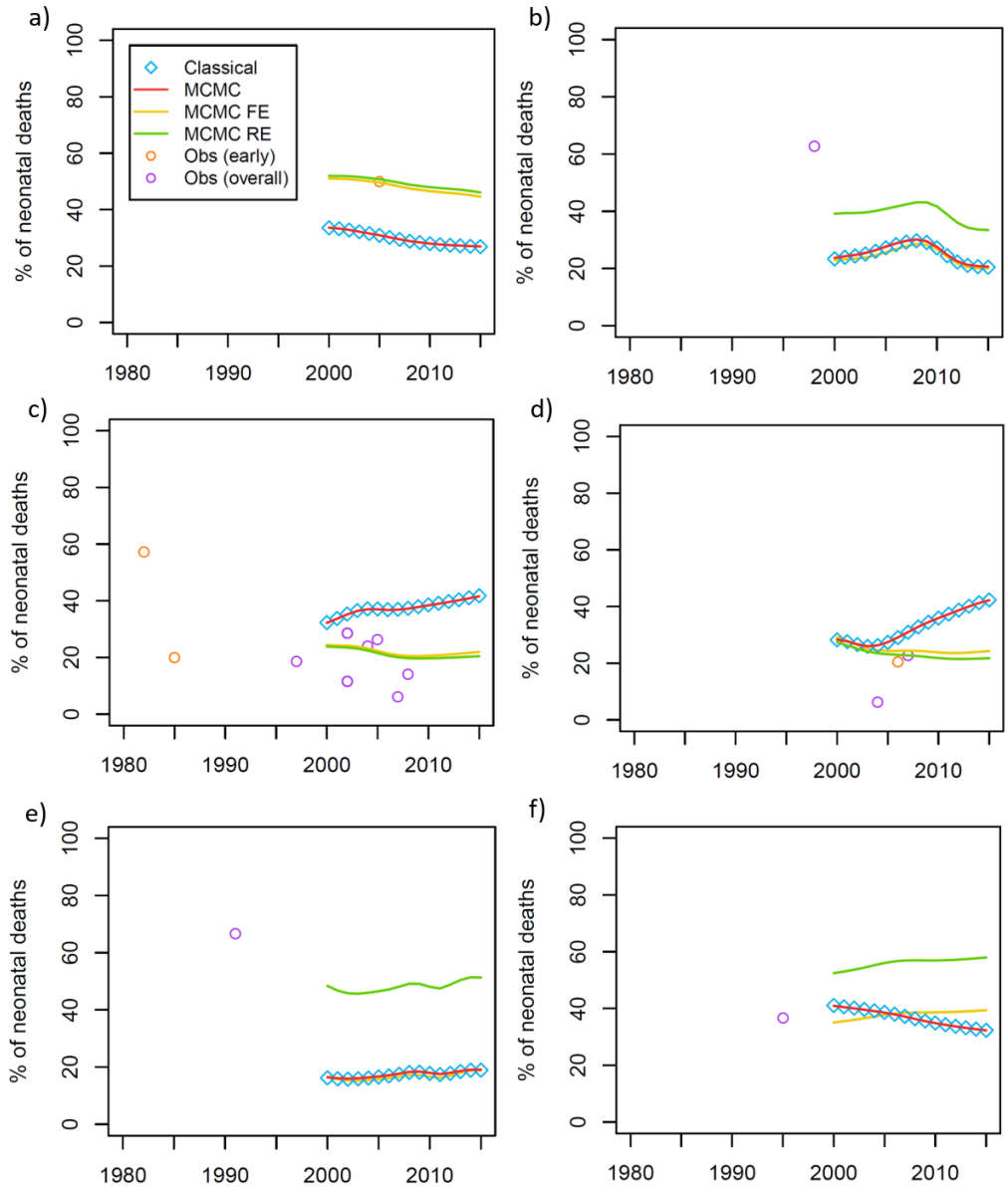
	Classical model	MCMC model[1]	MCMC FE[2] model[1]
Parameter estimates for ln(preterm/intrapartum)[3]			
GFR	-44.5	-44.6	-12.2
GFR^2	142.8	142.9	81.4
LBW (S1)	11.5×10^{-3}	11.4×10^{-3}	11.0×10^{-3}
LBW (S1)	9.0×10^{-3}	9.1×10^{-3}	4.7×10^{-3}
Constant	3.0	3.0	0.01
Parameters estimates for ln(congenital/intrapartum)[3]			
U5MR	-38.9×10^{-3}	-38.8×10^{-3}	5.2×10^{-3}
U5MR^2	13.7×10^{-5}	13.7×10^{-5}	-1.9×10^{-5}
SSA	-0.02	0.02	-1.49
GFR	-37.6	-37.7	-36.4
GFR^2	140.6	140.6	168.7
Constant	2.9	2.9	0.5

[1] for 50,000 iterations; [2] fixed effects are the same for the MCMC FE and MCMC RE models; [3] see Table 4.2 for covariate acronym definitions

Comparison of the classical, MCMC, and MCMC ME models

The country-level random effects appeared to work as expected, with the MCMC RE estimates pulled towards the observed values for countries with input data. However, the relationships between the MCMC (i.e. non-ME), MCMC FE, and MCMC RE estimates varied. Some representative examples are shown in Figure 7.5 for the high mortality model. For some countries/causes, the MCMC RE and MCMC FE estimate levels were similar to each other but different from the non-ME level, with similar time trends for all (e.g. Figure 7.5a). Some had similar time trends but the MCMC RE level was moderately different from the MCMC FE and non-ME levels (e.g. Figure 7.5b). A few countries had multiple input datapoints. For some, the MCMC RE and MCMC FE time trends were different from those for the non-ME model and this change was visually consistent with the input data (e.g. Figure 7.5c). A few others had different MCMC RE and MCMC FE time trends compared to the non-ME model, but the trend was not visually obvious from the input data (e.g. Figure 7.5d). In these cases, observations with larger sample sizes likely influenced the MCMC RE estimates more heavily. Finally, some MCMC RE estimate levels (but not trends) appeared to be substantially influenced by older input datapoints (e.g. Figure 7.5e), while for others the trend was also altered by older data (e.g. Figure 7.5f). This issue of how much older data should influence current estimates is part of a larger discussion of how to choose the tuning strength of the RE (see section 7.4 for further discussion).

Figure 7.5: Examples of the % of neonatal deaths by country and cause with classical, MCMC, MCMC FE, and MCMC RE models for high mortality countries with input data. a) Congo/preterm; b) Guinea/infection, c) Bangladesh/preterm, d) Nepal/preterm, e) Nigeria/intrapartum, f) Bolivia/intrapartum

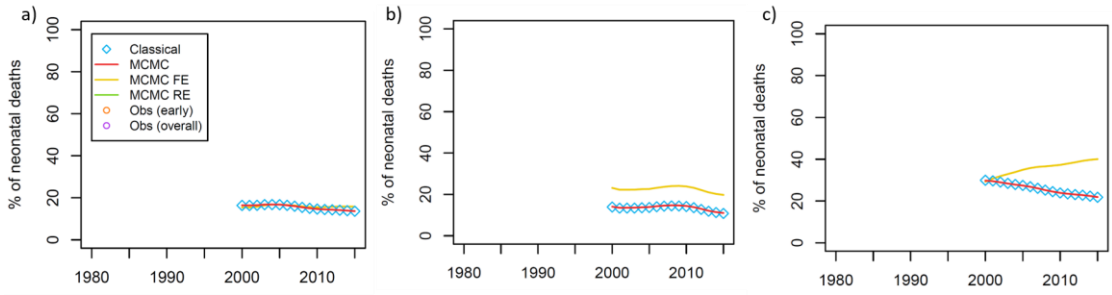


Notes: Country-level random effects were applied to all studies and not only nationally representative ones for the purpose of this proof-of-concept exercise (section 7.2.3); MCMC FE estimates (yellow lines) are alongside MCMC estimates (red lines) for Figures 7.5b and 7.5e.

The MCMC FE results varied widely at times from the non-ME and MCMC RE results (when assuming a strong random effect). Figure 7.6 shows representative examples for countries which did not have studies in the input dataset. For some countries/causes with no input data, the MCMC FE estimates were consistent with those from the non-ME model (e.g. Figure 7.6a). For others, the MCMC FE and non-ME trends were the same but the levels were different (e.g.

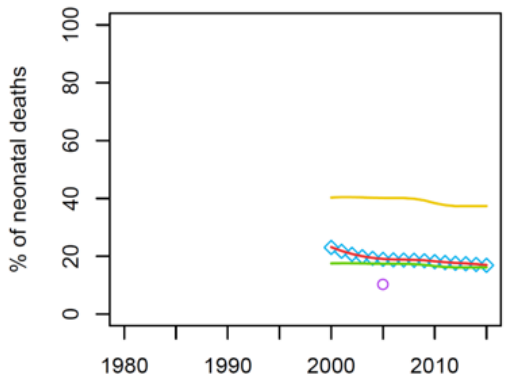
Figure 7.6b). There were several countries/causes in which the MCMC FE estimate had a different trend from the non-ME model (e.g. Figure 7.6c).

Figure 7.6: Examples of the % of neonatal deaths by country and cause with classical, MCMC, and MCMC FE models for high mortality countries without input data. a) Namibia/infection, b) Myanmar/infection, c) Bhutan/intrapartum



The differences between the MCMC RE and MCMC FE estimates are important to consider. Figure 7.7 shows an example of a country (Iran) with one observed datapoint, where the MCMC RE and non-ME results were very similar (~20% of neonatal deaths from intrapartum) but the MCMC FE result was remarkably different (~40%). Assuming this choice of random effect strength, Iran's reported estimate would be the MCMC FE result if there had been no input data for the country.

Figure 7.7: Example of differences in the % of neonatal deaths estimated by the MCMC FE and MCMC RE models (Iran, intrapartum)



Such differences are important because the majority of countries in the high mortality model have no nationally representative studies in the input dataset, and therefore their estimates would come from the fixed component of an ME model. Later in this section, I discuss our work to tune the strength of the random effects to address what are sometimes large differences between the MCMC FE and non-ME MCMC models. Furthermore, the broader implications of random effects strength tuning are included in the discussion (section 7.4).

When the models were applied to the input dataset for validation, the MCMC RE model with a strong random effect performed better (i.e. predicted estimates were closer to the observed values) than the non-ME MCMC model for all causes (Table 7.6). At this random effect strength, the non-ME MCMC model performed better than the MCMC FE model for all but the "other" cause category (Table 7.6). However, as the allowed random effect was weakened, the χ^2 values for the three models moved towards convergence and the MCMC FE outperformed the non-ME model for three causes (Table 7.6).

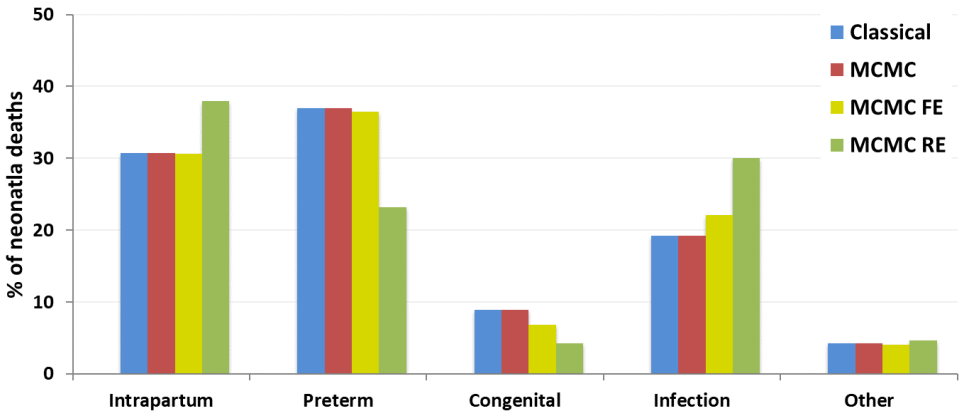
Table 7.6: Comparison of χ^2 values for the Bayesian MCMC, MCMC RE, and MCMC FE models when allowing strong- versus medium-strength random effects

	MCMC model (non-ME)	Strong random effect[1]		Medium random effect[1]	
		MCMC RE	MCMC FE	MCMC RE	MCMC FE
Intrapartum	1504	376	2774	649	1483
Preterm	723	385	1684	562	982
Congenital	584	325	1146	407	671
Infection	831	612	1907	685	731
Other	1159	780	1107	1004	1079

[1] Standard deviation (SD) of random effects prior allowed up to 1.6 for strong prior and 0.7 for medium prior (section 7.3.2)

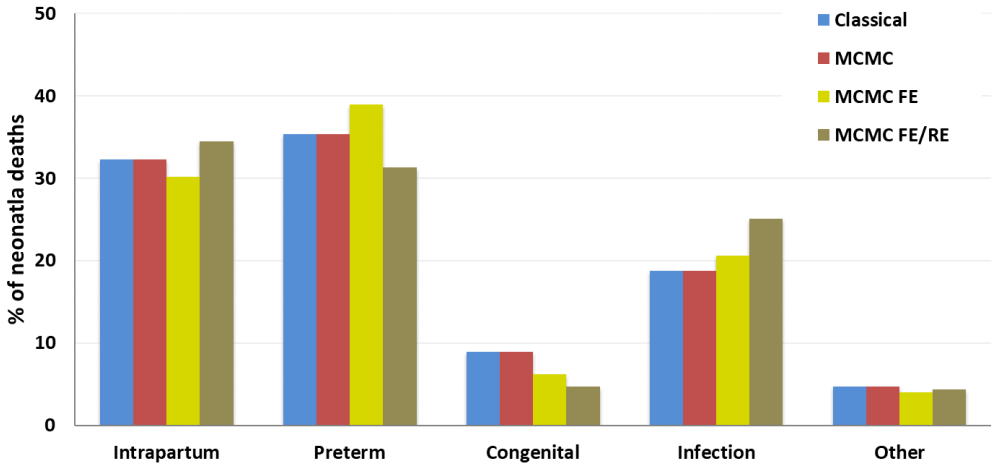
Figure 7.8 shows the proportional COD distribution for the 20 countries with studies in the input database. Overall, the MCMC and MCMC FE models were closer to each other in their aggregated results than to the MCMC RE results when assuming a strong random effect. The intrapartum and infection results tended to be higher for the MCMC RE model, while they were lower for preterm and congenital. Note that the country-level random effects were estimated using all studies for this proof-of-concept exercise, and not only the nationally representative ones.

Figure 7.8: % of neonatal deaths by cause and model type (classical, MCMC, MCMC FE, and MCMC RE) for 20 countries with input data in the high mortality dataset



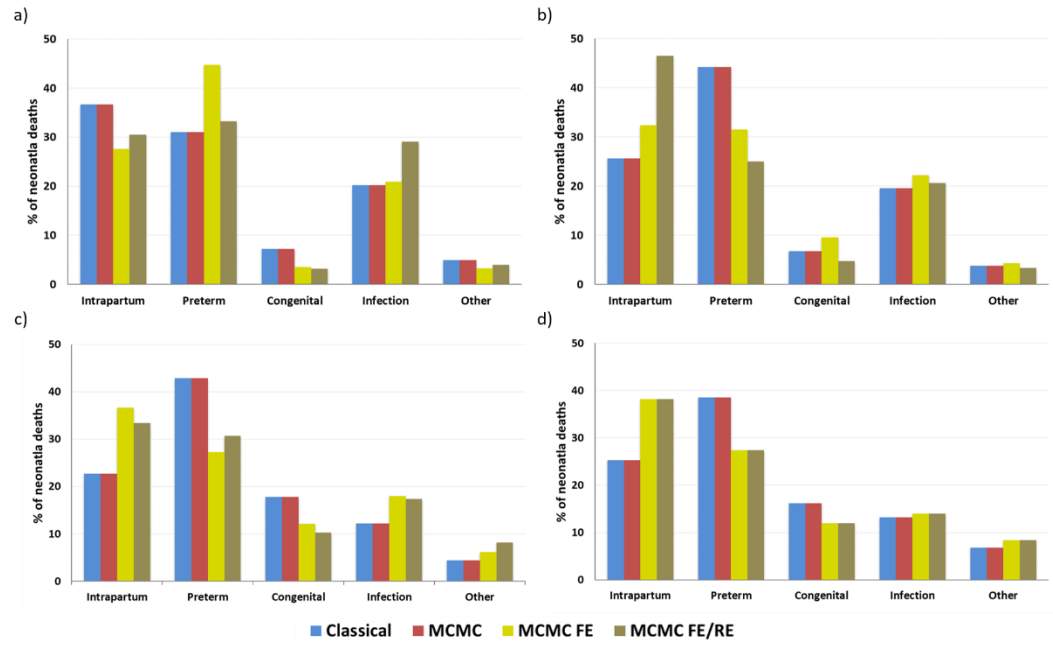
This pattern (intrapartum and infection higher; preterm and congenital lower) was replicated in the aggregated results for all modelled countries (Figure 7.9), though to a lesser extent. The difference was less pronounced because the last category (brown bar in Figure 7.9) combines countries with no input data (i.e. MCMC FE results only) with those with input data (i.e. MCMC RE results).

Figure 7.9: % of neonatal deaths by cause and model type (classical, MCMC, MCMC FE, and MCMC FE/RE) for 80 high mortality model countries



The regional estimates show some more substantial differences between the model types (Figure 7.10). For example, Sub-Saharan Africa had a higher proportion of preterm deaths in the MCMC FE model than the other models and compared to the other regions. The random effects dramatically increased the intrapartum estimate while decreasing preterm in South Asia (and to a lesser extent in East Asia and the Pacific). The Europe and Central Asia region had no input studies, and therefore the MCMC FE and MCMC FE/RE results were the same (Figure 7.10d).

Figure 7.10: % of neonatal deaths by cause and model type (classical, MCMC, MCMC FE, and MCMC FE/RE) for four regions. a) Sub-Saharan Africa (n=43), b) South Asia (n=5), c) East Asia and the Pacific (n=14), d) Europe and Central Asia (n=6)



Tuning the strength of the random effects

We were able to adjust the strength of the country-level random effects by tuning the priors for the random effects intercepts. The choices of the SD of the random effects priors to allow weak and strong random effects strengths are explained in section 7.2.3. Briefly, the strong prior (i.e. weak random effects strength) allowed the random effects to be drawn from a Gaussian which had an SD of the log of the odds ratio that could vary up to 0.01, while the weak prior (i.e. strong random effects strength) allowed an SD up to 1.6. We tested several SDs in between these two, and chose a medium random effect strength by setting the SD to 0.7 (i.e. the odds ratios between countries could vary between 1/2 to 2). Figure 7.11 shows an example of the MCMC, MCMC FE, and MCMC RE estimates for 2000-2015 with these strong (Figure 7.11a), medium (Figure 7.11b), and weak (Figure 7.11c) random effects. When the random effect strength is allowed to be strong, the MCMC RE estimates are pulled very close to the observed data, while in this case the MCMC FE is substantially higher than the MCMC RE, and moderately higher than the non-ME MCMC estimate. With a medium strength random effect, both the MCMC FE and MCMC RE estimates came closer to the non-ME estimate. Finally, when the random effect was constrained to be weak, the three estimates were nearly the same. Thus, adjusting the random effect prior can substantially affect both the MCMC RE and MCMC FE estimates.

Figure 7.11: Example of the % of neonatal deaths with a) strong, b) medium, and c) weak country-level random effects (Afghanistan, preterm)

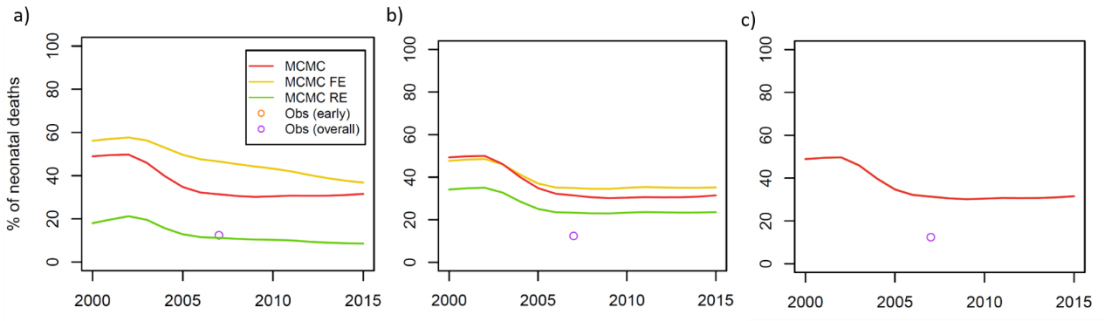
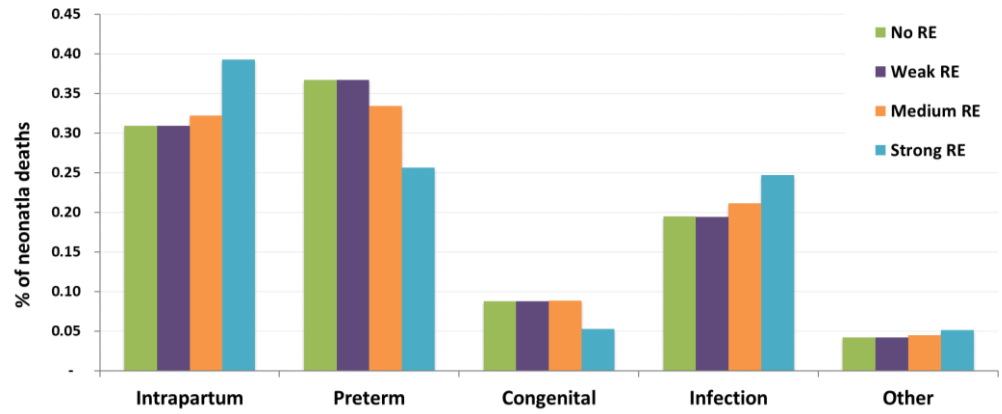


Figure 7.12 shows how the proportional COD distribution changed with the weak, medium, and strong random effects for the 20 high mortality countries with input data. The weak random effect effectively gave the same result as the MCMC model (i.e. no random effects), while the proportions for intrapartum and preterm had differences of about 10 percentage points with the strong compared to no random effect. The medium strength random effect can be tuned to fall anywhere between these two.

Figure 7.12: % of neonatal deaths by cause, model type (classical, MCMC, MCMC FE, and MCMC RE), and random effects strength for 20 countries with input data in the high mortality dataset



Note: RE = random effect

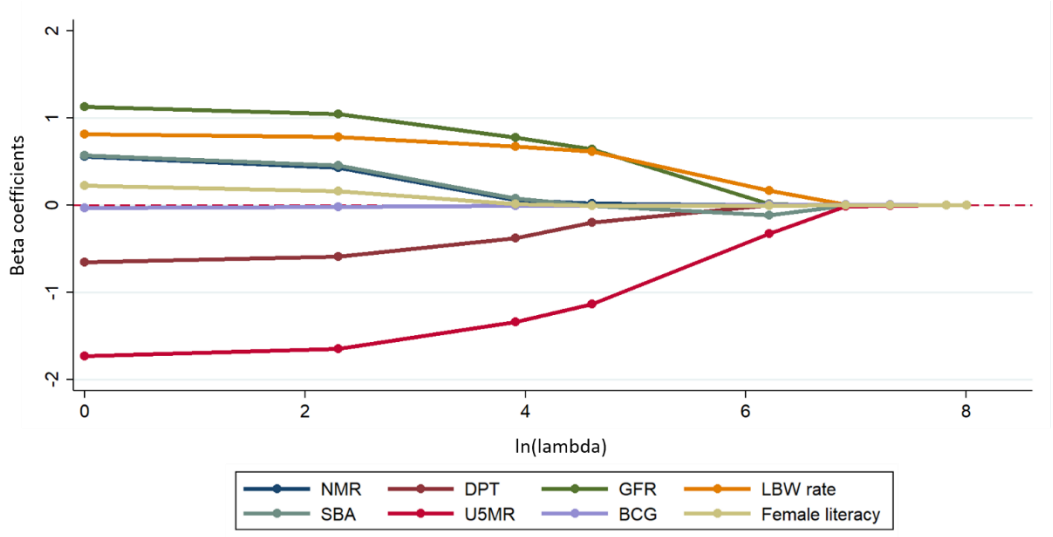
Overall, the results here indicate that we were able to successfully tune the strength of the random effects, with the implications of this discussed in section 7.4.

7.3.3 Implementing the lasso for Bayesian covariate selection

We tested a range of lambda values (from 1 to 3000) for the lasso penalty while implementing the Bayesian lasso covariate selection in the non-ME model. Overall, the covariate (beta)

coefficients moved towards zero as the lambda value increased (e.g. Figure 7.13). This is the expected result of a lasso regression, thus suggesting that our Bayesian lasso implementation was successful.

Figure 7.13: Example of covariate (beta) coefficients moving towards zero with an increasing Bayesian lasso penalty (other, high mortality model)



Across the causes, we found that approximately six to eight covariates were retained in the model with a low penalty (e.g. lambda: 10), while a high penalty (e.g. lambda: 3000) resulted in 0-1 covariates. A medium penalty (e.g. lambda: 1000) typically resulted in three to five covariates in the model, which may represent a reasonable compromise between complex models and underfitting.

Table 7.7 shows the covariates chosen by Bayesian lasso using two medium-range penalties (lambdas of 1000 and 1500) compared to those selected using our classical approach (section 4.2.3). Several of the covariates were the same between the classical model and Bayesian lasso. In general, the higher lambda penalty (1500) appears to be closer to the classical model. Only two covariates showed up in the classical covariate selection but not the Bayesian lasso at a lambda of 1000 or higher: GFR in infection and LBW in preterm. The former was dropped from the lasso between a lambda of 100-500, and the latter was dropped between a lambda of 500-1000. In general, there is overlap between the classical and Bayesian lasso results, but there are also several differences in selected covariates.

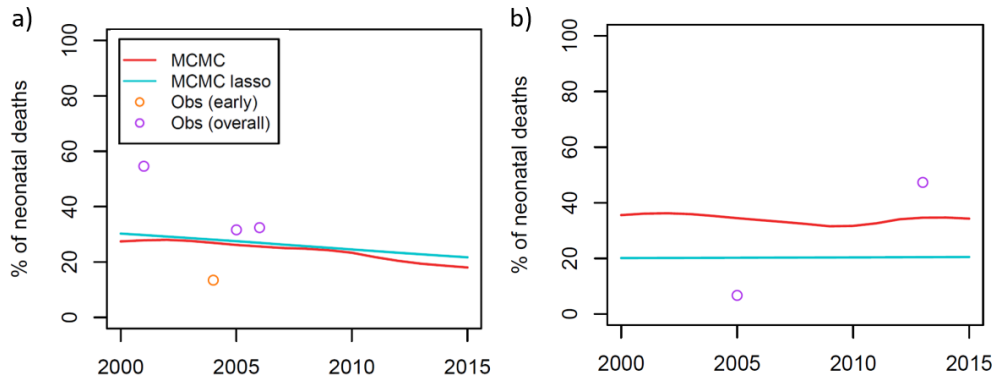
Table 7.7: Comparison of covariates (and their parameter estimates) selected using the classical approach versus two Bayesian lasso models with medium-range penalties

	Classical[1]	Bayesian lasso[2] λ=1000	Bayesian lasso[2] λ=1500
Preterm			
	Constant: 0.08	Constant: 0.27	Constant: 0.33
	GFR: -0.50	GFR: -0.08	
	LBW: 0.10		
		SBA: 0.16	SBA: 0.15
		U5MR: 0.02	U5MR: 0.01
Congenital			
	Constant: -1.69	Constant: -1.35	Constant: -1.23
	U5MR: -1.21	U5MR: -0.61	U5MR: -0.44
		DPT: 0.05	DPT: 0.02
		NMR: -0.01	
		BCG: 0.01	
Infection			
	Constant: -0.08	Constant: -0.05	Constant: -0.05
	BCG: 0.47	BCG: 0.27	BCG: 0.21
	PAB: 0.29	PAB: 0.09	PAB: 0.01
	LBW: 0.45	LBW: 0.01	
	GFR: 0.54		
		SBA: -0.51	SBA: -0.48
Other			
	Constant: -2.03	Constant: -1.62	Constant: -1.48
	PAB: 1.36	PAB: 0.63	PAB: 0.39
		BCG: 0.01	BCG: 0.01
		LBW: 0.01	
		U5MR: -0.01	

[1] Covariates were scaled to be comparable with the Bayesian lasso approach;
[2] Coefficients below 0.01 were considered to be 0 since coefficients in Bayesian lasso cannot go to exactly 0.

The COD predictions produced by the Bayesian lasso model varied in their similarity to the standard MCMC model presented in section 7.3.1. For example, some lasso results were relatively similar to their standard counterparts (e.g. Figures 7.14a) while others had fairly different levels (e.g. 7.14b) for a medium lasso penalty of 1000. This is unsurprising given that the covariates selected by Bayesian lasso were similar but not the same. The neonatal COD distributions were similar for lasso penalties of 1000 versus 1500.

Figure 7.14: Examples of the % of neonatal deaths by country and cause between the MCMC and MCMC lasso models. a) Pakistan/infection, b) Ethiopia/intrapartum



In general, the proportional COD distribution did not have substantial differences between the lasso and classical covariate selection approaches at the regional level (Table 7.8), with most differences within 5 percentage points. The most substantial difference was for intrapartum-related deaths in Sub-Saharan Africa, where the percentage of neonatal deaths predicted using the Bayesian lasso versus classical covariate selection approach were 24% versus 37%, respectively. Since intrapartum is the baseline cause, it is difficult to identify a specific reason (e.g. covariate) that may be leading to this difference. The percentage differences across the other four causes suggest that the Bayesian lasso had a higher estimate for these causes by two to five percentage points each, which likely decreased the intrapartum proportion.

Table 7.8: Neonatal proportional cause-of-death distribution by region using Bayesian lasso versus classical covariate selection methods (differences ≥ 5 percentage points italicized)

	Preterm		Intrapartum		Infection		Congenital		Other	
Region[1]	**Lasso**	**Classic**	**Lasso**	**Classic**	**Lasso**	**Classic**	**Lasso**	**Classic**	**Lasso**	**Classic**
CCA	0.41	0.38	0.27	0.26	0.14	0.13	0.11	0.16	0.07	0.07
LAC	0.38	0.36	0.25	0.29	0.20	0.16	0.10	0.14	0.07	0.05
NA	0.41	0.37	0.26	0.24	0.15	0.14	*0.11*	*0.19*	0.07	0.06
SEA	0.40	0.43	0.26	0.22	0.17	0.12	*0.10*	*0.18*	0.07	0.04
SA	*0.37*	*0.44*	0.24	0.26	0.22	0.19	0.10	0.07	0.07	0.04
SSA	0.36	0.31	*0.24*	*0.37*	0.24	0.20	0.10	0.07	0.07	0.05
WA	0.38	0.34	*0.25*	*0.33*	*0.20*	*0.14*	0.10	0.14	0.07	0.04

[1] CCA = Caucasus and Central Asia; LAC = Latin America and Caribbean; NA = North Africa; SEA = Southeast Asia; SA = Southern Asia; SSA = Sub-Saharan Africa; WA = Western Asia

The results in this subsection indicate that we were able to successfully implement Bayesian lasso in our model. The final estimates were remarkably similar given that the classical version is based on single-cause (instead of multinomial) covariate selection without lasso. The next

section includes details on how the Bayesian lasso implementation can be furthered for inclusion in a future model.

7.4 Discussion

In this chapter, I presented the results of our work on transitioning from the classical neonatal cause-of-death models in Stata to a Bayesian framework in R. We were able to successfully reproduce the classical results, and to extend the multinomial model by implementing country-level random effects. We also tested the Bayesian lasso for covariate selection. This proof-of-concept work suggests that shifting to a Bayesian framework is feasible for these models and allows us to add features that we were unable to implement in the existing classical models (e.g. random effects).

A key finding was that the Bayesian multinomial model could reproduce the classical multinomial results assuming weakly informative parameter priors and the same covariates in the cause equations. Bayesian covariate selection using lasso regression with a medium penalty could also produce estimates which were relatively similar to the classical results. Additionally, we could tune the strength of the random effects by changing the prior distribution for the between-country variation (by varying the standard deviation of the prior). We found that the goodness-of-fit of the proportional COD distribution improved with inclusion of random effects. This is not surprising since the expected outcome of incorporating random effects in our model is to pull the modelled estimates closer to the observed data.

This result also largely explains the differences in the goodness-of-fit comparison described by IHME for our previous work [202]. There, IHME compared their modelled estimates (using mixed effects [203]) against their input data versus our modelled estimates (with no mixed effects) against our input data. This is an unequal comparison, and including hierarchical effects in such validation can mask whether the fixed components of the models are performing well. This also leads to the broader issue of the interpretation and use of random effects in models such as these.

In theory, one includes random effects because of a belief that while some variations in outcomes can be explained by factors that influence all groups in the same way, some real variation (as opposed to stochastic variation) cannot be explained by measured covariates. Thus, applying country-level random effects is justifiable as an acknowledgement that there may be real between-country differences which the other covariates in our model are unable

to capture. But this interpretation of what country-level random effects represent does not capture the full picture when dealing with real-world data such as those in our input dataset. We have demonstrated in this chapter that allowing a strong random effect can pull a country's modelled estimate substantially towards the observed data. At first glance this seems reasonable, especially given our ultimate goal of shifting from modelled estimates to empirical data when possible.

But several important data issues complicate this, particularly in the high mortality model. First, the input studies in this model vary widely in quality and generally use VA methods for COD attribution. VA studies are often the only viable option for COD determination in many low-resource settings but have important biases which can influence their results (section 2.3.2). Additionally, most of the high mortality input studies are cross-sectional and report data for a short period of time (as opposed to our modelled estimates over 15+ years). Finally, the majority of input studies are local studies with relatively small sample sizes. These factors can affect the quality and generalizability of the observed data.

Given these issues, there is a question about whether the unexplained variation identified between countries from the input studies and modelled by the random effects represents real, epidemiological between-country differences or is due to data quality issues. Or in other words, are the random effects capturing true between-country variation or methodological challenges/differences between studies? Additionally, how much should one (often small and several years old) input study affect the current estimates for a country? We can try to mitigate the generalisability issue by limiting the use of random effects to those derived from nationally representative studies and/or newer studies, but these are still limited in number and may range widely in quality. Thus, a subjective decision must be made on how much a country's modelled estimate should be pulled towards what is often at present a single nationally representative study datapoint with quality concerns. Modelled estimates and empirical data both have a range of issues, and there is a delicate balance to be struck on how much weight to give to each one.

Moreover, this concern carries over to the fixed component of the ME model as well. We demonstrated that when there is a strong random effect, the estimate from the MCMC FE model can be substantially different from that of the MCMC RE model (and thus the empirical data as well). But there is a similar question of whether the difference between this MCMC FE estimate and the empirical data represents a real difference in the cause distribution or is due

to errors (random or systematic) in the empirical data. This is an especially important issue to consider since the majority of countries in our models have no nationally representative input data (and therefore the fixed effect component of the ME model would be used for their estimate).

An additional element is how deaths are treated within an observation (i.e. study or VR datapoint). Without any weighting scheme, all input deaths are treated independently and given the same weighting in the model. However, this is likely to be incorrect since it is probable that there is "clustering" of causes of deaths within observations. For the classical multinomial models, we applied an ad hoc fix such that individual deaths from larger observations carried less weight than those from smaller ones. Random effects act similarly by assuming clustering of causes of deaths within observations. The strength of the random effects influence how much "clustering" is allowed, and thus this issue should also be considered when selecting the strength of the random effects.

Ultimately, addressing these issues in a mixed effects model comes down to the strength of the random effects. This requires careful consideration and subjective decisions based on the issues discussed above, as well as the overall goal of the project. In our case, we sought to include country-level random effects in acknowledgement that modelled estimates are non-ideal and are unlikely to capture all variation between countries, and also because we believe in the principle of shifting towards empirical data when possible. However, the data quality concerns of the input studies (whether as individual studies or when combined into one dataset for analysis [section 5.1]) are real and need to be considered as well. Given these issues with both the modelled and empirical data, I would recommend a medium random effects strength, which serves as a compromise between the modelled estimate and empirical data for countries with input studies. This strength could also be adjusted over time (e.g. if input data increase in standardization/quality).

The work presented in this chapter shows that it is possible to implement a Bayesian multinomial neonatal COD model. Taking the model beyond this stage involves additional model building and analysis steps. First, a way to deal with unreported causes in the input data needs to be developed and implemented so that all studies, including those that do not report on all causes in our model, can be included in the analysis[2]. Importantly, any model used to

[2] As mentioned earlier, Dr. David Prieto-Merino has now developed a method to account for missing causes for our models in the Bayesian framework.

produce estimates must also include credible intervals. Though we did not do this at this stage, it is a necessity for actual estimates since uncertainty is one of the most important concerns for such modelling exercises. This can be done by taking a sample of n iterations (e.g. n = 1000) after convergence of the MCMC chains and estimating the COD distribution for each of these iterations. The iterations should be taken with sufficient spacing to avoid autocorrelation. The 95% credible interval can then be computed from these n samples for each country/year/cause. The models should also be tested for predictive accuracy and stability, particularly once all of the model features are implemented (see sections 5.1 and 8.6 for more details).

Further work on Bayesian covariate selection would also be useful. We tested only the Bayesian lasso as a more complete review of covariate selection was beyond the scope of this thesis. We chose the Bayesian lasso because our initial review of the literature suggested that the lasso was a good fit both for multinomial covariate selection and for the potential stability issues in our models (section 5.1). However, other methods, including Bayesian ridge regression and various information criterion, may also be useful to investigate. Additionally, we implemented Bayesian lasso by testing a range of lambda values (i.e. the penalty term), which is a common method for tuning the penalty. However, selecting the lambda parameter using out-of-sample cross-validation is more robust (though it will also substantially increase the time required to run the model). The final Bayesian covariate selection method will also need to be implemented in the mixed effects model.

The work I have presented here on shifting our classical models to a Bayesian framework fits an increasing trend on these topics. Over the last decade, Bayesian methods have become more popular in global health, including for mortality and COD estimation. For example, UN-IGME now uses Bayesian methods for neonatal, under-5, and maternal all-cause mortality estimates [204-206]. IHME shifted to using Bayesian methods for their 2010 Global Burden of Disease estimates for 235 causes of death in 187 countries, including for the neonatal age group [203, 207]. Some key differences between the IHME and MCEE approaches for COD estimation, especially regarding the modelling approaches and input data, are described in sections 2.3.4 and 8.2.

Based on the overall results of this exercise, I believe that shifting the neonatal COD models to the Bayesian framework is feasible and appropriate. The inclusion of country-level random effects is an important modelling feature which we were unable to include in the classical models. In addition, we believe that the estimates are unlikely to have substantial unintentional

(or inexplicable) differences from the classical approach given that we were able to recreate the classical results, and that we demonstrated the ability to tune the random effects strength and the Bayesian lasso penalty. Thus, the Bayesian neonatal COD framework in this proof-of-concept exercise offers additional flexibility without many drawbacks. The key disadvantage of this approach is that it can require substantially greater computing resources and time, especially with the addition of any out-of-sample cross-validation. The remaining tasks to finalize a model that can be used to produce publishable national-level COD estimates will require some time but are doable. I therefore recommend that the neonatal COD models be shifted to the Bayesian framework.

Theme 3: Discussion and recommendations

8 Discussion

The overall goal of this thesis was to help improve our understanding of the temporal and causal distributions of deaths within the neonatal period, including both conducting analyses to fill knowledge gaps and making improvements to existing statistical models. Here, I provide a summary of our findings, a comparison with other work, a discussion of overall strengths and limitations, and an overview of future work. I also provide recommendations on various topics related to neonatal estimation modelling, including on input data improvements, core modelling principles, and uses/limits of such models.

8.1 Summary of main findings

The main findings are divided into three parts: 1) estimating risk of death by day within the neonatal period (chapter 3), 2) estimating cause-of-death distributions for the early and late neonatal periods (chapter 4), and 3) identifying limitations of the current cause-of-death modelling approach and investigating model improvements (chapters 5-7).

8.1.1 Estimating risk of death by day within the neonatal period

The days immediately after birth are the riskiest for human survival, yet neonatal mortality risks are generally not reported by day. Early neonatal deaths are sometimes under-reported or might be misclassified by day of death or as stillbirths. We modelled daily neonatal mortality risk and estimated the proportion of deaths on the day of birth and in week 1 for 186 countries in 2013. For 57 countries with high-quality vital registration data, we used the data as reported. For the remaining 129 countries, we applied an exponential model to data from 206 Demographic and Health Surveys (DHS) in 79 countries to estimate the proportions of neonatal deaths per day and used bootstrap sampling to develop uncertainty estimates.

We found that 36% (uncertainty range: 34-38%) of all neonatal deaths occurred on the day of birth (day 0) and that 73% (72-74%) occurred in the first week. Thus, in 2013, an estimated 1.00 million (0.94 million-1.05 million) neonatal deaths occurred on day 0 and 2.02 million (1.99 million-2.05 million) occurred in week 1. Strikingly, these proportions had little variation by neonatal mortality rate, income, or region. Of all neonatal deaths, Sub-Saharan Africa had the highest risk of neonatal death and, therefore, had the highest risk of death on day 0 (11.2 [10.6-11.8] per 1,000 livebirths); the highest number of deaths on day 0 was seen in southern Asia

(n=392,300 [369,100-412,500]). We also developed simple analytical methods to identify DHS where substantial misclassification of deaths between days 0 and 1 or underreporting of very early neonatal deaths was likely to have occurred.

These results represent a starting point for understanding the burden of day 0 and week 1 deaths in each country. On average, our results likely represent the day 0 and week 1 proportions for many of the modelled countries, but our approach will mask any variation that does exist between countries. We hope that our estimates will be improved upon as better data become available, including with external validation if high-quality day-of-death data are made available.

Our results highlight the fact that the risk of early neonatal death is very high across a range of countries and contexts. Timely, cost-effective, and feasible interventions to improve neonatal and maternity care could save many lives if implemented during these riskiest hours and days after birth.

8.1.2 Estimating cause-of-death distributions for the early and late neonatal periods

Understanding the cause-of-death (COD) distribution is important for selecting appropriate interventions to reduce mortality and morbidity. Both available data and our understanding of biology and pathology suggest that the COD distribution differs substantially between the early (days 0-6) and late (days 7-27) neonatal periods. Our work furthered the previous neonatal COD modelling by producing separate COD distribution estimates by neonatal period. We estimated these early and late neonatal period COD distributions for 194 countries between 2000 and 2013. For 65 countries with high-quality vital registration (VR) data, we used each country's observed proportional cause distributions as reported. For the remaining 129 countries, we used multinomial logistic models to estimate these distributions.

We found that the neonatal COD distribution differed between the early and late periods and varied with neonatal mortality rate level. Although preterm, intrapartum, and infection were the leading cause categories of neonatal death in both the early and late neonatal periods, their distributions were substantially different between the two periods. Preterm birth (41%) and intrapartum complications (27%) accounted for most early neonatal deaths while infections caused nearly half of late neonatal deaths. Generally, low mortality countries had higher proportions of deaths from congenital disorders and lower proportions from intrapartum and infections, while the opposite was true in high mortality settings. Between 2000 and 2013,

neonatal deaths decreased for most causes. Of the approximately 2.8 million neonatal deaths in 2013, 0.99 (0.70–1.31) million deaths were estimated to be caused by preterm birth complications, 0.64 (0.46-0.84) million by intrapartum complications and 0.43 (0.22-0.66) million by sepsis and other severe infections. Preterm birth complications were the leading cause of death in all regions of the world.

As with all such modelling exercises, our estimates should be viewed as an interim measure to help policymakers, particularly in settings with little or no data currently. Such results are not a panacea for actual data collection, and we applaud the many efforts underway currently to improve CRVS systems and other COD data collection throughout the world. However, given the current lack of ubiquity in adequate CRVS systems, statistical modelling remains necessary to estimate COD distributions for the majority of countries.

The results from this work reinforce the idea that COD distributions are different across the early and late neonatal periods. Thus, both timing and causes of neonatal deaths should be considered carefully when implementing interventions.

8.1.3 Improving the current neonatal cause-of-death models

The work on improving the existing neonatal COD models had two key steps: 1) identifying potential issues and limitations with our models and 2) investigating and implementing potential strategies to address these concerns. I focused on model stability and predictive accuracy, as well as whether and how to weight empirical data for country-specific COD estimates.

First, I identified several factors which could contribute to model performance issues. Some of these, such as the amount and quality of the input data, are outside of our direct control. Others, such as the covariate selection method and statistical framework, are modelling choices we made and can refine to improve model performance. I demonstrated that certain strategies, such as ensemble methods, may be able to improve stability. The lasso regression appeared to be a promising technique for multinomial covariate selection and stability improvement, but we were unable to fully implement it in our classical framework. Second, I discussed practical and philosophical reasons for adding country-level random effects to our models. However, we were unable to incorporate these into our existing classical multinomial models.

We tested two key alternative modelling strategies to help address some of the issues described above. First, we implemented the COD models in a binomial modelling framework, including with random effects. However, we found that this framework did not have a substantial impact on improving model performance. We also encountered some challenges while implementing the random effects which suggested we may be faced with further difficulties if increasing the complexity of the modelling strategy (e.g. with lasso regression or ensemble methods).

Next, we successfully implemented a proof-of-concept Bayesian model with country-level random effects, and demonstrated the feasibility of applying Bayesian lasso as a multinomial covariate selection strategy within this framework. We were able to recreate the classical neonatal COD results in the Bayesian framework with weakly informative priors, indicating that shifting to this model would not have unintended or unexplainable differences from the original models. I recommended applying a medium random effects strength at the country level based both on our analyses of testing weak, medium, and strong random effect priors, as well as the nuances around what random effects mean for models such as ours (section 7.4). Finally, we demonstrated a successful implementation of the Bayesian lasso. Ultimately, I believe that shifting the neonatal COD models to the Bayesian framework is feasible and advisable since it allows us to incorporate modelling features which we were unable to implement in the classical framework. In section 8.4, I discuss some future work to finalize these proof-of-concept models.

There were many lessons learned about statistical modelling and neonatal estimates while conducting the work presented in this thesis. I have included a discussion of implications and recommendations later in this chapter.

8.2 Comparison with other work

We chose to conduct the risk by day of death and COD by neonatal period analyses as these were gaps that we believed needed to be addressed at the time. For the risk by day of death analysis, individual studies had been conducted in various countries, but no systematic, nationally comparable estimates had yet been published when we conducted our analysis. We are still unaware of comparable work. A brief history of neonatal COD estimation is included in section 2.3.4. Our group (under the Maternal Child Epidemiology Estimation [MCEE] umbrella) and the Institute for Health Metrics and Evaluation (IHME) both produce nationally comparable neonatal COD. MCEE (previously CHERG) first published neonatal COD estimates in 2005 [13], but these were not separated into the early and late neonatal periods prior to the work presented in this thesis [141]. IHME first published COD estimates by neonatal period in 2012

[93], with consistent publication of period-specific neonatal COD estimates beginning in their GBD study published in 2017 [103]. Here, I briefly describe how our approach compares to that of IHME.

Previous work has compared the under-5 COD mortality estimates between MCEE (including our neonatal estimates) and IHME [208]. These results showed that IHME and MCEE neonatal COD estimates were similar for several causes, including intrapartum, sepsis, pneumonia, and congenital. The two key causes with differences were complications of preterm birth and other. The 2013 global COD estimates for preterm were 0.965 (0.615-1.537) million by MCEE and 0.693 (0.554-0.854) million by IHME, and the "other" estimates were 0.232 (0.145-0.373) million by MCEE and 0.452 (uncertainty not available) million by IHME [208]. When estimates are different between the two groups, it is difficult to tease apart the reasons because there are several major methodological and input data differences between the IHME and MCEE models.

IHME models causes of death for the neonatal age group as part of their larger global burden of disease modelling strategy. This involves modelling 250+ causes of death using an ensemble approach within a Bayesian mixed effects single-cause framework [203]. While the different statistical approaches will certainly have an impact on the results, the input data decisions are likely to play a larger role in explaining the MCEE versus IHME differences. One major difference is that IHME combines all of the input data into one model, whereas we separate our models into two categories (low and high mortality models). We do this for two key reasons: the COD distributions are different for low versus high mortality countries, and the amount of data from low mortality countries (through VR data) is so much greater that it would vastly overpower the high mortality input study data.

A consequence of this modelling difference is that IHME models over 250 causes while we model eight major cause categories. Their use of low mortality VR inputs for high mortality countries allows for ICD-level cause categories. Producing estimates by finer cause categories would be desirable if the input data supported it. However, almost no high mortality countries have reliable VR data, and verbal autopsy (VA) studies report far fewer cause categories (often even less than eight). Thus, estimating hundreds of causes for countries with no relevant supporting data means drawing the inputs mostly from low mortality countries which tend to have substantially different COD distributions. This conveys an unwarranted level of certainty about COD modelling. The reason the imperfect solution of verbal autopsies even exists is because many countries lack reliable CRVS systems. While the data issues with VA are well known, they

are at present one of our only ways to obtain empirical COD data without CRVS systems. These studies only reporting a handful of death categories is also an implicit acknowledgement of how much uncertainty there is around COD assignment without conventional autopsies or well-conducted medical certification. To go from such empirical data to modelling hundreds of causes seems beyond what the data can support. Even hierarchical models that give weight to neighbouring and/or regional data in the absence of input data from a given country cannot solve the problem that many countries lacking adequate VR data are clustered (e.g. Sub-Saharan African countries).

Ultimately, the differences in the results between MCEE and IHME are indicative of the central issue with modelling, and the reason it is done: we lack reliable empirical data. Modelling decisions around input data inclusion/exclusion and processing, and statistical method choices, will influence over the results. As data inputs improve in quantity and quality, modelled estimates between different groups will ideally converge. I discuss several of these issues in later subsections, including improving input data (section 8.5), core principles for modelling (section 8.6), and uses and limitations of models (section 8.7).

8.3 Strengths and limitations

The work presented in this thesis has several strengths and limitations, most of which are described in detail in the previous chapters. In this section, I summarize some of the key overall strengths and limitations of our modelling approaches, with a particular focus on the cause-of-death estimation work since it constitutes the majority of this thesis.

8.3.1 Strengths

A key strength of our modelling work is that it is part of a larger effort that has been underway for over fifteen years to improve our understanding of the neonatal mortality burden and associated issues such as stillbirths, preterm birth, and morbidity. As such, the work presented in this thesis has benefited from the knowledge, data, and modelling processes that have been built up over time through this wider effort. This includes a strong expertise and institutional memory of various data-related issues, including around specific input studies, nuances around timing and causes of death (including mapping of ICD codes to broader cause categories and understanding implausible model outputs), and reasons for historical modelling decisions. The neonatal COD database began in 2003, and has been added to with updated literature reviews (such as those conducted for the work presented here). Additionally, this long-standing effort

allowed us to use networks to obtain additional input data for inclusion in our models, such as finer-grained COD data than were included in publications.

Our neonatal COD estimation work falls under the CHERG/MCEE umbrella. This means that we have collaborations with several research partners (e.g. Johns Hopkins Bloomberg School of Public Health, University of Edinburgh) and multilateral organizations (e.g. WHO, UNICEF). This provides us with the opportunity of formally presenting our work one to two times per year to individuals with widespread areas of expertise (e.g. technical, policy) and viewpoints (e.g. researchers, policymakers, donors). Thus, the work presented in this thesis was not done within a research silo, but rather as part of an interactive process with feedback from a variety of perspectives. This will ideally make the results of our work more useful and credible to policymakers and others who need such estimates for decision making. Through this collaborative process, for example, it became clear that there was a desire from countries to have country-specific data play a larger role in their COD predictions.

The neonatal estimates we produce are part of the UN system's official cause-of-death estimates, and are published by the WHO and UNICEF. This increases the likelihood of impact, and has allowed us to engage outside of the research community. An important benefit is that our estimates go through the WHO's formal country consultation process. This process involves providing the preliminary estimates to health ministries of all WHO member countries for feedback. Through this, we have received feedback from several countries. Typically, these have been countries with medium-to-high quality VR data. However, the mechanism for countries to feedback to us before we finalize our estimates is an important one, and one which could become even more valuable as countries enhance their data collection systems. More generally, the longstanding nature of this work and the UN collaborations have meant that this work is not a one-off, but is part of an ongoing process of trying to enhance and improve child cause-of-death modelling.

Additionally, this connection with the WHO has allowed us to directly feedback to their teams working on covariate time series updates. This resulted in a useful back-and-forth through which they corrected (or smoothed) various time series issues which we identified (e.g. unrealistic spikes in otherwise stable covariates).

We have aimed to make our data and code transparent and available. To do this, I revised our code to make it user-friendly; our full models can be run to reproduce the published results by

"one-click" once the statistical software (Stata) and code are downloaded. To improve accessibility, our data inputs and modelling files are hosted on the WHO's Global Health Observatory (https://www.who.int/gho/). We will follow the same transparency processes for the revised modelling strategy proposed in this thesis, which has the added advantage of being written in the free software programme R. We also believe strongly in the Guidelines for Accurate Transparent Health Estimates Reporting (GATHER) [209] and have strived to adhere to them for our work.

8.3.2 Limitations

A key limitation for our modelling work is the quality and quantity of input data. I have discussed this in detail throughout the thesis (particularly in chapters 2 and 5). This is a limitation but also the very reason for the work we have done; the availability and/or quality of the data needed for relevant decision-making is still lacking in many countries. Thus, in many cases modelling remains the only viable option for such estimates (see section 2.3 for limitations of other options such as hospital death records). We tried to mitigate data quality issues where possible. In the risk by day of death analysis, we sought to correct for day-of-death misclassification with our model, and also developed simple analytical methods to identify some types of misclassification. In the neonatal COD death analysis, a key goal was to identify data issues and build robustness into our models. While the lack of sufficient empirical data also meant external validation of our models was not feasible, we conducted internal validation to assess model performance.

Our models are somewhat limited in scope because they mostly produce estimates at the national level or higher. This fails to capture variation within countries. However, developing subnational estimates is still challenging because the data quantity and quality are typically even more limited than at the national level. India is the only country for which we have produced state-level estimates [210]. Expanding the set of countries with reliable subnational modelled estimates may become feasible if more high-quality subnationally representative data are available. However, maintaining consistent and updated subnational input data (covariates and outcomes) for several countries will be challenging and time consuming. This is because, unlike national covariate time series, such data are not yet standardized, collated, and available for immediate use.

A couple additional points about the type of COD modelling we have presented are worth highlighting. First, one issue is the attribution of death to a single cause. This does not allow for co-morbidities, which are a frequent occurrence in neonatal deaths, and may thus

underestimate the impact of a given cause. At present, the majority of input studies and VR countries only provide data with deaths attributed to one cause, and thus we have little choice over this for our modelling. Second, there is the risk that modelled time series estimates can be interpreted as "tracking" changes in causes of death. As noted earlier in the thesis, our estimates are predictions of what might be occurring in countries. To track changes in burden due to specific causes of death requires each country to collect representative and consistent data on cause of death on a continuing basis. Our estimates are not a panacea for actual data collection.

The work in this thesis focuses almost entirely on neonatal mortality. Yet, neonates are part of a broader continuum, from pregnancy to birth to childhood and beyond. In particular, maternal health and complications (especially around the time of birth) and causes of intrapartum stillbirths are closely related to neonatal mortality. This is a well understood continuum [211, 212], and integrated service delivery and health packages have been designed with this in mind [213, 214]. The WHO perinatal death certificate and the recent introduction of the Application of ICD-10 to Deaths during the Perinatal Period (ICD-PM) classification system constitute efforts to better capture deaths within the continuum of antepartum stillbirths, intrapartum stillbirths, and the neonatal period (with inclusion of information about contributing maternal conditions) [215]. However, incorporating this idea into the COD modelling strategy is challenging for a number of reasons, including inadequate input data with relevant joint information (e.g. maternal and neonatal outcomes/co-morbidities).

Finally, we have tried to incorporate or account for uncertainty in various ways throughout the modelling process. Aside from publishing estimates with uncertainty intervals, we investigated ways to further account for uncertainty in our models. For example, we tested how to account for uncertainty in the covariate selection process (e.g. through simple model averaging) and the input dataset (e.g. through bagging). However, there are sources of uncertainty which we have not yet addressed, including in the covariate time series data. The covariate time series data are provided to us with no uncertainty intervals, and we use these point estimates as-is in our input and prediction datasets. However, modelling strategies can be used such as perturbing the covariate data which can help to better account for such uncertainty. Carefully thinking through all sources of input and modelling uncertainty is important, and statistical approaches can be used when the data provided lack uncertainty intervals. I discuss this more in sections 8.4 and 8.6.

Overall, the limitations described here are common to this type of modelling. One positive development is that the quantity and quality of relevant data are increasing, which can help improve the models (and ideally, eventually replace them). Later in this chapter, I describe some efforts underway by others to improve the types of data we use in our models (section 8.5) and discuss uses and limits of these types of models (section 8.7).

8.4 Future work on neonatal cause-of-death estimation

Here, I discuss some possible future work, including steps to finalize the Bayesian neonatal COD models described in chapter 7, addressing a few remaining methodological issues, and suggestions for analyses to help improve our understanding of variations in neonatal COD distributions.

Taking the Bayesian neonatal COD models beyond the proof-of-concept stage requires some additional work. As noted in chapter 7, a way to deal with unreported causes in the input data within the Bayesian framework has already been developed by a team member. I also discussed a method for deriving credible intervals (section 7.4), which are essential for publishable estimates. While I demonstrated the feasibility of the Bayesian lasso, the covariate selection method needs more work. Key steps include implementing cross-validation to select the lambda penalty value with the best out-of-sample goodness-of-fit and implementing the Bayesian lasso within the mixed effects framework. Additionally, a broader review of Bayesian covariate selection could yield other alternative strategies for further testing. Once the core modelling strategy is finalized, model performance should be evaluated using internal cross-validation (such as described in section 5.1). Alternative covariate selection methods could also be evaluated against the lasso results for predictive accuracy. Finally, methods to improve model performance (e.g. ensemble methods) could be considered for implementation based on the results of the model performance exercise.

A few methodological issues could also be addressed by future work. First, there are no data from modelled countries in the low mortality input dataset at present. Lower-quality VR data from modelled countries should be considered for inclusion in the input dataset for the Bayesian mixed effects model. This would allow a country's estimates in the low mortality model to also be influenced by the country's own data (via a random effect), as is proposed for nationally representative VA data in the high mortality model. These VR data could be included by lowering the quality threshold currently used to select high-quality VR countries (appendix B.1.1). For example, we currently include a criterion of ≥80% CRVS coverage for high-quality VR

data which could be reduced (e.g. ≥50% CRVS coverage). Retaining a quality threshold provides a balance between the desire for inclusion of relevant data with concerns about their lower quality and reliability. Second, some of the countries for which we produce modelled estimates are improving their CRVS systems, while others are newly introducing such systems. One question is how the time series should be compiled for the years before the data are available. Plausible options are to perform regression-based imputations or somehow incorporate the modelled estimates for those years. Third, ways to better build accurate amounts of uncertainty into the modelling process could be investigated. As mentioned in section 8.3, we have not accounted for all relevant sources of uncertainty (e.g. uncertainty in covariate values). Methods like perturbing the data or drawing covariate values from an assumed distribution may be useful to investigate. How such methods would work within the full modelling framework would be important to consider so that the number of total iterations and time needed to run the model do not become infeasible. Finally, better ways to communicate uncertainty in the final results are needed. Publishing numbers with uncertainty intervals has been insufficient because it is too easy to focus only on the point estimate. Innovative data visualization solutions are emerging [216], including maps which make it impossible to ignore uncertainty by using colour grids which integrate uncertainty with final estimates [217]. While not essential, addressing these types of methodological issues can enhance the neonatal COD models.

On a broader level, there is substantial variation in the neonatal COD distribution across countries, even amongst countries with high-quality VR. This is true for all of the major causes, including direct complications of preterm birth (one of the leading causes of neonatal death). For instance, amongst high-quality VR countries with 100 or more deaths in 2015, the proportion of deaths from preterm complications ranged from 20% (Japan) to 71% (Macedonia) [218]. Such variation appears to be greater than could be plausibly attributed to epidemiological variation, and thus likely reflects at least some variations in coding practices.

Death certificate data in which multiple causes of death are recorded provide an opportunity to better understand these differences across countries. This is important for two main reasons. First, given the high burden of neonatal deaths in many countries, a clearer picture is needed for why countries have reported such wide differences in proportions. Are there genuine differences in causes of death, or are at least some of the differences an artefact of coding practices? Such an analysis could also investigate whether countries are adhering to the relevant ICD-10 coding rules. For those that are not, the results can help guide them on where their coding practices are not in line with the rules. This is useful because correctly determining

the sequence of causes related to a death, from immediate to underlying, can help target interventions more appropriately. Second, VA studies have also reported wide ranges for the proportion of deaths from the main neonatal causes of death. VA data quality varies between studies, but is generally worse than high-quality VR data. By carefully evaluating VR data, crude but useful metrics could be developed for gauging the plausibility of VA data. For example, a reasonable lower bound could be calculated for absolute risk of neonatal death from direct complications of preterm birth based on VR data from low mortality countries. This could serve as a way to gauge the plausibility of this aspect of VA data, and may help study investigators choose appropriate causal hierarchies when analysing their VA data.

8.5 How can "input" data be improved?

Input data are arguably the most important component of a model and can have a substantial impact on the reliability of model outputs. In our case, input data have meant data on national covariate time series and neonatal outcomes (i.e. COD distributions and timing of deaths). There are two types of input-related improvements relevant for our work: 1) those relating directly to the quantity and quality of individual datasets (e.g. data from a CRVS system or study) and 2) those related to the aggregating of multiple such datasets for use in models like ours. What I call "input" data throughout this thesis are of course just data; they were not primarily collected for use in our models. This means both that we have little to no direct influence over them and that they are designed to be used for purposes different than ours. With this in mind, I briefly discuss some key topics related to input data improvements.

First, major efforts are currently underway throughout the health and development sectors to improve each of the main data sources we use for our models. For example, momentum to introduce or improve CRVS systems in countries has been increasing. Between 2013 and 2018, nine countries moved from being modelled in our estimates to having high-quality VR that could be used as reported. Various programmes exist to scale up CRVS systems in the coming decade across many countries [27], including a joint WHO/World Bank global CRVS investment plan [28] and the Bloomberg Data for Health Initiative [27]. Similarly, more countries and organizations are investing in large nationally representative VA studies in countries without adequate CRVS system. A number of improvements to VA methods have been proposed and tested, including standardization of VA tools [67, 73, 219] and probabilistic algorithms for assigning deaths [73, 75]. Additionally, many discussions and proposals have been made around selecting and improving child and maternal health-related covariates as the SDG era has begun [220-222]. None of these efforts are simple or short-term challenges; at minimum they require substantial

amounts of careful planning, funding, and commitments. But they are based on years of lessons learned and are promising attempts to work on these long-standing data issues. Such investments in on-the-ground data collection are essential for improving modelled estimates and to ultimately replace estimates with empirical data.

A number of innovative ideas have also been proposed that relate to these topics. For example, minimally invasive autopsy (MIA; sometimes called minimally invasive tissue sampling) is a promising alternative to conventional autopsies [84] (section 2.3.3). While unlikely to replace VA studies in the short term, the method could be coupled with at least some of them to help improve the accuracy of empirical COD data in countries lacking adequate CRVS systems. Various ideas to link different data collection methods have also been proposed, including integrating VA studies with CRVS systems [30] and linking different data systems within countries [29, 223]. Other methods, such as maternal and perinatal audits [81] and sample registration systems like the one used in India [38] also have the potential to improve data if implemented more widely.

Some improvements are particularly useful when aggregating various datasets. First, standardization is key. Certain standards already exist for various data collection methods (e.g. ICD coding for VR data [224]) and reporting of studies/results (e.g. STROBE guidelines for observational studies [225]). For neonatal causes of deaths, agreeing and adhering to specific case definitions and causal hierarchies is particularly important. For example, there still appears to be some confusion or disagreement amongst researchers about which causal hierarchies to use when assigning neonatal causes of death. Some VA studies publish results comparing multiple hierarchies [226, 227], but the majority select one (which is not always consistent between different studies). Resolving these issues will help make empirical neonatal COD data more comparable. Publishing VA data in a format that allows subsequent application of causal hierarchies is a solution that can help resolve this discrepancy while also adding transparency to the COD attribution process. At minimum, more complete reporting of the methods used, including the specific case definitions and hierarchies selected, is needed. Finally, improving the timeliness of published COD data would be useful. The lag between data collection and publication is sometimes five or more years. A potential solution is a central repository where data can be added once data collection is complete, though the well-established concerns of researchers about such a repository must be addressed for it to be successful.

The hope of input data improvement in relation to modelling is two-fold: that 1) the COD models will become more reliable as input data increases in quantity and quality, and 2) data will improve to the point that models are no longer needed.

8.6 What core principles should be followed when producing modelled estimates?

Data are expanding, demands for modelled estimates are growing, and building complex models is becoming easier with advances in statistical methods and software. Thus, this is an apt time to map out some core principles around the production of modelled estimates. While there are no perfect models, accuracy and reliability should be the main goals. This requires a carefully planned modelling strategy and a deep understanding of the input data. Equally important are the ways in which the modelled results are reported and shared. With this in mind, I include some recommendations here on core principles to follow when conducting such analyses. I have based these on lessons learned through the course of this thesis.

Note: throughout the model planning phase, it is important to revisit whether the research question being asked is a reasonable one to answer through modelling. This would include consideration of existing knowledge on the topic, availability of appropriate input data, limitations of feasible modelling approaches, and implications of the findings given any modelling limitations. Further discussion of risks of inappropriate modelling are included in section 8.7.

Carefully developing a thorough and comprehensive modelling strategy should be the first step of the modelling process. This includes thinking through the core components of the full modelling approach, including the input data, outputs, statistical methods and software, model performance, uncertainty, and reporting and sharing of results. In Table 8.1, I include some recommendations for each of these topics.

Table 8.1: Core components of a systematic strategy for producing modelled estimates

Component	Recommendations
Overall	Consider the intended audience for results; engage some of them if possible for feedback on the modelling strategy and planned outputs
	Put together an advisory panel with knowledge on the range of topics relevant to the modelling exercise (e.g. types of input data, statistical methods, understanding the outputs, dissemination)
	Develop a full initial modelling strategy before conducting analyses
Input data	Use as many data gathering methods as feasible, including literature reviews (with multiple languages if possible), grey literature searches, and contacting individuals and organizations
	Seek a deep understanding of the data sources, including possible sources of measurement error (systematic and random) to understand data biases and uncertainty
	Try to be as consistent as possible with the data (e.g. extract information using the same rules, apply the same data cleaning procedures)
	Save raw versions of the input data (i.e. create new files after data aggregation or cleaning; do not overwrite raw files)
	Document all decisions around changes to the input data and cleaning
Desired outputs	Consider whether the desired outputs/results are feasible and appropriate to produce given the input data
	Select appropriate output measure(s) depending on the data and the intended use of the results (e.g. risks versus proportions)
	Once the model is run, check outputs in detail to ensure results are sensible
Uncertainty	Consider all sources of uncertainty (e.g. in input data, statistical methods)
	Incorporate methods to account for as much uncertainty as possible
	Include uncertainty intervals in all published estimates
Statistical methods	Investigate, test, and choose an appropriate statistical model based on the type of data (e.g. data generating process), intended outputs (e.g. proportional distribution), and desired modelling features (e.g. hierarchical structure)
	Review the literature to fully understand advantages, disadvantages, and any recent advances for the selected methods
	Revise methods based on model performance checks and feedback
Model performance	Incorporate strategies for testing aspects of model performance (e.g. predictive accuracy, model instability) into the core modelling approach
	Test for model performance with whatever data are available. Use internal validation (e.g. out-of-sample cross-validation) at a minimum, but external validation if data rich
	Consider additional checks based on uncertainty in model inputs or processes (e.g. perturb input data, test results with model variations)
Software	Select software with the flexibility and/or packages needed to complete the analysis based on the desired statistical methods
	Automate as much as possible to reduce error, improve consistency, and ensure reproducibility (e.g. use scripts for even minor analyses; avoid Excel)
	Seek answers (including user-written packages) for encountered challenges through online statistical software user forums or other websites
Reporting results	Adhere to all aspects of the GATHER guidelines [209]
	Consider creative ways to report uncertainty of inputs and results (e.g. maps overlaying results with uncertainty using a colour grid [217])
	Make data and code fully accessible, publicly available, and easy to use (e.g. “one-click” automation to run code)

There are a few additional recommendations that I want to highlight based on our experiences with the modelling exercises in this thesis. First, I do not believe that modelled time series estimates such as ours should be updated every one to two years. Historically, such estimates were produced at wide intervals, with several years between estimates. In the last five years, however, updates every one to two years have become the norm. This was partially driven by IHME's decision to update their GBD results annually [228]. The CHERG/MCEE updates are generally now done every two years (though literature reviews may occur every 2-3 years instead). However, we are modelling outcomes which do not change that quickly, and our models are unlikely to be able to capture rapid changes if they did occur. Frequently updating the estimates can give the illusion of both "tracking" outcomes and of there being more certainty in the results than is true. Such assumptions are not justified and can be risky, including if they give the impression that on-the-ground data collection is no longer essential. Thus, I recommend releasing updated estimates every three to five years instead. This provides enough time for there to be real changes (e.g. in a country's COD distribution), and also serves as a reminder that these are modelled estimates and not regularly collected empirical data.

Second, it is the responsibility of modellers to provide clear explanations to non-modellers on how to use and interpret both the modelled results and the models themselves. These are complex models with many different components (e.g. range of input data of differing quality, various statistical methods), but the outputs are typically simple (e.g. COD proportions by country and year). Thus, it is easy to forget that these relatively straightforward results are outputs of a complex statistical machinery using a diverse set of input data. Uncertainty intervals are required with modelled estimates, but are not by themselves enough to make clear the uncertainty and complexity of the whole process. Therefore, modellers should publish a guidance document alongside their data and code which explains the input data and the limits of the model (see Appendix D for an example based on our neonatal COD modelling exercise). Such a document is especially important given the push to have models be made publicly available. This is a positive development and is providing much-needed transparency. But it also means that others can adapt the models for their data, even if the models are not suitable for such an analysis. For example, some have tried to use our models to produce subnational COD estimates with prediction covariate data that were well outside the range of our input data. This required extensive discussions to explain why their proposed analysis was beyond the scope of our models, at least without performing rigorous checks and validation exercises.

In general, careful consideration is needed when trying to extend models beyond their original purpose. The first question is whether the model can support the desired extension (as described in the example above). Additionally, models can sometimes be extended by doing sub-models within the model. MCEE did this for the child COD models (including ours) by adding subnational estimates for India. This required a separate process to identify, collate, and clean state-level prediction data. Thus, incorporating this into a regularly updated modelling system (currently every one to two years) requires an active effort to update not only the main prediction database each time, but this additional one as well. If more countries are added for subnational estimates, the effort required will burgeon. Just the database updating process then transforms into one with many moving parts, which increases the risk of diminishing the reliability of the outputs. Thus, the consequences of adding components to the model should be thought through as carefully as the original modelling strategy itself.

The GATHER guidelines state that authors should "provide the results of an evaluation of model performance, if done…" [209]. I would go further to suggest that testing model performance should be an essential component of such modelling exercises. Even though having the data for external validation is still uncommon in this field, internal validation using approaches like cross-validation is almost always feasible. In this thesis, I have demonstrated why testing model performance is useful, including for helping to identify where robustness is needed and can be built into the model.

Finally, I believe that such modelled estimates should not be published in academic journals unless there is a genuinely new research contribution. These estimates have been published for years in academic journals, sometimes with little change between the previous versions. "Salami slicing" is also an issue, where one large study is published in many small parts rather than as one paper. Groups that produce these estimates simultaneously host them on online platforms (e.g. WHO's Global Health Observatory, IHME's Global Health Data Exchange site); the results are thus already publicly available outside of the academic journal system. These online versions are typically more complete as well (e.g. code, input data, full country-year estimates). UN-IGME serves as an example of how nationally comparable estimates can be published. They produce an annual report (i.e. Levels and Trends in Child Mortality) about the given year's mortality estimates, and provide detailed country-specific estimates. Their academic journal articles are largely restricted to detailing methodological updates or innovations [105, 229], comparisons with other estimates [54], and larger summary papers at

key times (e.g. end of the MDGs) [206]. I believe that it should be the exception, not the rule, that regular updated estimates are published in academic journals.

8.7 What are reasonable uses and limits of such modelling?

The question of uses and limits of modelled estimates is a complex discussion whose full breadth is beyond the scope of this thesis (numerous insightful articles have been written regarding this topic [55, 230-235]). Here, I touch on two interrelated questions that are of particular relevance to the work presented in this thesis: what is the role of modelled estimates and how far can/should models be pushed to fill knowledge gaps.

There is increasing pressure to do more with existing data. For mortality estimation, this has meant modelling estimates with finer-grained categories of causes, ages, sex, and location (e.g. subnational, local area) and over longer periods of time. The main limitation for achieving such finer gradation is not a technical one. It is possible to design a model that will extrapolate from any number of data points to as fine-grained a distribution as desired. However, just because models *can* be used to produce such estimates does not mean that they *should* be. Key limitations are the availability and quality of relevant input data that are the foundation of modelled estimates.

Thus, a balance is needed between what is desired versus what is appropriate to model given various limitations (e.g. input data availability and quality). Ignoring this greatly increases the risk of false precision. Such overconfidence is particularly dangerous in a field where having enough data for external validation is rare. These issues have arisen in our work, as seen by the lack of external validation for our modelled estimates. We tried to mitigate this as best as possible through internal validation, but there are limits to how much this can prevent concerns such as overfitting. These limits become more severe as we go towards finer-grained estimates for which there are even fewer available empirical data.

There are serious potential risks to "over-modelling". First, to the extent that modelled estimates are seen by some as a substitute for data collection, excessive confidence can lead to undervaluing primary data collection. Detrimental consequences of this include reallocation of funds meant for on-the-ground data collection and reduction of in-country capacity for health statistics, both of which are thought to already be occurring [231, 232]. Second, to the extent that modelled estimates are intended to inform policy decisions, false precision can be dangerous. Reliability is likely to be low for models with very limited relevant input data,

thereby increasing the chances of inaccurate results. It is better to tell a policymaker that the desired numbers are unavailable or highly uncertain than to provide estimates which may be far from reality or describe trends that do not exist.

These issues have not gone unnoticed by health ministries and other users of modelled estimates. Some key concerns and criticisms of modelled estimates include how to interpret discrepancies between different estimates (e.g. MCEE versus IHME) or between estimates and empirical data (e.g. DHS study), difficulty in understanding the perceived "black box" nature of modelling methods (including when estimates change due to methodology or data updates), and the disconnect between national-level estimates and local-level decision-making needs [230, 231, 233]. These concerns highlight the fact that more work is needed to improve the usefulness and credibility of modelled estimates. Stretching models beyond their reasonable limits is likely to exacerbate these challenges.

Questions on the uses and limits of modelled estimates are not going away soon. In the last two decades, the field of metrics and burden estimation has grown dramatically. This is due at least in part to the Millennium Development Goals (MDGs) [55], which had 60 country-specific indicators for 21 targets across eight goals [236]. This necessitated a monitoring framework that could track progress over time for a range of metrics, including in data-poor countries. This type of monitoring framework, which lends itself naturally to nationally comparable time trend estimates, is being reinforced by the even more metrics-heavy Sustainable Development Goals (SDGs), which have 232 indicators across 169 targets for 17 goals [237].

The push for more measurement, evaluation, and progress tracking is generally a positive one, but questions remain on how best to achieve these. Modelled estimates appear well-suited to fill the gaps left by the lack of adequate empirical data, especially when needing regularly updated country-specific comparable numbers such as for the SDGs. Such estimates may seem especially attractive because, compared to ground-level data collection, they are cheaper, faster, and can yield more comprehensive sets of numbers. These advantages can make it easy to forget or overlook the fact that modelled estimates are an imperfect solution, and can even be an inappropriate one at times. For example, unlike continuously collected empirical data, current covariate-based models are not capable of "tracking". They predict estimates based on regressions and are unable to capture rapid changes on the ground or changes not captured by available covariates. However, words like "monitoring" and "surveillance" [228], alongside frequent (e.g. every 1-2 year) updates, can give the misleading impression that models are in

fact tracking progress in countries. Modelled estimates are at their best, I believe, when they are appropriate, robust, and communicated to be temporary solutions until good empirical data can replace them. In almost all cases, modelled estimates cannot and should not replace empirical data. However, if communicated properly, they can in fact play an important role in highlighting the gaps in and need for strong empirical data.

In these discussions, it is important to be clear about how the demands for and uses of modelled estimates vary at different levels and amongst different consumers of health information. At the global level, modelled estimates have frequently been used for awareness and advocacy (including to seek funding for specific conditions), as nationally comparable indicators (e.g. for the MDGs/SDGs), and as tools to help donors and organizations "track" progress for their own purposes [230]. These uses also partly drive the pressure for annual estimates and finer-grained details. Such updates can be used to maintain attention on topics of interest and to satisfy the MDG/SDG monitoring framework mentioned earlier. The fact that many international donors share these priorities helps reinforce the pressures and perceived value of producing and expanding frequently updated modelled estimates [55, 230]. In contrast, such modelled estimates have been less in demand at the national, subnational, and programme levels [238]. At these levels, where the focus is on making specific funding and policy decisions, decisionmakers have generally preferred empirical data which are easier to understand and justify (see concerns/criticisms of modelled estimates earlier in this section). The demand for and uses of modelled estimates are thus often substantially different at the global level compared to the national level or below. In the end, there is value when discussing modelled estimates in asking who desires them and for what purpose, and in being humble about their utility in different situations. The article by AbouZahr and colleagues titled "Global estimates of country health indicators: useful, unnecessary, inevitable?" [230] is a particularly eloquent summary of these issues.

Whether, when, and how modelled estimates should be used are important questions that continue to be debated in the field. Empirical data is often still so limited that models may currently serve an important purpose, including for a broad perspective on the levels of burden and risk. But this same sparsity of empirical data also limits what can be reasonably modelled. Given these limitations, sometimes modelling is not the right choice, even if it is the easier or only choice. Making these determinations should be the joint responsibility of those who produce, fund, and use models.

8.8 Overall conclusions

The work presented in this thesis constitutes a modest attempt to fill some gaps in our understanding of neonatal deaths and to improve neonatal cause-of-death modelling. The findings on daily risk of death (chapter 3) and causes of deaths in the early/late neonatal periods (chapter 4) highlight how quickly temporal and causal distributions change within the first four weeks of life. An understanding of such changes is useful for decision making, including to help prioritize implementation of the right interventions at the right time. With our deeper dive into the neonatal cause-of-death models (chapters 5-7), we were able to identify potential ways to improve the models, including to make them more robust, incorporate multinomial covariate selection, and add country-level random effects.

Beyond wanting the results themselves to be useful, I hope that the discussions included here can contribute to conversations about the role of modelling and estimation in global health. As the field of health metrics continues to grow, modelled estimates are likely to become even more common. Such estimates can play an important role in global health, especially in filling knowledge gaps where adequate empirical data are lacking. However, the desire to eliminate remaining gaps must be balanced by an understanding of what can be feasibly and appropriately addressed through modelling exercises, including an assessment of whether the available input data can support the intended models. Otherwise, modelled estimates can potentially do more harm than good, especially if they lead to the reduction of national-level capacity or ground-level data collection efforts.

Over the last two decades, the decline of neonatal mortality and the rise of health metrics have each been impressive. These are both positive developments, but with caveats. In 2018, an astounding 2.5 million babies were still estimated to have died in their first month of life [3]. More rapid progress is needed if we are to reach the Sustainable Development Goals target of 12 or fewer neonatal deaths per 1,000 livebirths for all countries by 2030 [138]. At the same time, modelled estimates – and more generally, the field of health metrics – are still finding their place between researchers, donors, policymakers, and implementers. While neither the challenges nor the solutions to any of this are simple, there is both great momentum and hope for simultaneously improving neonatal health, data collection systems, and health metrics.

9 References

[1] United Nations, "The Millennium Development Goals Report 2015," United Nations, 2015. Accessed: 22 March 2018.

[2] N. Akseer *et al.*, "Ending preventable newborn deaths in a generation," (in eng), *Int J Gynaecol Obstet,* vol. 131 Suppl 1, pp. S43-8, Oct 2015, doi: 10.1016/j.ijgo.2015.03.017.

[3] UNICEF, "Levels and Trends in Child Mortality - Report 2019. Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation," UNICEF, 2019.

[4] V. Thirani and A. Gupta, "The value of data," World Economic Forum (Available at: https://www.weforum.org/agenda/2017/09/the-value-of-data/), 2017.

[5] L. Mikkelsen *et al.*, "A global assessment of civil registration and vital statistics systems: monitoring data quality and progress," (in eng), *Lancet,* vol. 386, no. 10001, pp. 1395-1406, Oct 2015, doi: 10.1016/S0140-6736(15)60171-4.

[6] J. E. Lawn *et al.*, "Every Newborn: progress, priorities, and potential beyond survival," (in eng), *Lancet,* vol. 384, no. 9938, pp. 189-205, Jul 2014, doi: 10.1016/S0140-6736(14)60496-7.

[7] M. Molyneux and E. Molyneux, "Reaching Millennium Development Goal 4," (in eng), *Lancet Glob Health,* vol. 4, no. 3, pp. e146-7, Mar 2016, doi: 10.1016/S2214-109X(16)00009-7.

[8] G. L. Darmstadt, J. Shiffman, and J. E. Lawn, "Advancing the newborn and stillbirth global agenda: priorities for the next decade," (in eng), *Arch Dis Child,* vol. 100 Suppl 1, pp. S13-8, Feb 2015, doi: 10.1136/archdischild-2013-305557.

[9] Z. A. Bhutta *et al.*, "Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost?," (in eng), *Lancet,* vol. 384, no. 9940, pp. 347-70, Jul 2014, doi: 10.1016/S0140-6736(14)60792-3.

[10] CDC Wonder, "Linked Birth/Infant Death Records for 2007-2010 with ICD10 codes.," vol. 2014, no. July 31, 2014.

[11] V. Fauveau, "New indicator of quality of emergency obstetric and newborn care," (in eng), *Lancet,* vol. 370, no. 9595, p. 1310, Oct 2007, doi: 10.1016/S0140-6736(07)61571-2.

[12] Save the Children, *Surviving the first day: State of the World's Mothers 2013*. Save the Children, 2013.

[13] J. E. Lawn, S. Cousens, J. Zupan, and L. N. S. S. Team, "4 million neonatal deaths: when? where? why?," (in eng), *Lancet,* vol. 365, no. 9462, pp. 891-900, 2005 Mar 5-11 2005, doi: 10.1016/S0140-6736(05)71048-5.

[14] R. E. Black *et al.*, "Global, regional, and national causes of child mortality in 2008: a systematic analysis," (in eng), *Lancet,* vol. 375, no. 9730, pp. 1969-87, Jun 2010, doi: 10.1016/S0140-6736(10)60549-1.

[15] L. Liu *et al.*, "Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000," (in eng), *Lancet,* vol. 379, no. 9832, pp. 2151-61, Jun 2012, doi: 10.1016/S0140-6736(12)60560-1.

[16] J. E. Backer, "Population statistics and population registration in Norway. Part 1. The vital statistics of Norway: An historical review," *Population studies,* vol. 1, no. 2, pp. 212-226, 1947.

[17] C. AbouZahr *et al.*, "Civil registration and vital statistics: progress in the data revolution for counting and accountability," *The Lancet,* vol. 386, no. 10001, pp. 1373-1385, 2015.

[18] D. E. Phillips *et al.*, "Are well functioning civil registration and vital statistics systems associated with better health outcomes?," *The Lancet,* vol. 386, no. 10001, pp. 1386-1394, 2015.

[19] D. C. Muñoz, C. Abouzahr, and D. de Savigny, "The 'Ten CRVS Milestones' framework for understanding civil registration and vital statistics systems," *BMJ global health,* vol. 3, no. 2, p. e000673, 2018.
[20] Y. Ye, M. Wamukoya, A. Ezeh, J. B. Emina, and O. Sankoh, "Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa?," *BMC public health,* vol. 12, no. 1, p. 741, 2012.
[21] C. Rao, D. Bradshaw, and C. D. Mathers, "Improving death registration and statistics in developing countries: Lessons from sub-Saharan Africa," *Southern African Journal of Demography,* pp. 81-99, 2004.
[22] J. Joubert, C. Rao, D. Bradshaw, R. E. Dorrington, T. Vos, and A. Lopez, "Characteristics, availability and uses of vital registration and other mortality data sources in post-democracy South Africa," *Global health action,* vol. 5, no. 1, p. 19263, 2012.
[23] P. Mahapatra *et al.*, "Civil registration systems and vital statistics: successes and missed opportunities," (in eng), *Lancet,* vol. 370, no. 9599, pp. 1653-63, Nov 2007, doi: 10.1016/S0140-6736(07)61308-7.
[24] C. D. Mathers, D. M. Fat, M. Inoue, C. Rao, and A. D. Lopez, "Counting the dead and what they died from: an assessment of the global status of cause of death data," (in eng), *Bull World Health Organ,* vol. 83, no. 3, pp. 171-7, Mar 2005, doi: /S0042-96862005000300009.
[25] D. E. Phillips *et al.*, "A composite metric for assessing data on mortality and causes of death: the vital statistics performance index," (in eng), *Popul Health Metr,* vol. 12, p. 14, 2014, doi: 10.1186/1478-7954-12-14.
[26] World Health Organization, "World health statistics 2017: Monitoring health for the SDGs.," 2017.
[27] A. D. Lopez and P. W. Setel, "Better health intelligence: a new era for civil registration and vital statistics?," *BMC medicine,* vol. 13, no. 1, p. 73, 2015.
[28] World Health Organization and World Bank, "Global Civil Registration and Vital Statistics: A Scaling Up Investment Plan 2015-2024," 2014.
[29] R. Silva and C. AbouZahr, "Towards the next generation of record-linkage studies to advance data quality assessment of civil registration systems in low-and middle-income countries," ed: IUSSP Panel on Innovations in Strengthening Civil Registration & Vital Statistics Systems, 2016.
[30] D. de Savigny *et al.*, "Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): system-level considerations," *Global health action,* vol. 10, no. 1, p. 1272882, 2017.
[31] O. Sankoh and P. Byass, "Time for civil registration with verbal autopsy," *The Lancet Global Health,* vol. 2, no. 12, pp. e693-e694, 2014.
[32] C. AbouZahr, D. De Savigny, L. Mikkelsen, P. W. Setel, R. Lozano, and A. D. Lopez, "Towards universal civil registration and vital statistics systems: the time is now," *The Lancet,* vol. 386, no. 10001, pp. 1407-1418, 2015.
[33] World Health Organization, *The World Health Report 2005: Make every mother and child count*. World Health Organization, 2005.
[34] P. W. Setel *et al.*, "A scandal of invisibility: making everyone count by counting everyone," (in eng), *Lancet,* vol. 370, no. 9598, pp. 1569-77, Nov 2007, doi: 10.1016/S0140-6736(07)61307-5.
[35] M. Målqvist *et al.*, "Unreported births and deaths, a severe obstacle for improved neonatal survival in low-income countries; a population based study," *BMC international health and human rights,* vol. 8, no. 1, p. 4, 2008.
[36] J. Arudo *et al.*, "Comparison of government statistics and demographic surveillance to monitor mortality in children less than five years old in rural western Kenya," (in eng), *Am J Trop Med Hyg,* vol. 68, no. 4 Suppl, pp. 30-7, Apr 2003.
[37] K. Hill, A. D. Lopez, K. Shibuya, P. Jha, and M. o. V. E. (MoVE), "Interim measures for meeting needs for health sector data: births, deaths, and causes of death," (in eng),

Lancet, vol. 370, no. 9600, pp. 1726-35, Nov 2007, doi: 10.1016/S0140-6736(07)61309-9.

[38] P. Mahapatra, "An overview of the sample registration system in India," in *Prince Mahidol Award Conference & Global Health Information Forum*, 2010.

[39] E. D. Pratiwi and S. Kosen, "Development of an Indonesian sample registration system: a longitudinal study," *The Lancet,* vol. 381, p. S118, 2013.

[40] K. Z. Ahsan *et al.*, "Production and use of estimates for monitoring progress in the health sector: the case of Bangladesh," *Global health action,* vol. 10, no. sup1, p. 1298890, 2017.

[41] O. Sankoh and P. Byass, "The INDEPTH Network: filling vital gaps in global epidemiology," (in eng), *Int J Epidemiol,* vol. 41, no. 3, pp. 579-88, Jun 2012, doi: 10.1093/ije/dys081.

[42] T. Boerma, "Moving towards better cause of death registration in Africa and Asia," (in eng), *Glob Health Action,* vol. 7, p. 25931, 2014, doi: 10.3402/gha.v7.25931.

[43] C. L. Bose *et al.*, "The Global Network Maternal Newborn Health Registry: a multi-national, community-based registry of pregnancy outcomes," *Reproductive Health,* vol. 12, no. 2, p. S1, 2015.

[44] N. Upadhaya *et al.*, "Information systems for mental health in six low and middle income countries: cross country situation analysis," *International journal of mental health systems,* vol. 10, no. 1, p. 60, 2016.

[45] K. Tull, "Designing and Implementing Health Management Information Systems," *K4D Helpdesk Report,* 2018.

[46] I. Asangansi *et al.*, "Improving the routine HMIS in Nigeria through mobile technology for community data collection," *Journal of Health Informatics in Developing Countries,* vol. 7, no. 1, pp. 76-87, 2013.

[47] World Health Organization, "Covering every birth and death: Improving civil registration and vital statistics (CRVS): Report of the technical discussions, New Delhi, 16–17 June 2014," WHO Regional Office for South-East Asia, 2015.

[48] C. Mathers and T. Boerma, "Mortality measurement matters: improving data collection and estimation methods for child and adult mortality," *PLoS medicine,* vol. 7, no. 4, p. e1000265, 2010.

[49] DHS, "Demographic and Health Surveys: Topics," Available from: https://dhsprogram.com/Topics/ (Accessed on 19/09/2019), 2019.

[50] J. Boerma and A. Sommerfelt, "Demographic and health surveys (DHS): contributions and limitations," *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales,* vol. 46, no. 4, pp. 222-226, 1993.

[51] K. Hill, E. Brady, L. Zimmerman, L. Montana, R. Silva, and A. Amouzou, "Monitoring change in child mortality through household surveys," *PloS one,* vol. 10, no. 11, p. e0137713, 2015.

[52] L. Hug, M. Alexander, D. You, L. Alkema, and U. I.-a. G. for Child, "National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis," *The Lancet Global Health,* vol. 7, no. 6, pp. e710-e720, 2019.

[53] D. Dicker *et al.*, "Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: a systematic analysis for the Global Burden of Disease Study 2017," *The lancet,* vol. 392, no. 10159, pp. 1684-1735, 2018.

[54] L. Alkema and D. You, "Child mortality estimation: a comparison of UN IGME and IHME estimates of levels and trends in under-five mortality rates and deaths," *PLoS medicine,* vol. 9, no. 8, p. e1001288, 2012.

[55] T. Boerma, C. Victora, and C. Abouzahr, "Monitoring country progress and achievements by making global predictions: is the tail wagging the dog?," *The Lancet,* vol. 392, no. 10147, pp. 607-609, 2018.

[56] T. Kircher and R. E. Anderson, "Cause of death: proper completion of the death certificate," *Jama,* vol. 258, no. 3, pp. 349-352, 1987.
[57] World Health Organization, "World Health Statistics data visualizations dashboard | SDG target 17.19 | Death registration," Available from: http://apps.who.int/gho/data/view.sdg.17-19-data-reg?lang=en (Accessed on 24 June 2019), 2019.
[58] A. E. S. Sehdev and G. M. Hutchins, "Problems with proper completion and accuracy of the cause-of-death statement," *Archives of internal medicine,* vol. 161, no. 2, pp. 277-284, 2001.
[59] K. A. Myers and D. R. Farquhar, "Improving the accuracy of death certification," (in eng), *CMAJ,* vol. 158, no. 10, pp. 1317-23, May 1998.
[60] F. Janssen and A. E. Kunst, "ICD coding changes and discontinuities in trends in cause-specific mortality in six European countries, 1950-99," (in eng), *Bull World Health Organ,* vol. 82, no. 12, pp. 904-13, Dec 2004, doi: /S0042-96862004001200006.
[61] C. J. Murray, S. C. Kulkarni, and M. Ezzati, "Understanding the coronary heart disease versus total cardiovascular mortality paradox: a method to enhance the comparability of cardiovascular death statistics in the United States," (in eng), *Circulation,* vol. 113, no. 17, pp. 2071-81, May 2006, doi: 10.1161/CIRCULATIONAHA.105.595777.
[62] G. Rey *et al.*, "Cause-specific mortality time series analysis: a general method to detect and correct for abrupt data production changes," (in eng), *Popul Health Metr,* vol. 9, p. 52, 2011, doi: 10.1186/1478-7954-9-52.
[63] M. Garenne and V. Fauveau, "Potential and limits of verbal autopsies," (in eng), *Bull World Health Organ,* vol. 84, no. 3, p. 164, Mar 2006, doi: /S0042-96862006000300004.
[64] P. W. Setel *et al.*, "Core verbal autopsy procedures with comparative validation results from two countries," (in eng), *PLoS Med,* vol. 3, no. 8, p. e268, Aug 2006, doi: 10.1371/journal.pmed.0030268.
[65] J. Leitao *et al.*, "Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low- and middle-income countries: systematic review," (in eng), *BMC Med,* vol. 12, p. 22, 2014, doi: 10.1186/1741-7015-12-22.
[66] J. E. Lawn, D. Osrin, A. Adler, and S. Cousens, "Four million neonatal deaths: counting and attribution of cause of death," (in eng), *Paediatr Perinat Epidemiol,* vol. 22, no. 5, pp. 410-6, Sep 2008, doi: 10.1111/j.1365-3016.2008.00960.x.
[67] N. Soleman, D. Chandramohan, and K. Shibuya, "Verbal autopsy: current practices and challenges," (in eng), *Bull World Health Organ,* vol. 84, no. 3, pp. 239-45, Mar 2006, doi: /S0042-96862006000300020.
[68] F. Baiden *et al.*, "Setting international standards for verbal autopsy," (in eng), *Bull World Health Organ,* vol. 85, no. 8, pp. 570-1, Aug 2007.
[69] N. Thatte, H. D. Kalter, A. H. Baqui, E. M. Williams, and G. L. Darmstadt, "Ascertaining causes of neonatal deaths using verbal autopsy: current methods and challenges," (in eng), *J Perinatol,* vol. 29, no. 3, pp. 187-94, Mar 2009, doi: 10.1038/jp.2008.138.
[70] D. J. Corsi, M. Neuman, J. E. Finlay, and S. V. Subramanian, "Demographic and health surveys: a profile," (in eng), *Int J Epidemiol,* vol. 41, no. 6, pp. 1602-13, Dec 2012, doi: 10.1093/ije/dys184.
[71] R. Snow *et al.*, "Childhood deaths in Africa: uses and limitations of verbal autopsies," *The Lancet,* vol. 340, no. 8815, pp. 351-355, 1992.
[72] S. Herrera *et al.*, "A systematic review and synthesis of the strengths and limitations of measuring malaria mortality through verbal autopsy," *Malaria journal,* vol. 16, no. 1, p. 421, 2017.
[73] E. K. Nichols *et al.*, "The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0," *PLoS medicine,* vol. 15, no. 1, p. e1002486, 2018.

[74] World Health Organization, "Verbal autopsy standards: ascertaining and attributing cause of death," Available from: https://www.who.int/healthinfo/statistics/verbalautopsystandards/en/, 2016.
[75] P. Byass *et al.*, "Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia," (in eng), *J Glob Health,* vol. 5, no. 1, p. 010402, Jun 2015, doi: 10.7189/jogh.05.010402.
[76] AMANHI study group, "Burden, timing and causes of maternal and neonatal deaths and stillbirths in sub–Saharan Africa and South Asia: protocol for a prospective cohort study," *Journal of global health,* vol. 6, no. 2, 2016.
[77] P. Byass *et al.*, "An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model," *BMC medicine,* vol. 17, no. 1, p. 102, 2019.
[78] A. D. Flaxman *et al.*, "Collecting verbal autopsies: improving and streamlining data collection processes using electronic tablets," *Population health metrics,* vol. 16, no. 1, p. 3, 2018.
[79] Civil Registration and Vital Statistics Improvement Group and Bloomberg Philanthropies Data for Health Initiative, "Introducing verbal autopsies into civil registration and vital statistics systems: guiding principles," Available from: https://crvsgateway.info/file/9806/52 (Accessed on 16/06/2019), 2017.
[80] R. Rampatige, S. Gamage, S. Peiris, and A. D. Lopez, "Assessing the reliability of causes of death reported by the Vital Registration System in Sri Lanka: medical records review in Colombo," *Health Information Management Journal,* vol. 42, no. 3, pp. 20-28, 2013.
[81] K. J. Kerber *et al.*, "Counting every stillbirth and neonatal death through mortality audit to improve quality of care for every pregnant woman and her baby," *BMC pregnancy and childbirth,* vol. 15, no. 2, p. S9, 2015.
[82] B. D. Nicholson *et al.*, "Death audits and reviews for reducing maternal, perinatal and child mortality," *The Cochrane Database of Systematic Reviews,* vol. 2018, no. 3, 2018.
[83] R. Pattinson *et al.*, "Perinatal mortality audit: counting, accountability, and overcoming challenges in scaling up in low‐and middle‐income countries," *International Journal of Gynecology & Obstetrics,* vol. 107, no. Supplement, pp. S113-S122, 2009.
[84] Q. Bassat *et al.*, "Validity of a minimally invasive autopsy tool for cause of death determination in pediatric deaths in Mozambique: An observational study," *PLoS medicine,* vol. 14, no. 6, p. e1002317, 2017.
[85] A. Feroz *et al.*, "Perceptions of parents and healthcare professionals regarding minimal invasive tissue sampling to identify the cause of death in stillbirths and neonates: a qualitative study protocol," *Reproductive health,* vol. 15, no. 1, p. 179, 2018.
[86] A. Kone *et al.*, "Using Participatory Workshops to Assess Community Alignment or Tension for Child Mortality Surveillance Involving Minimally Invasive Tissue Sampling," in *American Journal of Tropical Medicine and Hygiene*, 2018, vol. 99, no. 4: Amer Soc Trop Med & Hygiene, pp. 476-476.
[87] P. Byass, "Minimally invasive autopsy: a new paradigm for understanding global health?," *PLoS medicine,* vol. 13, no. 11, p. e1002173, 2016.
[88] D. T. Jamison, W. H. Mosley, A. R. Measham, and J. L. Bobadilla, *Disease control priorities in developing countries*. Oxford University Press, 1993.
[89] D. Jamison, W. Mosley, A. Measham, and J. Bobadilla, "World development report: Investing in health," *Washington, DC: World Bank,* 1993.
[90] C. J. Murray and A. D. Lopez, *The Global Burden of Disease, Vol. 1 of Global Burden of Disease and Injury Series*. Boston: Havard University School of Public Health, 1996.
[91] C. J. Murray and A. D. Lopez, "Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study," (in eng), *Lancet,* vol. 349, no. 9063, pp. 1436-42, May 1997, doi: 10.1016/S0140-6736(96)07495-8.

[92] C. J. Murray and A. D. Lopez, "Mortality by cause for eight regions of the world: Global Burden of Disease Study," (in eng), *Lancet,* vol. 349, no. 9061, pp. 1269-76, May 1997, doi: 10.1016/S0140-6736(96)07493-4.
[93] R. Lozano *et al.*, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," (in eng), *Lancet,* vol. 380, no. 9859, pp. 2095-128, Dec 2012, doi: 10.1016/S0140-6736(12)61728-0.
[94] S. S. Morris, R. E. Black, and L. Tomaskovic, "Predicting the distribution of under-five deaths by cause in countries without adequate vital registration systems," (in eng), *Int J Epidemiol,* vol. 32, no. 6, pp. 1041-51, Dec 2003.
[95] R. E. Black, S. S. Morris, and J. Bryce, "Where and why are 10 million children dying every year?," (in eng), *Lancet,* vol. 361, no. 9376, pp. 2226-34, Jun 2003, doi: 10.1016/S0140-6736(03)13779-8.
[96] D. T. Jamison *et al.*, *Disease control priorities in developing countries*. World Bank Publications, 2006.
[97] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, *Global burden of disease and risk factors* (Disease Control Priorities Project). Washington DC: World Bank, 2006.
[98] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," (in eng), *Lancet,* vol. 367, no. 9524, pp. 1747-57, May 2006, doi: 10.1016/S0140-6736(06)68770-9.
[99] C. Mathers, D. M. Fat, and J. Boerma, *The global burden of disease: 2004 update*. World Health Organization, 2008.
[100] L. Liu *et al.*, "Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis," (in ENG), *Lancet,* Sep 2014, doi: 10.1016/S0140-6736(14)61698-6.
[101] L. Liu *et al.*, "Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals," *The Lancet,* vol. 388, no. 10063, pp. 3027-3035, 2016.
[102] H. Wang *et al.*, "Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013," (in ENG), *Lancet,* May 2014, doi: 10.1016/S0140-6736(14)60497-9.
[103] M. Naghavi *et al.*, "Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet,* vol. 390, no. 10100, pp. 1151-1210, 2017.
[104] World Bank, "Births attended by skilled health staff (% of total)," Available from: https://data.worldbank.org/indicator/SH.STA.BRTC.ZS, 2018.
[105] L. Alkema and J. R. New, "Global estimation of child mortality using a Bayesian B-spline bias-reduction model," *The Annals of Applied Statistics,* pp. 2122-2149, 2014.
[106] World Health Organization, "Disease burden and mortality estimates: child causes of death, 2000-2017," Available from: https://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html, 2018.
[107] S. Oza, S. N. Cousens, and J. E. Lawn, "Estimation of daily risk of neonatal death, including the day of birth, in 186 countries in 2013: a vital-registration and modelling-based study," (in eng), *Lancet Glob Health,* vol. 2, no. 11, pp. e635-44, Nov 2014, doi: 10.1016/S2214-109X(14)70309-2.
[108] J. E. Lawn *et al.*, "Newborn survival: a multi-country analysis of a decade of change," (in eng), *Health Policy Plan,* vol. 27 Suppl 3, pp. iii6-28, Jul 2012, doi: 10.1093/heapol/czs053.
[109] J. Lawn *et al.*, "Every Newborn: Progress, priorities, and potential beyond survival.," *Lancet.,* vol. 383, 2014.

[110] UNICEF, *Levels and Trends in Child Mortality - Report 2014. Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation*. UNICEF, 2014.
[111] WHO, UNICEF, UNFPA, and World Bank, "Trends in maternal mortality: 1990 to 2013," 2014.
[112] J. E. Lawn *et al.*, "Stillbirths: Where? When? Why? How to make the data count?," (in eng), *Lancet,* vol. 377, no. 9775, pp. 1448-63, Apr 2011, doi: 10.1016/S0140-6736(10)62187-3.
[113] J. E. Lawn, J. Mwansa-Kambafwile, F. C. Barros, B. L. Horta, and S. Cousens, "'Kangaroo mother care' to prevent neonatal deaths due to pre-term birth complications," (in ENG), *Int J Epidemiol,* Nov 2010, doi: 10.1093/ije/dyq172.
[114] Global Health Observatory of the WHO. "Demographic and socioeconomic statistics: Census and civil registration coverage by country." http://apps.who.int/gho/data/node.main.121 (accessed February 7, 2013).
[115] M. Z. Oestergaard *et al.*, "Neonatal mortality levels for 193 countries in 2009 with trends since 1990: a systematic analysis of progress, projections, and priorities," (in eng), *PLoS Med,* vol. 8, no. 8, p. e1001080, Aug 2011, doi: 10.1371/journal.pmed.1001080.
[116] ICF International, "Demographic and Health Surveys Methodology - Questionnaires: Household, Woman's, and Man's.," MEASURE DHS Phase III, Calverton, Maryland, USA, 2011.
[117] K. Hill and Y. Choi, "Neonatal mortality in the developing world," *Demographic Research,* vol. 14, no. 18, pp. 429-452, 2006.
[118] S. Neal, "The measurement of neonatal mortality: how reliable is Demographic and Household Survey Data?," *CPC Working Paper,* 2012.
[119] O. M. National Statistical Office (NSO) [Malawi], "Malawi Demographic and Health Survey 2004," Calverton, Maryland, 2005.
[120] G. Msemo *et al.*, "Newborn mortality and fresh stillbirth rates in Tanzania after helping babies breathe training," (in eng), *Pediatrics,* vol. 131, no. 2, pp. e353-60, Feb 2013, doi: 10.1542/peds.2012-1795.
[121] H. Blencowe *et al.*, "National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications," (in eng), *Lancet,* vol. 379, no. 9832, pp. 2162-72, Jun 2012, doi: 10.1016/S0140-6736(12)60820-4.
[122] R. Lozano *et al.*, "Progress towards Millennium Development Goals 4 and 5 on maternal and child mortality: an updated systematic analysis," (in eng), *Lancet,* vol. 378, no. 9797, pp. 1139-65, Sep 2011, doi: 10.1016/S0140-6736(11)61337-8.
[123] Global Health Observatory of the WHO. "Life Tables." http://www.who.int/gho/mortality_burden_disease/life_tables/life_tables/en/ (accessed May 24, 2013).
[124] M. Witten, "A return to time, cells, systems, and aging: IV. Further thoughts on Gompertzian survival dynamics—The neonatal years," *Mechanisms of ageing and development,* vol. 33, no. 2, pp. 177-190, 1986.
[125] M. Witten, "A return to time, cells, systems, and aging: V. Further thoughts on Gompertzian survival dynamics—the geriatric years," *Mechanisms of ageing and development,* vol. 46, no. 1, pp. 175-200, 1988.
[126] Global Health Observatory of the WHO, "Demographic and socioeconomic statistics: Census and civil registration coverage by country," Available from: http://apps.who.int/gho/data/node.main.121 (Accessed on 10 July 2013).
[127] G. G. Venter, "Mortality trend models," in *Casualty Actuarial Society E-Forum, Winter 2011 Volume 2*, 2011.
[128] J. Sullivan, "An Assessment of the Credibility of Child Mortality Declines Estimated from DHS Mortality Rates (Working Draft)," *Inter-Agency Coordination Group on Child Mortality Estimation: UNICEF,* 2008.

[129] S. L. Curtis, *Assessment of the quality of data used for direct estimation of infant and child mortality in DHS-II surveys*. Macro International Incorporated, 1995.
[130] K. Hill and R. Pande, "The recent evolution of child mortality in the developing world," *Arlington, VA, Partnership for Child Health Care, Basic Support for Institutionalizing Child Survival [BASICS]. Current Issues in Child Survival Series,* 1997.
[131] J. Lawn, M. Gravett, T. Nunes, C. Rubens, and C. Stanton, "Global report on preterm birth and stillbirth (1 of 7): definitions, description of the burden and opportunities to improve data," *BMC pregnancy and childbirth,* vol. 10, no. Suppl 1, p. S1, 2010.
[132] A. H. Baqui *et al.*, "Rates, timing and causes of neonatal deaths in rural India: implications for neonatal health programmes," (in eng), *Bull World Health Organ,* vol. 84, no. 9, pp. 706-13, Sep 2006, doi: S0042-96862006000900013 [pii].
[133] A. Leach *et al.*, "Neonatal mortality in a rural area of The Gambia," (in eng), *Ann Trop Paediatr,* vol. 19, no. 1, pp. 33-43, Mar 1999.
[134] A. Prost *et al.*, "Women's groups practising participatory learning and action to improve maternal and newborn health in low-resource settings: a systematic review and meta-analysis," (in eng), *Lancet,* vol. 381, no. 9879, pp. 1736-46, May 2013, doi: 10.1016/S0140-6736(13)60685-6.
[135] Z. S. Lassi, B. A. Haider, and Z. A. Bhutta, "Community-based intervention packages for reducing maternal and neonatal morbidity and mortality and improving neonatal outcomes," (in eng), *Cochrane Database Syst Rev,* no. 11, p. CD007754, 2010, doi: 10.1002/14651858.CD007754.pub2.
[136] R. Pattinson *et al.*, "Stillbirths: how can health systems deliver for mothers and babies?," (in eng), *Lancet,* vol. 377, no. 9777, pp. 1610-23, May 2011, doi: 10.1016/S0140-6736(10)62306-9.
[137] WHO and UNICEF, *Every Newborn: an action plan to end preventable deaths.* Geneva: World Health Organization, 2014.
[138] A. Coll-Seck, H. Clark, R. Bahl, S. Peterson, A. Costello, and T. Lucas, "Framing an agenda for children thriving in the SDG era: a WHO-UNICEF-Lancet Commission on Child Health and Wellbeing," (in eng), *Lancet,* vol. 393, no. 10167, pp. 109-112, 01 2019, doi: 10.1016/S0140-6736(18)32821-6.
[139] World Health Organization, "World health statistics 2014," 2014.
[140] UNICEF, "Committing to Child Survival: A Promise Renewed. Progress Report 2014," *UNICEF. New York,* 2014.
[141] S. Oza, J. E. Lawn, D. R. Hogan, C. Mathers, and S. N. Cousens, "Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000-2013," (in eng), *Bull World Health Organ,* vol. 93, no. 1, pp. 19-28, Jan 2015, doi: 10.2471/BLT.14.139790.
[142] UNICEF, "Levels and Trends in Child Mortality - Report 2013. Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation," UNICEF, 2013.
[143] J. Lawn *et al.*, "Every Newborn: survival and beyond.," *Lancet.,* vol. in press., 2014.
[144] Every Newborn. "Every newborn: An action plan to end preventable deaths." www.everynewborn.org (accessed March 30, 2014).
[145] CDC Wonder. "Linked Birth/Infant Death Records for 2007-2010 with ICD10 codes." http://wonder.cdc.gov/lbd.html (accessed August 20, 2013).
[146] H. R. Chowdhury, S. Thompson, M. Ali, N. Alam, M. Yunus, and P. K. Streatfield, "Causes of neonatal deaths in a rural subdistrict of Bangladesh: implications for intervention," (in eng), *J Health Popul Nutr,* vol. 28, no. 4, pp. 375-82, Aug 2010.
[147] Global Health Observatory of the WHO. "Demographic and socioeconomic statistics: Census and civil registration coverage by country." http://apps.who.int/gho/data/node.main.121 (accessed July 10, 2013).
[148] J. E. Lawn, "4 million neonatal deaths: an analysis of available cause-of-death data and systematic country estimates with a focus on “birth asphyxia”," UCL (University College London), 2009.

[149] J. S. Wigglesworth, "Monitoring perinatal mortality. A pathophysiological approach," (in eng), *Lancet,* vol. 2, no. 8196, pp. 684-6, Sep 1980.
[150] I. G. Winbo, F. H. Serenius, G. G. Dahlquist, and B. A. Källén, "NICE, a new cause of death classification for stillbirths and neonatal deaths. Neonatal and Intrauterine Death Classification according to Etiology," (in eng), *Int J Epidemiol,* vol. 27, no. 3, pp. 499-504, Jun 1998.
[151] J. E. Lawn, K. Wilczynska-Ketende, and S. N. Cousens, "Estimating the causes of 4 million neonatal deaths in the year 2000," (in eng), *Int J Epidemiol,* vol. 35, no. 3, pp. 706-18, Jun 2006, doi: 10.1093/ije/dyl043.
[152] World Health Organization, "The Global Health Observatory data repository," Available from: www.who.int/gho/, 2014.
[153] J. E. Lawn, K. Kerber, C. Enweronu-Laryea, and O. Massee Bateman, "Newborn survival in low resource settings--are we delivering?," (in eng), *BJOG,* vol. 116 Suppl 1, pp. 49-59, Oct 2009, doi: 10.1111/j.1471-0528.2009.02328.x.
[154] J. Katz *et al.*, "Mortality risk in preterm and small-for-gestational-age infants in low-income and middle-income countries: a pooled country analysis," (in eng), *Lancet,* vol. 382, no. 9890, pp. 417-25, Aug 2013, doi: 10.1016/S0140-6736(13)60993-9.
[155] X. L. Feng *et al.*, "China's facility-based birth strategy and neonatal mortality: a population-based epidemiological study," (in eng), *Lancet,* vol. 378, no. 9801, pp. 1493-500, Oct 2011, doi: 10.1016/S0140-6736(11)61096-9.
[156] I. Rudan *et al.*, "Causes of deaths in children younger than 5 years in China in 2008," (in eng), *Lancet,* vol. 375, no. 9720, pp. 1083-9, Mar 2010, doi: 10.1016/S0140-6736(10)60060-8.
[157] D. T. Jamison *et al.*, "Global health 2035: a world converging within a generation," (in eng), *Lancet,* vol. 382, no. 9908, pp. 1898-955, Dec 2013, doi: 10.1016/S0140-6736(13)62105-4.
[158] O. Nelles, *Nonlinear system identification : from classical approaches to neural networks and fuzzy models*. New York: Springer, 2000.
[159] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," (in eng), *BMC Genomics,* vol. 13 Suppl 4, p. S2, Jun 2012, doi: 10.1186/1471-2164-13-S4-S2.
[160] Y. Akachi and M. E. Kruk, "Quality of care: measuring a neglected driver of improved health," (in eng), *Bull World Health Organ,* vol. 95, no. 6, pp. 465-472, Jun 2017, doi: 10.2471/BLT.16.180190.
[161] A. B. Moller *et al.*, "Measures matter: A scoping review of maternal and newborn indicators," (in eng), *PLoS One,* vol. 13, no. 10, p. e0204763, 2018, doi: 10.1371/journal.pone.0204763.
[162] N. Fantom and U. Serajuddin, *The World Bank's classification of countries by income*. The World Bank, 2016.
[163] I. Davidson and W. Fan, "When efficient model averaging out-performs boosting and bagging," in *Knowledge Discovery in Databases: PKDD 2006*: Springer, 2006, pp. 478-486.
[164] P. Bühlmann, "Bagging, boosting and ensemble methods," in *Handbook of Computational Statistics*: Springer, 2012, pp. 985-1022.
[165] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 41, no. 3, pp. 552-568, 2010.
[166] A. Ullah and H. Wang, "Parametric and Nonparametric Frequentist Model Selection and Model Averaging," *Econometrics,* vol. 1, no. 2, pp. 157-179, 2013.
[167] R. Maclin and D. Opitz, "Popular ensemble methods: An empirical study," *arXiv preprint arXiv:1106.0257,* 2011.
[168] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[169] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.
[170] R. Maclin and D. Opitz, "An empirical evaluation of bagging and boosting," *AAAI/IAAI,* vol. 1997, pp. 546-551, 1997.
[171] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *Journal of the American Statistical Association,* vol. 99, no. 465, pp. 279-290, 2004.
[172] I. J. Myung, "The importance of complexity in model selection," *Journal of Mathematical Psychology,* vol. 44, no. 1, pp. 190-204, 2000.
[173] L. McCann, "Robust model selection and outlier detection in linear regressions," Massachusetts Institute of Technology, 2006.
[174] M. Yuan and Y. Lin, "On the non‐negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 69, no. 2, pp. 143-161, 2007.
[175] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288, 1996.
[176] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 70, no. 1, pp. 53-71, 2008.
[177] D. Drukker, "Inference after lasso model selection," in *2019 Stata Conference*, 2019, no. 3: Stata Users Group.
[178] T. A. Snijders, "Fixed and random effects," *Encyclopedia of statistics in behavioral science,* 2005.
[179] S. Rabe-Hesketh, "GLLAMM: Stata program to fit generalised linear latent and mixed models," Available from: https://econpapers.repec.org/software/bocbocode/s401701.htm, 2011.
[180] A. R. Hole, "Mixed logit modeling in Stata--an overview," in *United Kingdom Stata Users' Group Meetings 2013*, 2013, no. 23: Stata Users Group.
[181] C. Huber, "Introduction to Structural Equation Modeling Using Stata," *California Association for Institutional Research,* 2014.
[182] W. Gould, "Stata 15 announced, available now.," ed. Available from: https://blog.stata.com/2017/06/06/stata-15-announced-available-now/, 2017.
[183] R. Williams, "Review of Alan Acock's Discovering Structural Equation Modeling Using Stata, Revised Edition," *The Stata Journal,* vol. 15, no. 1, pp. 309-315, 2015.
[184] R. Pope, "In the spotlight: meet Stata's new xtmlogit command.," ed. Available from: https://www.stata.com/stata-news/news29-2/xtmlogit/, 2014.
[185] Statalist forum post, "Random effects in mlogit (no panel data)," ed. Available from:https://www.statalist.org/forums/forum/general-stata-discussion/general/1426914-random-effects-in-mlogit-no-panel-data, 2018.
[186] R. G. Gutierrez, "Recent developments in multilevel modeling, including models for binary and count responses," *North American Stata users group. Boston,* 2007.
[187] A. Agresti, *Categorical Data Analysis*. Wiley, 2003.
[188] S. B. McGrayne, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press, 2011.
[189] M. J. Zyphur and F. L. Oswald, "Bayesian estimation and inference: A user's guide," *Journal of Management,* vol. 41, no. 2, pp. 390-420, 2015.
[190] Science and Technology Columns, "In praise of Bayes," in *The Economist*, ed. Available from: https://www.economist.com/science-and-technology/2000/09/28/in-praise-of-bayes, 2000.
[191] M. Richey, "The evolution of Markov chain Monte Carlo methods," *The American Mathematical Monthly,* vol. 117, no. 5, pp. 383-413, 2010.

[192] M. A. Tanner and W. H. Wong, "From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s," *Statistical science,* vol. 25, no. 4, pp. 506-516, 2010.
[193] M. E. Glickman and D. A. Van Dyk, "Basic bayesian methods," in *Topics in Biostatistics*: Springer, 2007, pp. 319-338.
[194] J. O. Berger and R. L. Wolpert, "A conversation with James O. Berger," *Statistical science,* pp. 205-218, 2004.
[195] J. Hadfield, "MCMCglmm course notes," in "Available from: http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf," 2012.
[196] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician,* vol. 46, no. 3, pp. 167-174, 1992.
[197] R. B. O'Hara and M. J. Sillanpää, "A review of Bayesian variable selection methods: what, how and which," *Bayesian analysis,* vol. 4, no. 1, pp. 85-117, 2009.
[198] J. Piironen and A. Vehtari, "Comparison of Bayesian predictive methods for model selection," *Statistics and Computing,* vol. 27, no. 3, pp. 711-735, 2017.
[199] A. Vehtari and J. Ojanen, "A survey of Bayesian predictive methods for model assessment, selection and comparison," *Statistics Surveys,* vol. 6, pp. 142-228, 2012.
[200] B. S. Bloom, N. De Pouvourville, and S. Libert, "Classic or bayesian research design and analysis: Does it make a difference?," *International journal of technology assessment in health care,* vol. 18, no. 1, pp. 120-126, 2002.
[201] G. H. Skrepnek, "The contrast and convergence of Bayesian and frequentist statistical approaches in pharmacoeconomic analysis," *Pharmacoeconomics,* vol. 25, no. 8, pp. 649-664, 2007.
[202] T. Vos, R. Barber, D. E. Phillips, A. D. Lopez, and C. J. Murray, "Causes of child death: comparison of MCEE and GBD 2013 estimates - Authors' reply," (in eng), *Lancet,* vol. 385, no. 9986, pp. 2462-4, Jun 2015, doi: 10.1016/S0140-6736(15)61133-3.
[203] K. J. Foreman, R. Lozano, A. D. Lopez, and C. Murray, "Modeling causes of death: an integrated approach using CODEm," University of Washington, 2011.
[204] M. Alexander and L. Alkema, "Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model," *Demographic Research,* vol. 38, pp. 335-372, 2018.
[205] L. Alkema *et al.*, "Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group," (in eng), *Lancet,* vol. 387, no. 10017, pp. 462-74, Jan 2016, doi: 10.1016/S0140-6736(15)00838-7.
[206] D. You *et al.*, "Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation," (in eng), *Lancet,* vol. 386, no. 10010, pp. 2275-86, Dec 2015, doi: 10.1016/S0140-6736(15)00120-8.
[207] A. D. Flaxman, D. T. Vos, and C. J. Murray, *An integrative metaregression framework for descriptive epidemiology*. University of Washington Press, 2015.
[208] L. Liu, R. E. Black, S. Cousens, C. Mathers, J. E. Lawn, and D. R. Hogan, "Causes of child death: comparison of MCEE and GBD 2013 estimates," *The Lancet,* vol. 385, no. 9986, pp. 2461-2462, 2015.
[209] G. A. Stevens *et al.*, "Guidelines for accurate and transparent health estimates reporting: the GATHER statement," *PLoS medicine,* vol. 13, no. 6, p. e1002056, 2016.
[210] L. Liu *et al.*, "National, regional, and state-level all-cause and cause-specific under-5 mortality in India in 2000–15: a systematic analysis with implications for the Sustainable Development Goals," *The Lancet Global Health,* vol. 7, no. 6, pp. e721-e734, 2019.

[211] K. J. Kerber, J. E. de Graft-Johnson, Z. A. Bhutta, P. Okong, A. Starrs, and J. E. Lawn, "Continuum of care for maternal, newborn, and child health: from slogan to service delivery," *The Lancet,* vol. 370, no. 9595, pp. 1358-1369, 2007.
[212] D. Chou, B. Daelmans, R. R. Jolivet, M. Kinney, and L. Say, "Ending preventable maternal and newborn mortality and stillbirths," *Bmj,* vol. 351, p. h4255, 2015.
[213] K. Kikuchi *et al.*, "Effective linkages of continuum of care for improving neonatal, perinatal, and maternal mortality: a systematic review and meta-analysis," *PloS one,* vol. 10, no. 9, p. e0139288, 2015.
[214] O. Pasha *et al.*, "Communities, birth attendants and health facilities: a continuum of emergency maternal and newborn care (the Global Network's EmONC trial)," *BMC pregnancy and childbirth,* vol. 10, no. 1, p. 82, 2010.
[215] World Health Organization, "The WHO application of ICD-10 to deaths during the perinatal period: ICD-PM," 2016.
[216] K. Seipp, F. Gutiérrez, X. Ochoa, and K. Verbert, "Towards a visual guide for communicating uncertainty in visual analytics," *Journal of Visual Languages and Computing,* vol. 50, pp. 1-18, 2019.
[217] L. Dwyer-Lindgren *et al.*, "Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017," *Nature,* vol. 570, no. 7760, p. 189, 2019.
[218] World Health Organization, "WHO-MCEE estimates for child causes of death 2000-2015.," ed. Available from: http://www.who.int/entity/healthinfo/global_burden_disease/childCOD_estimates_2000_2015.xls (Accessed 23 October 2016). 2016.
[219] World Health Organization, "Verbal autopsy standards: ascertaining and attributing cause of death," 2007.
[220] T. Boerma *et al.*, "Countdown to 2030: tracking progress towards universal coverage for reproductive, maternal, newborn, and child health," *The Lancet,* vol. 391, no. 10129, pp. 1538-1548, 2018.
[221] A. Moran *et al.*, "'What gets measured gets managed': revisiting the indicators for maternal and newborn health programmes," *Reproductive health,* vol. 15, no. 1, p. 19, 2018.
[222] C. J. Murray, "Choosing indicators for the health-related SDG targets," *The Lancet,* vol. 386, no. 10001, pp. 1314-1317, 2015.
[223] C. W. Kabudula *et al.*, "Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa," *Population health metrics,* vol. 12, no. 1, p. 23, 2014.
[224] I. M. Moriyama, R. M. Loy, A. H. T. Robb-Smith, H. M. Rosenberg, and D. L. Hoyert, "History of the statistical classification of diseases and causes of death," 2011.
[225] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies," *Annals of internal medicine,* vol. 147, no. 8, pp. 573-577, 2007.
[226] A. Baqui *et al.*, "Effect of community-based newborn care on cause-specific neonatal mortality in Sylhet district, Bangladesh: findings of a cluster-randomized controlled trial," *Journal of Perinatology,* vol. 36, no. 1, p. 71, 2016.
[227] H. D. Kalter, J. Perin, and R. E. Black, "Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death," (in eng), *J Glob Health,* vol. 6, no. 1, p. 010601, Jun 2016, doi: 10.7189/jogh.06.010601.
[228] C. J. Murray and A. D. Lopez, "Measuring global health: motivation and evolution of the Global Burden of Disease Study," *The Lancet,* vol. 390, no. 10100, pp. 1460-1464, 2017.
[229] L. Alkema, J. R. New, J. Pedersen, and D. You, "Child mortality estimation 2013: an overview of updates in estimation methods by the United Nations Inter-agency Group for Child Mortality Estimation," *PloS one,* vol. 9, no. 7, p. e101112, 2014.

[230] C. AbouZahr, T. Boerma, and D. Hogan, "Global estimates of country health indicators: useful, unnecessary, inevitable?," *Global health action,* vol. 10, no. sup1, p. 1290370, 2017.
[231] E. Pisani and M. Kok, "In the eye of the beholder: to make global health estimates useful, make them more socially robust," *Global health action,* vol. 10, no. sup1, p. 1266180, 2017.
[232] M. Mahajan, "The IHME in the shifting landscape of global health metrics," *Global Policy,* vol. 10, pp. 110-120, 2019.
[233] C. AbouZahr, T. Boerma, and P. Byass, "Bridging the data gaps: do we have the right balance between country data and global estimates?," ed: Taylor & Francis, 2017.
[234] W. H. Organization, "The utility of estimates for health monitoring and decision-making: global, regional and country perspectives: report of a technical meeting, WHO, Glion sur Montreux, Switzerland 24–25 June 2015," World Health Organization, 2015.
[235] C. AbouZahr, S. Adjei, and C. Kanchanachitra, "From data to policy: good practices and cautionary tales," *The Lancet,* vol. 369, no. 9566, pp. 1039-1046, 2007.
[236] UNICEF, "Millennium Development Goals (MDG) monitoring," Available from: https://www.unicef.org/statistics/index_24304.html, 2014.
[237] P. W. Setel, C. AbouZahr, A. Karpati, and M. Bratschi, "Civil Registration and Vital Statistics (CRVS), and the Sustainable Development Goals (SDGs) - Keynote Paper," *International Conference on Civil Registration and Vital Statistics (Dhaka, Bangladesh),* 2018.
[238] World Health Organization, "The utility of estimates for health monitoring and decision-making: global, regional and country perspectives: report of a technical meeting, WHO, Glion sur Montreux, Switzerland 24–25 June 2015," World Health Organization, 2015.
[239] K. Adazu *et al.*, "Health and demographic surveillance in rural western Kenya: a platform for evaluating interventions to reduce morbidity and mortality from infectious diseases," (in eng), *Am J Trop Med Hyg,* vol. 73, no. 6, pp. 1151-8, Dec 2005.
[240] A. Aguilar, R. Alvarado, D. Cordero, P. Kelly, A. Zamora, and R. Salgado, "Mortality Survey in Bolivia: The Final Report. Investigating and Identifying Causes of Death for Children Under Five.," Published for the USAID by the Basic Support for Institutionalizing Child Survival (BASICS) Project., Arlington, VA., 1998.
[241] S. Akgün, M. Colak, and C. Bakar, "Identifying and verifying causes of death in Turkey: National verbal autopsy survey," (in eng), *Public Health,* vol. 126, no. 2, pp. 150-8, Feb 2012, doi: 10.1016/j.puhe.2011.09.031.
[242] J. Aleman, I. Brännström, J. Liljestrand, R. Peña, L. A. Persson, and J. Steidinger, "Saving more neonates in hospital: an intervention towards a sustainable reduction in neonatal mortality in a Nicaraguan hospital," (in eng), *Trop Doct,* vol. 28, no. 2, pp. 88-92, Apr 1998.
[243] K. Anand, S. Kant, G. Kumar, and S. K. Kapoor, ""Development" is not essential to reduce infant mortality rate in India: experience from the Ballabgarh project," (in eng), *J Epidemiol Community Health,* vol. 54, no. 4, pp. 247-53, Apr 2000.
[244] K. Asling-Monemi, R. Peña, M. C. Ellsberg, and L. A. Persson, "Violence against women increases the risk of infant and child mortality: a case-referent study in Nicaragua," (in eng), *Bull World Health Organ,* vol. 81, no. 1, pp. 10-6, 2003.
[245] S. Awasthi and V. K. Pande, "Cause-specific mortality in under fives in the urban slums of Lucknow, north India," (in eng), *J Trop Pediatr,* vol. 44, no. 6, pp. 358-61, Dec 1998.
[246] F. Baiden, A. Hodgson, M. Adjuik, P. Adongo, B. Ayaga, and F. Binka, "Trend and causes of neonatal mortality in the Kassena-Nankana district of northern Ghana, 1995-2002," (in eng), *Trop Med Int Health,* vol. 11, no. 4, pp. 532-9, Apr 2006, doi: 10.1111/j.1365-3156.2006.01582.x.

[247] E. Balci, E. Kucuk, I. Gun, M. Gulgun, B. Kilic, and K. Cetinkara, "Neonatal deaths at Melikgazi at Kayseri in 2006.," *Fırat University Medical Journal of Health Sciences,* vol. 22, no. 6, pp. 323-326, 2008

[248] A. T. Bang, R. A. Bang, S. B. Baitule, M. H. Reddy, and M. D. Deshmukh, "Effect of home-based neonatal care and management of sepsis on neonatal mortality: field trial in rural India," (in eng), *Lancet,* vol. 354, no. 9194, pp. 1955-61, Dec 1999, doi: 10.1016/S0140-6736(99)03046-9.

[249] U. Bapat, G. Alcock, N. S. More, S. Das, W. Joshi, and D. Osrin, "Stillbirths and newborn deaths in slum settlements in Mumbai, India: a prospective verbal autopsy study," (in eng), *BMC Pregnancy Childbirth,* vol. 12, p. 39, 2012, doi: 10.1186/1471-2393-12-39.

[250] F. C. Barros, C. G. Victora, J. P. Vaughan, and H. J. Estanislau, "Perinatal mortality in southern Brazil: a population-based study of 7392 births," (in eng), *Bull World Health Organ,* vol. 65, no. 1, pp. 95-104, 1987.

[251] A. J. Barros, A. Matijasevich, I. S. Santos, E. P. Albernaz, and C. G. Victora, "Neonatal mortality: description and effect of hospital of birth after risk adjustment," (in eng), *Rev Saude Publica,* vol. 42, no. 1, pp. 1-9, Feb 2008.

[252] D. G. Bassani *et al.*, "Causes of neonatal and child mortality in India: a nationally representative mortality survey," (in eng), *Lancet,* vol. 376, no. 9755, pp. 1853-60, Nov 2010, doi: 10.1016/S0140-6736(10)61461-4.

[253] A. Bezzaoucha, A. El Kebboub, and A. Aliche, "[Evolution of neonatal mortality at the Blida University Teaching Hospital (Algeria) between 1999 and 2006]," (in fre), *Bull Soc Pathol Exot,* vol. 103, no. 1, pp. 29-36, Feb 2010, doi: 10.1007/s13149-009-0001-z.

[254] S. Bhatia, "Patterns and causes of neonatal and postneonatal mortality in rural Bangladesh," (in eng), *Stud Fam Plann,* vol. 20, no. 3, pp. 136-46, 1989 May-Jun 1989.

[255] Z. Bhutta, "Hala Community-Based Trial. Report of Baseline Analysis.," Pakistan: Aga Khan University, 2003.

[256] O. Campbell *et al.*, "The Egypt National Perinatal/Neonatal Mortality Study 2000," (in eng), *J Perinatol,* vol. 24, no. 5, pp. 284-9, May 2004, doi: 10.1038/sj.jp.7211084.

[257] A. Chowdhury, K. Aziz, A. de Francisco, and M. Khan, "Differences in neonatal mortality by religious and socioeconomic covariates in rural Bangladesh," *The Journal of Family Welfare,* vol. 42, no. 2, pp. 31-40, 1996.

[258] M. E. Chowdhury, H. H. Akhter, V. Chongsuvivatwong, and A. F. Geater, "Neonatal mortality in rural Bangladesh: an exploratory study," (in eng), *J Health Popul Nutr,* vol. 23, no. 1, pp. 16-24, Mar 2005.

[259] G. L. Darmstadt *et al.*, "Evaluation of a cluster-randomized controlled trial of a package of community-based maternal and newborn interventions in Mirzapur, Bangladesh," (in eng), *PLoS One,* vol. 5, no. 3, p. e9696, 2010, doi: 10.1371/journal.pone.0009696.

[260] N. Datta, M. Mand, and V. Kumar, "Validation of causes of infant death in the community by verbal autopsy," (in eng), *Indian J Pediatr,* vol. 55, no. 4, pp. 599-604, 1988 Jul-Aug 1988.

[261] A. Deribew, F. Tessema, and B. Girma, "Determinants of under-five mortality in Gilgel Gibe Field Research Center, Southwest Ethiopia," *Ethiopian Journal of Health Development,* vol. 21, no. 2, pp. 117-124, 2007.

[262] NIPORT/Bangladesh, Mitra and Associates, and ORC Macro, "Bangladesh Demographic and Health Survey," 2005.

[263] Honduras Secretaría de Salud, Instituto Nacional de Estadística, and Macro International, *Encuesta nacional de demografía y salud: ENDESA 2005-2006*. Instituto Nacional de Estadística, 2006.

[264] S. Khanal, V. S. Gc, P. Dawson, and R. Houston, "Verbal autopsy to ascertain causes of neonatal deaths in a community setting: a study from Morang, Nepal," (in eng), *JNMA J Nepal Med Assoc,* vol. 51, no. 181, pp. 21-7, 2011 Jan-Mar 2011.

[265] NIPS/Pakistan, "Pakistan Demographic and Health Survey 2006-07 Islamabad," *Pakistan: National Institute of Population Studies and Macro International Inc,* 2008.

[266] Mozambique Ministry of Health, UNCIEF Mozambique, and LSHTM, "Mozambique - National Child Mortality Study 2009 - Summary," Maputo, Mozambique Ministry of Health, 2009.
[267] S. Djaja and S. Soemantri, "The cause of neonatal death and the attributed health care system in Indonesia: mortality study of household health survey, 2001," *Jakarta: National of Health Research and Development, Ministry of Health Indonesia,* 2003.
[268] J. Dommisse, "The causes of perinatal deaths in the greater Cape Town area. A 12-month survey," (in eng), *S Afr Med J,* vol. 80, no. 6, pp. 270-5, Sep 1991.
[269] K. M. Edmond *et al.*, "Aetiology of stillbirths and neonatal deaths in rural Ghana: implications for health programming in developing countries," (in eng), *Paediatr Perinat Epidemiol,* vol. 22, no. 5, pp. 430-7, Sep 2008, doi: 10.1111/j.1365-3016.2008.00961.x.
[270] E. E. Ekanem, A. A. Asindi, and O. U. Okoi, "Community-based surveillance of paediatric deaths in Cross River State, Nigeria," (in eng), *Trop Geogr Med,* vol. 46, no. 5, pp. 305-8, 1994.
[271] M. Y. el-Zibdeh, S. A. Al-Suleiman, and M. H. Al-Sibai, "Perinatal mortality at King Fahd Hospital of the University Al-Khobar, Saudi Arabia," (in eng), *Int J Gynaecol Obstet,* vol. 26, no. 3, pp. 399-407, Jun 1988.
[272] M. Fantahun, "Patterns of childhood mortality in three districts of north Gondar Administrative Zone. A community based study using the verbal autopsy method," (in eng), *Ethiop Med J,* vol. 36, no. 2, pp. 71-81, Apr 1998.
[273] V. Fauveau, B. Wojtyniak, G. Mostafa, A. M. Sarder, and J. Chakraborty, "Perinatal mortality in Matlab, Bangladesh: a community-based study," (in eng), *Int J Epidemiol,* vol. 19, no. 3, pp. 606-12, Sep 1990.
[274] F. F. Fikree, S. I. Azam, and H. W. Berendes, "Time to focus child survival programmes on the newborn: assessment of levels and causes of infant mortality in rural Pakistan," (in eng), *Bull World Health Organ,* vol. 80, no. 4, pp. 271-6, 2002.
[275] P. Fonseka, K. Wijewardene, D. G. Harendra de Silva, C. Goonaratna, and W. A. Wijeyasiri, "Neonatal and post-neonatal mortality in the Galle district," (in eng), *Ceylon Med J,* vol. 39, no. 2, pp. 82-5, Jun 1994.
[276] M. Garenne, N. Darkaoui, M. Braikat, and M. Azelmat, "Changing cause of death profile in Morocco: the impact of child-survival programmes," (in eng), *J Health Popul Nutr,* vol. 25, no. 2, pp. 212-20, Jun 2007.
[277] C. J. Gill *et al.*, "Effect of training traditional birth attendants on neonatal mortality (Lufwanyama Neonatal Survival Project): randomised controlled study," (in eng), *BMJ,* vol. 342, p. d346, 2011.
[278] J. d. O. Gomes and A. H. Santo, "Mortalidade infantil em município da região Centro-Oeste Paulista, Brasil, 1990 a 1992," *Rev Saúde Pública,* vol. 31, no. 4, pp. 330-341, 1997.
[279] A. M. Greenwood *et al.*, "A prospective survey of the outcome of pregnancy in a rural area of the Gambia," (in eng), *Bull World Health Organ,* vol. 65, no. 5, pp. 635-43, 1987.
[280] A. K. Halder *et al.*, "Causes of early childhood deaths in urban Dhaka, Bangladesh," (in eng), *PLoS One,* vol. 4, no. 12, p. e8145, 2009, doi: 10.1371/journal.pone.0008145.
[281] S. G. Hinderaker *et al.*, "Avoidable stillbirths and neonatal deaths in rural Tanzania," (in eng), *BJOG,* vol. 110, no. 6, pp. 616-23, Jun 2003.
[282] I. Jehan *et al.*, "Neonatal mortality, risk factors and causes: a prospective population-based cohort study in urban Pakistan," (in eng), *Bull World Health Organ,* vol. 87, no. 2, pp. 130-8, Feb 2009.
[283] A. Karabulut, B. İstanbullu, T. Karahan, and K. Özdemir, "Two year evaluation of infant and maternal mortality in Denizli," *Journal of the Turkish-German Gynecological Association,* vol. 10, pp. 95-98, 2009.

[284] N. Khalique, S. N. Sinha, M. Yunus, and A. Malik, "Early childhood mortality--a rural study," (in eng), *J R Soc Health,* vol. 113, no. 5, pp. 247-9, Oct 1993.
[285] S. R. Khan, F. Jalil, S. Zaman, B. S. Lindblad, and J. Karlberg, "Early child health in Lahore, Pakistan: X. Mortality," (in eng), *Acta Paediatr Suppl,* vol. 82 Suppl 390, pp. 109-17, Aug 1993.
[286] P. Khanjanasthiti, V. Benchakarn, A. Saksawad, S. Khantanaphar, and P. Posayanond, "Perinatal problems in rural Thailand," (in eng), *J Trop Pediatr,* vol. 30, no. 2, pp. 72-8, Apr 1984.
[287] J. Kishan, A. L. Soni, A. Y. Elzouki, N. A. Mir, and M. R. Magoub, "Perinatal outcome at Benghazi and implications for perinatal care in developing countries," (in eng), *Indian J Pediatr,* vol. 55, no. 4, pp. 611-5, 1988 Jul-Aug 1988.
[288] A. Krishnan, N. Ng, S. K. Kapoor, C. S. Pandav, and P. Byass, "Temporal trends and gender differentials in causes of childhood deaths at Ballabgarh, India - need for revisiting child survival strategies," (in eng), *BMC Public Health,* vol. 12, p. 555, 2012, doi: 10.1186/1471-2458-12-555.
[289] B. L. Liu, D. Z. Zhang, H. Q. Tao, and P. Huang, "Perinatal mortality rate in 11 Jiangsu cities," (in eng), *Chin Med J (Engl),* vol. 98, no. 3, pp. 157-60, Mar 1985.
[290] G. N. Lucas and R. C. Ediriweera, "Perinatal deaths at the Castle Street Hospital for Women in 1993," (in eng), *Ceylon Med J,* vol. 41, no. 1, pp. 10-2, Mar 1996.
[291] S. R. Manandhar *et al.*, "Causes of stillbirths and neonatal deaths in Dhanusha district, Nepal: a verbal autopsy study," (in eng), *Kathmandu Univ Med J (KUMJ),* vol. 8, no. 29, pp. 62-72, 2010 Jan-Mar 2010.
[292] R. M. Matendo *et al.*, "Challenge of reducing perinatal mortality in rural Congo: findings of a prospective, population-based study," (in eng), *J Health Popul Nutr,* vol. 29, no. 5, pp. 532-40, Oct 2011.
[293] A. Matijasevich *et al.*, "Perinatal mortality in three population-based cohorts from Southern Brazil: trends and differences," (in eng), *Cad Saude Publica,* vol. 24 Suppl 3, pp. S399-408, 2008.
[294] E. Mendieta, V. Battaglia, and B. Villalba, "Mortalidad neonatal en el Paraguay: análisis de los indicadores," *Pediatría (Asunción),* vol. 28, no. 1, pp. 12-18, 2001.
[295] D. Nandan *et al.*, "Social audits for community action: a tool to initiate community action for reducing child mortality," *Indian Journal of Community Medicine,* vol. 30, no. 3, p. 5, 2005.
[296] N. T. Nga, D. T. Hoa, M. Målqvist, L. Persson, and U. Ewald, "Causes of neonatal death: results from NeoKIP community-based trial in Quang Ninh province, Vietnam," (in eng), *Acta Paediatr,* vol. 101, no. 4, pp. 368-73, Apr 2012, doi: 10.1111/j.1651-2227.2011.02513.x.
[297] R. Pattinson, "Perinatal Problem Identification Programme data for South Africa," South Africa, 2013.
[298] H. Perry, "Causes of Neonatal Mortality in Urban Bangladesh and Rural Haiti," 2003.
[299] R. K. Phukan and J. Mahanta, "A study of neonatal deaths in the tea gardens of Dibrugarh district of upper Assam," (in eng), *J Indian Med Assoc,* vol. 96, no. 11, pp. 333-4, 337, Nov 1998.
[300] G. Pison, J. F. Trape, M. Lefebvre, and C. Enel, "Rapid decline in child mortality in a rural area of Senegal," (in eng), *Int J Epidemiol,* vol. 22, no. 1, pp. 72-80, Feb 1993.
[301] A. Pratinidhi, U. Shah, A. Shrotri, and N. Bodhani, "Risk-approach strategy in neonatal care," (in eng), *Bull World Health Organ,* vol. 64, no. 2, pp. 291-7, 1986.
[302] S. Rahman and F. Nessa, "Neo-natal mortality patterns in rural Bangladesh," (in eng), *J Trop Pediatr,* vol. 35, no. 4, pp. 199-202, Aug 1989.
[303] S. Rajindrajith, S. Mettananda, D. Adihetti, R. Goonawardana, and N. M. Devanarayana, "Neonatal mortality in Sri Lanka: timing, causes and distribution," (in eng), *J Matern Fetal Neonatal Med,* vol. 22, no. 9, pp. 791-6, Sep 2009, doi: 10.3109/14767050902994549.

[304] M. E. Samms-Vaughan, A. M. McCaw-Binns, D. C. Ashley, and K. Foster-Williams, "Neonatal mortality determinants in Jamaica," (in eng), *J Trop Pediatr,* vol. 36, no. 4, pp. 171-5, Aug 1990.
[305] R. Schumacher, E. Swedberg, M. Diallo, D. Keita, and H. Kalter, "Mortality study in Guinea. Investigating the causes of death in children under 5," *Basic Support for Institutionalizing Child Survival, USAID (BASICS II),* 2002.
[306] P. Setel, D. Whiting, and Y. Hemed, "Adult Mortality and Morbidity Project," Tanzania Ministry of Health, 2004.
[307] M. Shah, N. Khalique, Z. Khan, and A. Amir, "Verbal autopsy to determine causes of deaths among under-five children," *Current Pediatric Research,* vol. 14, no. 1, 2010.
[308] G. R. Sharifzadeh, "An Epidemiological Study on Infant Mortality and Factors Affecting it in Rural Areas of Birjand, Iran," *Iranian Journal of Pediatrics,* vol. 18, no. 4, 2008.
[309] S. P. Shrivastava, A. Kumar, and A. Kumar Ojha, "Verbal autopsy determined causes of neonatal deaths," (in eng), *Indian Pediatr,* vol. 38, no. 9, pp. 1022-5, Sep 2001.
[310] P. K. Singhal, G. P. Mathur, S. Mathur, and Y. D. Singh, "Neonatal morbidity and mortality in ICDS urban slums," (in eng), *Indian Pediatr,* vol. 27, no. 5, pp. 485-8, May 1990.
[311] C. Sivagnanasundram, N. Sivarajah, and A. Wijayaratnam, "Infant deaths in a health unit area of Northern Sri Lanka," (in eng), *J Trop Med Hyg,* vol. 88, no. 6, pp. 401-6, Dec 1985.
[312] S. S. Tikmani, H. J. Warraich, F. Abbasi, A. Rizvi, G. L. Darmstadt, and A. K. Zaidi, "Incidence of neonatal hyperbilirubinemia: a population-based prospective study in Pakistan," (in eng), *Trop Med Int Health,* vol. 15, no. 5, pp. 502-7, May 2010, doi: 10.1111/j.1365-3156.2010.02496.x.
[313] E. Turnbull *et al.*, "Causes of stillbirth, neonatal death and early childhood death in rural Zambia by verbal autopsy assessments," (in eng), *Trop Med Int Health,* vol. 16, no. 7, pp. 894-901, Jul 2011, doi: 10.1111/j.1365-3156.2011.02776.x.
[314] A. Vaid, A. Mammen, B. Primrose, and G. Kang, "Infant mortality in an urban slum," (in eng), *Indian J Pediatr,* vol. 74, no. 5, pp. 449-53, May 2007.
[315] P. Waiswa, K. Kallander, S. Peterson, G. Tomson, and G. W. Pariyo, "Using the three delays model to understand why newborn babies die in eastern Uganda," (in eng), *Trop Med Int Health,* vol. 15, no. 8, pp. 964-72, Aug 2010, doi: 10.1111/j.1365-3156.2010.02557.x.
[316] G. Walraven, "Farafenni Neonatal Deaths.," ed. Gambia: MRC, 2003.
[317] D. Woods, "Perinatal Audit System Database, 1 January-31 December 2001," Cape Town Metropolitan Area, South Africa, 2001.
[318] K. M. Yassin, "Indices and sociodemographic determinants of childhood mortality in rural Upper Egypt," (in eng), *Soc Sci Med,* vol. 51, no. 2, pp. 185-97, Jul 2000.
[319] G. T. Debelew, M. F. Afework, and A. W. Yalew, "Determinants and causes of neonatal mortality in Jimma Zone, Southwest Ethiopia: a multilevel analysis of prospective follow up study," (in eng), *PLoS One,* vol. 9, no. 9, p. e107184, 2014, doi: 10.1371/journal.pone.0107184.
[320] APHI/MoPH, CSO/Afghanistan, ICF Macro, IIHMR, and WHO/EMRO, "Afghanistan Mortality Survey 2010," APHI/MoPH, CSO/Afghanistan, ICF Macro, IIHMR, and WHO/EMRO, Calverton, Maryland, USA, 2011. [Online]. Available: http://dhsprogram.com/pubs/pdf/FR248/FR248.pdf
[321] NIPORT/Bangladesh, Mitra and Associates/Bangladesh, and ICF International, "Bangladesh Demographic and Health Survey 2011," NIPORT, Mitra and Associates, and ICF International, Dhaka, Bangladesh, 2013. [Online]. Available: http://dhsprogram.com/pubs/pdf/FR265/FR265.pdf
[322] V. Dogra, R. Khanna, A. Jain, A. M. Kumar, H. D. Shewade, and S. S. Majumdar, "Neonatal mortality in India's rural poor: Findings of a household survey and verbal

autopsy study in Rajasthan, Bihar and Odisha," (in eng), *J Trop Pediatr,* vol. 61, no. 3, pp. 210-4, Jun 2015, doi: 10.1093/tropej/fmv013.
[323] E. Fottrell *et al.*, "The effect of increased coverage of participatory women's groups on neonatal mortality in Bangladesh: A cluster randomized trial," *JAMA pediatrics,* vol. 167, no. 9, pp. 816-825, 2013.
[324] Y. Jain, M. Bansal, R. Tiwari, and P. K. Kasar, "Causes of neonatal mortality: A community based study using verbal autopsy tool," *Nat. J. Comm. Med,* vol. 4, no. 3, pp. 498-502, 2013.
[325] Y. Ma *et al.*, "Cause of death among infants in rural western China: a community-based study using verbal autopsy," (in eng), *J Pediatr,* vol. 165, no. 3, pp. 577-84, Sep 2014, doi: 10.1016/j.jpeds.2014.04.047.

Appendix A: Further details on timing of neonatal deaths

A.1 Description of included and excluded input data

A.1.1 Vital registration data

In the analysis, vital registration (VR) data from 57 countries were used as reported. Key exclusion criteria for VR data included: 1) <50 neonatal deaths, 2) <20% of neonatal deaths on day 0, and 3) <80% adult VR coverage. Table A.1 shows the details of the VR data used for this work, and Table A.2 includes information about VR data that we excluded.

Table A.1: Vital registration data included in the timing of neonatal deaths analysis

Country	Year	MDG region[1]	Neonatal deaths (#)	Country	Year	MDG region	Neonatal deaths (#)
Argentina	2010	LAC	5891	**Kuwait**	2009	WA	363
Australia	2006	DR	774	**Kyrgyzstan**	2010	CCA	2396
Austria	2010	DR	214	**Lithuania**	2010	DR	83
Bahrain	2009	WA	56	**Luxembourg**	2006-10	DR	52
Belgium	2006	DR	315	**Macedonia**	2010	DR	134
Belize	2007	LAC	68	**Maldives**	2008	SA	59
Brazil	2010	LAC	27687	**Malta**	2008-10	DR	60
Bulgaria	2010	DR	394	**Mauritius**	2010	SSA	121
Canada	2009	DR	1404	**Mexico**	2010	LAC	18002
Chile	2009	LAC	1359	**Moldova**	2010	DR	299
Colombia	2009	LAC	5945	**Netherlands**	2010	DR	509
Costa Rica	2009	LAC	462	**New Zealand**	2008	DR	188
Croatia	2010	DR	145	**Norway**	2010	DR	104
Cyprus	2008-10	DR	59	**Panama**	2009	LAC	459
Czech Republic	2010	DR	196	**Paraguay**	2009	LAC	1158
Denmark	2006	DR	174	**Poland**	2010	DR	1454
Dominica	2006-10	LAC	56	**Serbia**	2010	DR	316
Ecuador	2010	LAC	1809	**Slovakia**	2010	DR	217
Estonia	2009-10	DR	80	**Slovenia**	2009-10	DR	74
Finland	2010	DR	91	**South Africa**	2009	SSA	13360
France	2009	DR	1852	**Spain**	2010	DR	1025
Georgia	2010	CCA	506	**Sweden**	2010	DR	183
Germany	2010	DR	1541	**Switzerland**	2010	DR	510
Greece	2010	DR	568	**Trinidad / Tobago**	2008	LAC	219
Hungary	2010	DR	313	**United Kingdom**	2010	DR	2414
Iceland	2005-9	DR	57	**United States**	2008	DR	18091
Ireland	2010	DR	188	**Uruguay**	2009	LAC	209
Italy	2009	DR	1470	**Venezuela**	2007	LAC	5857
Japan	2010	DR	1167				

[1] LAC = Latin America and Caribbean; CCA = Caucasus and Central Asia; DR = Developed regions; SA = Southern Asia; SSA = Sub-Saharan Africa; WA = Western Asia

Table A.2: Vital registration data excluded from the timing of neonatal deaths analysis based on exclusion criteria

Country	Year	MDG region[1]	Reason for exclusion
Albania	2004	DR	<80% adult VR coverage
Antigua and Barbuda	2009	LAC	<20% deaths on day 0
Armenia	2010	CCA	<80% adult VR coverage
Azerbaijan	2007	CCA	<80% adult VR coverage
Bahamas	2008	LAC	<20% deaths on day 0
Barbados	2008	LAC	<20% deaths on day 0
Belarus	2009	DR	no day of death breakdown
Bosnia and Herzegovina	1991	DR	no day of death breakdown
Brunei Darussalam	2009	SEA	no day of death breakdown
Cuba	2010	LAC	<20% deaths on day 0
Dominican Republic	2005	LAC	<80% adult VR coverage
Egypt	2010	NA	no day of death breakdown
El Salvador	2009	LAC	<80% adult VR coverage
Fiji	2009	Oceania	<20% deaths on day 0
Grenada	2010	LAC	<20% deaths on day 0
Guatemala	2008	LAC	<20% deaths on day 0
Guyana	2008	LAC	<80% adult VR coverage
Haiti	2003	LAC	<80% adult VR coverage
Honduras	1990	LAC	<80% adult VR coverage
Iraq	2008	WA	<80% adult VR coverage
Israel	2009	DR	no day of death breakdown
Jamaica	2006	LAC	<80% adult VR coverage
Jordan	2009	WA	<80% adult VR coverage
Kazakhstan	2010	CCA	no day of death breakdown
Kiribati	2001	Oceania	no day of death breakdown
Latvia	2010	DR	no deaths on days 1-6
Malaysia	2008	SEA	<80% adult VR coverage
Mongolia	1994	EA	<80% adult VR coverage
Montenegro	2009	DR	<20% deaths on day 0
Morocco	2008	NA	<80% adult VR coverage
Nicaragua	2010	LAC	<80% adult VR coverage
Oman	2009	WA	<80% adult VR coverage
Peru	2007	LAC	<80% adult VR coverage
Philippines	2008	SEA	no day of death breakdown
Portugal	2010	DR	no day of death breakdown
Qatar	2009	WA	no day of death breakdown
Republic of Korea	2010	EA	no day of death breakdown
Romania	2010	DR	<20% deaths on day 0
Russian Federation	2010	DR	no day of death breakdown
Saint Lucia	2008	LAC	<20% deaths on day 0
St Vincent/Grenadines	2010	LAC	<20% deaths on day 0
Saudi Arabia	2009	WA	<80% adult VR coverage
Seychelles	2008	SSA	no day of death breakdown

Country	Year	MDG region[1]	Reason for exclusion
Singapore	2010	SEA	<80% adult VR coverage
Sri Lanka	2006	SA	<20% deaths on day 0
Suriname	2009	LAC	<20% deaths on day 0
Tajikistan	2005	CCA	<80% VR coverage
Thailand	2006	SEA	<80% adult VR coverage
Turkmenistan	1998	CCA	<80% adult VR coverage
Ukraine	2010	DR	no day of death breakdown
Uzbekistan	2005	CCA	no day of death breakdown
Zimbabwe	1990	SSA	<80% adult VR coverage

[1] LAC = Latin America and Caribbean; DR = CCA = Caucasus and Central Asia; Developed regions; EA = Eastern Asia; NA = Northern Africa; SA = Southern Asia; SEA = South-eastern Asia; SSA = Sub-Saharan Africa; WA = Western Asia

A.1.2 Demographic and Health Surveys

We used data from 206 Demographic and Health Surveys (DHS) as inputs into our model in this analysis. Table A.3 includes information about these included surveys. No DHS available at the time of this analysis were excluded.

Table A.3: Demographic and Health Surveys included in timing of neonatal deaths analysis

Country	Year	MDG region[1]	Neonatal deaths (#)	Country	Year	MDG region	Neonatal deaths (#)
Azerbaijan	2006	CCA	68	**Kyrgyzstan**	1997	CCA	61
Albania	2008	DR	16	**Lesotho**	2009	SSA	173
Armenia	2010	CCA	11	**Lesotho**	2004	SSA	158
Armenia	2005	CCA	25	**Liberia**	2007	SSA	174
Armenia	2000	CCA	33	**Liberia**	1986	SSA	338
Bangladesh	2007	SA	220	**Madagascar**	2008	SSA	304
Bangladesh	2004	SA	286	**Madagascar**	2003	SSA	190
Bangladesh	1999	SA	283	**Madagascar**	1997	SSA	256
Bangladesh	1996	SA	281	**Madagascar**	1992	SSA	215
Bangladesh	1993	SA	358	**Malawi**	2010	SSA	616
Benin	2006	SSA	490	**Malawi**	2004	SSA	282
Benin	2001	SSA	201	**Malawi**	2000	SSA	492
Benin	1996	SSA	184	**Malawi**	1992	SSA	187
Bolivia	2008	LAC	235	**Maldives**	2009	SA	39
Bolivia	2003	LAC	276	**Mali**	2006	SSA	633
Bolivia	1998	LAC	229	**Mali**	2001	SSA	737
Bolivia	1994	LAC	221	**Mali**	1995	SSA	609
Bolivia	1989	LAC	208	**Mali**	1987	SSA	178
Botswana	1998	SSA	72	**Mauritania**	2000	SSA	220
Brazil	1996	LAC	82	**Moldova**	2005	DR	8
Brazil	1986	LAC	116	**Morocco**	2003	NA	160
Burkina Faso	2010	SSA	421	**Morocco**	1992	NA	162
Burkina Faso	2003	SSA	312	**Morocco**	1987	NA	242
Burkina Faso	1998	SSA	239	**Mozambique**	2003	SSA	385
Burkina Faso	1993	SSA	262	**Mozambique**	1997	SSA	393

Country	Year	MDG region[1]	Neonatal deaths (#)	Country	Year	MDG region	Neonatal deaths (#)
Burundi	2010	SSA	241	**Namibia**	2006	SSA	116
Burundi	1987	SSA	134	**Namibia**	2000	SSA	79
Cambodia	2010	SEA	217	**Namibia**	1992	SSA	118
Cambodia	2005	SEA	208	**Nepal**	2011	SA	178
Cambodia	2000	SEA	298	**Nepal**	2006	SA	180
Cameroon	2011	SSA	352	**Nepal**	2001	SA	263
Cameroon	2004	SSA	227	**Nepal**	1996	SA	352
Cameroon	1998	SSA	151	**Nicaragua**	2001	LAC	107
Cameroon	1991	SSA	112	**Nicaragua**	1998	LAC	136
CAR[2]	1994	SSA	196	**Niger**	2006	SSA	327
Chad	2004	SSA	228	**Niger**	1998	SSA	359
Chad	1996	SSA	330	**Niger**	1992	SSA	284
Colombia	2010	LAC	183	**Nigeria**	2008	SSA	1102
Colombia	2005	LAC	169	**Nigeria**	2003	SSA	290
Colombia	2000	LAC	68	**Nigeria**	1999	SSA	215
Colombia	1995	LAC	93	**Nigeria**	1990	SSA	343
Colombia	1990	LAC	40	**Pakistan**	2006	SA	481
Colombia	1986	LAC	52	**Pakistan**	1990	SA	307
Comoros	1996	SSA	74	**Paraguay**	1990	LAC	74
Congo	2005	SSA	160	**Peru**	2007	LAC	85
Cote d'Ivoire	1998	SSA	135	**Peru**	2004	LAC	81
Cote d'Ivoire	1994	SSA	282	**Peru**	2000	LAC	225
DRC[3]	2007	SSA	368	**Peru**	1996	LAC	372
Dominican R[4]	2007	LAC	234	**Peru**	1991	LAC	211
Dominican R	2002	LAC	233	**Peru**	1986	LAC	111
Dominican R	1999	LAC	8	**Philippines**	2008	SEA	99
Dominican R	1996	LAC	117	**Philippines**	2003	SEA	119
Dominican R	1991	LAC	87	**Philippines**	1998	SEA	127
Dominican R	1986	LAC	171	**Philippines**	1993	SEA	152
Ecuador	1987	LAC	106	**Rwanda**	2010	SSA	245
Egypt	2008	NA	170	**Rwanda**	2007	SSA	157
Egypt	2005	NA	270	**Rwanda**	2005	SSA	313
Egypt	2000	NA	269	**Rwanda**	2000	SSA	345
Egypt	1995	NA	342	**Rwanda**	1992	SSA	214
Egypt	1992	NA	281	**Sao Tome/ Principe**	2008	SSA	34
Egypt	1988	NA	329	**Senegal**	2010	SSA	323
Eritrea	2002	SSA	148	**Senegal**	2005	SSA	353
Eritrea	1995	SSA	100	**Senegal**	1997	SSA	264
Ethiopia	2011	SSA	435	**Senegal**	1992	SSA	188
Ethiopia	2005	SSA	436	**Senegal**	1986	SSA	196
Ethiopia	2000	SSA	588	**Sierra Leone**	2008	SSA	210
Gabon	2000	SSA	121	**South Africa**	1998	SSA	96
Ghana	2008	SSA	89	**Sri Lanka**	1987	SA	62
Ghana	2003	SSA	155	**Sudan**	1989	SSA	285
Ghana	1998	SSA	93	**Swaziland**	2006	SSA	61
Ghana	1993	SSA	152	**Tanzania**	2010	SSA	212
Ghana	1988	SSA	180	**Tanzania**	2004	SSA	275
Guatemala	1998	LAC	110	**Tanzania**	1999	SSA	129
Guatemala	1995	LAC	232	**Tanzania**	1996	SSA	215
Guatemala	1987	LAC	147	**Tanzania**	1991	SSA	293
Guinea	2005	SSA	251	**Thailand**	1987	SEA	75

Country	Year	MDG region[1]	Neonatal deaths (#)	Country	Year	MDG region	Neonatal deaths (#)
Guinea	1999	SSA	284	**Timor-Leste**	2009	SEA	206
Guyana	2009	LAC	44	**Togo**	1998	SSA	272
Haiti	2005	LAC	138	**Togo**	1988	SSA	121
Haiti	2000	LAC	205	**Trinidad/Tobago**	1987	LAC	45
Haiti	1994	LAC	111	**Tunisia**	1988	NA	117
Honduras	2005	LAC	145	**Turkey**	1998	WA	87
India	2005	SA	2164	**Turkey**	1993	WA	108
India	1998	SA	2379	**Turkmenistan**	2000	CCA	125
India	1992	SA	2918	**Uganda**	2006	SSA	220
Indonesia	2007	SEA	310	**Uganda**	2000	SSA	246
Indonesia	2002	SEA	287	**Uganda**	1995	SSA	197
Indonesia	1997	SEA	348	**Uganda**	1988	SSA	215
Indonesia	1994	SEA	506	**Ukraine**	2007	DR	11
Indonesia	1991	SEA	450	**Uzbekistan**	1996	CCA	57
Indonesia	1987	SEA	221	**Viet Nam**	2002	SEA	29
Jordan	2009	WA	142	**Viet Nam**	1997	SEA	60
Jordan	2007	WA	134	**Yemen**	1997	WA	418
Jordan	2002	WA	91	**Yemen**	1991	WA	248
Jordan	1997	WA	120	**Zambia**	2007	SSA	220
Jordan	1990	WA	180	**Zambia**	2001	SSA	243
Kazakhstan	1999	CCA	47	**Zambia**	1996	SSA	247
Kazakhstan	1995	CCA	26	**Zambia**	1992	SSA	262
Kenya	2008	SSA	182	**Zimbabwe**	2010	SSA	170
Kenya	2003	SSA	201	**Zimbabwe**	2005	SSA	121
Kenya	1998	SSA	146	**Zimbabwe**	1999	SSA	98
Kenya	1993	SSA	155	**Zimbabwe**	1994	SSA	92
Kenya	1989	SSA	188	**Zimbabwe**	1988	SSA	88

[1] LAC = DR = CCA = Caucasus and Central Asia; Developed regions; EA = Eastern Asia; Latin America and Caribbean; NA = Northern Africa; SA = Southern Asia; SEA = South-eastern Asia; SSA = Sub-Saharan Africa; WA = Western Asia; [2] CAR = Central African Republic; [3] DRC = Democratic Republic of the Congo; [4] Dominican R = Dominican Republic

A.2 Country groupings by MDG region and income

We used MDG regions and World Bank income level categories to classify the 186 countries we included in this work into region and income groups, respectively. The countries in each category are below.

A.2.1 Country groupings by region

Developed regions (DR) – 46 countries

Albania, Australia, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Latvia, Lithuania, Luxembourg, Macedonia (TFYR of), Malta, Montenegro, Netherlands, New Zealand, Norway, Poland, Portugal, Republic of Moldova, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom, United States

North Africa (NA) – 5 countries

Algeria, Egypt, Libya, Morocco, Tunisia

Sub-Saharan Africa (SSA) – 49 countries

Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Democratic Republic of the Congo, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Sudan, Swaziland, Togo, Uganda, United Republic of Tanzania, Zambia, Zimbabwe

South-Eastern Asia (SEA) – 11 countries

Brunei Darussalam, Cambodia, Indonesia, Lao People's Democratic Republic, Malaysia, Myanmar, Philippines, Singapore, Thailand, Timor-Leste, Viet Nam

Eastern Asia – 4 countries

China, Democratic People's Republic of Korea, Mongolia, Republic of Korea

Southern Asia (SA) – 8 countries

Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, Sri Lanka

Western Asia (WA) – 14 countries

Bahrain, Iran, Iraq, Jordan, Kuwait, Lebanon, Occupied Palestinian Territory, Oman, Qatar, Saudi Arabia, Syrian Arab Republic, Turkey, United Arab Emirates, Yemen

Caucasus & Central Asia – 8 countries

Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan

Oceania – 9 countries

Fiji, Kiribati, Marshall Islands, Micronesia (Federated States of), Papua New Guinea, Samoa, Solomon Islands, Tonga, Vanuatu

Latin America & the Caribbean – 32 countries

Antigua and Barbuda, Argentina, Bahamas, Barbados, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Guyana, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, Uruguay, Venezuela

Source: Millennium Development Goals Report 2012
http://www.un.org/millenniumgoals/pdf/MDG%20Report%202012.pdf

A.2.2 Country groupings by income

We used the 2010 World Bank income categories, which were defined as a GNI per capita as follows: 1) low-income: $1,005 or less; 2) lower-middle-income: $1,006 to $3,975; 3) upper-middle-income: $3,976 to $12,275, and 4) high-income: $12,276 or more. The World Bank did not include the Occupied Palestinian Territory in their groupings. The full list of countries by income level is as follows:

Low-income – 35 countries

Afghanistan, Bangladesh, Benin, Burkina Faso, Burundi, Cambodia, Central African Republic, Chad, Comoros, Democratic Republic of the Congo, Democratic People's Republic of Korea, Eritrea, Ethiopia, Gambia, Guinea, Guinea-Bissau, Haiti, Kenya, Kyrgyzstan, Liberia, Madagascar, Malawi, Mali, Mozambique, Myanmar, Nepal, Niger, Rwanda, Sierra Leone, Somalia, Tajikistan, Togo, Uganda, United Republic of Tanzania, Zimbabwe

Lower-middle-income ($1,006 to $3,975) – 54 countries

Angola, Armenia, Belize, Bhutan, Bolivia, Cameroon, Cape Verde, Congo, Cote d'Ivoire, Djibouti, Egypt, El Salvador, Fiji, Georgia, Ghana, Guatemala, Guyana, Honduras, India, Indonesia, Iraq, Kiribati, Lao People's Democratic Republic, Lesotho, Marshall Islands, Mauritania, Micronesia (Federated States of), Mongolia, Morocco, Nicaragua, Nigeria, Pakistan, Papua New Guinea, Paraguay, Philippines, Republic of Moldova, Samoa, Sao Tome and Principe, Senegal, Solomon Islands, South Sudan, Sri Lanka, Sudan, Swaziland, Syrian Arab Republic, Timor-Leste, Tonga, Turkmenistan, Ukraine, Uzbekistan, Vanuatu, Viet Nam, Yemen, Zambia

Upper-middle-income ($3,976 to $12,275) – 50 countries

Albania, Algeria, Antigua and Barbuda, Argentina, Azerbaijan, Belarus, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Chile, China, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, Gabon, Grenada, Iran, Jamaica, Jordan, Kazakhstan, Latvia, Lebanon, Libya, Lithuania, Macedonia (TFYR of), Malaysia, Maldives, Mauritius, Mexico, Montenegro, Namibia, Panama, Peru, Romania, Russian Federation, Saint Lucia, Saint Vincent and the Grenadines, Serbia, Seychelles, South Africa, Suriname, Thailand, Tunisia, Turkey, Uruguay, Venezuela

High-income ($12,276) – 46 countries

Australia, Austria, Bahamas, Bahrain, Barbados, Belgium, Brunei Darussalam, Canada, Croatia, Cyprus, Czech Republic, Denmark, Equatorial Guinea, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Luxembourg, Malta, Netherlands, New Zealand, Norway, Oman, Poland, Portugal, Qatar, Republic of Korea, Saudi Arabia, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, Trinidad and Tobago, United Arab Emirates, United Kingdom, United States

Source: World Bank, http://data.worldbank.org/indicator; http://www.un.org/en/development/desa/policy/wesp/wesp_current/2012country_class.pdf

Appendix B: Further details on cause-of-death estimation by neonatal period

B.1 Country groupings by estimation method and MDG region

B.1.1 Country groupings by estimation method

Grouping by estimation method

The 194 countries in this analysis were separated into 3 groups based on the quality if their vital registration (VR) data and under-5 mortality rates (U5MR). U5MR was used instead of the neonatal mortality rate (NMR) in order to be consistent with Child Health Epidemiology Reference Group (CHERG) researchers working in parallel to develop estimates for children 1-59 months old. We kept the same country groupings as previous work [15], with the exception of a few countries (see "Changes to country groupings" in section 4.2.6).

As defined in the previous work [15], countries were considered to have high-quality VR data if they 1) had 80% or higher coverage, 2) did not have excessive use of non-specific/ill-defined VR codes, and 3) provided sufficient details in the coding such that the deaths could be grouped in the programmatically-relevant categories used in this work. Countries that lacked high-quality VR data were included in the low mortality model if their U5MR from 2000-2010 was ≤35 (per 1,000 live births) and in the high mortality model if their U5MR was >35.

High-quality VR – 65 countries

Antigua and Barbuda, Argentina, Australia, Austria, Bahamas, Bahrain, Barbados, Belgium, Belize, Brazil, Bulgaria, Chile, Colombia, Costa Rica, Croatia, Cuba, Czech Republic, Denmark, Dominica, Estonia, Finland, France, Germany, Greece, Grenada, Guyana, Hungary, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Latvia, Lithuania, Luxembourg, Macedonia (TFYR of), Malta, Mauritius, Mexico, Montenegro, Netherlands, New Zealand, Norway, Panama, Poland, Republic of Korea, Republic of Moldova, Romania, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Serbia, Singapore, Slovakia, Slovenia, South Africa, Spain, Suriname, Sweden, Trinidad and Tobago, United Kingdom, United States, Uruguay, Venezuela

Low mortality model – 49 countries

Albania, Andorra, Armenia, Belarus, Bosnia and Herzegovina, Brunei Darussalam, Canada, Cabo Verde, China, Cook Islands, Cyprus, Ecuador, Egypt, El Salvador, Fiji, Georgia, Honduras, Jamaica, Jordan, Lebanon, Libya, Malaysia, Maldives, Monaco, Nicaragua, Niue, Oman, Palau, Paraguay, Peru, Portugal, Qatar, Russian Federation, Samoa, San Marino, Saudi Arabia, Seychelles, Sri

Lanka, Switzerland, Syrian Arab Republic, Thailand, Tonga, Tunisia, Turkey, Tuvalu, Ukraine, United Arab Emirates, Vanuatu, Vietnam

High mortality model – 80 countries

Afghanistan, Algeria, Angola, Azerbaijan, Bangladesh, Benin, Bhutan, Bolivia, Botswana, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Democratic People's Republic of Korea, Democratic Republic of the Congo, Djibouti, Dominican Republic, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guatemala, Guinea, Guinea-Bissau, Haiti, India, Indonesia, Iran, Iraq, Kazakhstan, Kenya, Kiribati, Kyrgyzstan, Lao People's Democratic Republic, Lesotho, Liberia, Madagascar, Malawi, Mali, Marshall Islands, Mauritania, Micronesia, Mongolia, Morocco, Mozambique, Myanmar, Namibia, Nauru, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Philippines, Rwanda, Sao Tome and Principe, Senegal, Sierra Leone, Solomon Islands, Somalia, South Sudan, Sudan, Swaziland, Tajikistan, Timor-Leste, Togo, Turkmenistan, Uganda, United Republic of Tanzania, Uzbekistan, Yemen, Zambia, Zimbabwe

B.1.2 Country groupings by region

We used MDG regions as described in Appendix A.2.1, with the following changes:

- Developed regions – 49 countries instead of 46 (added Andorra, Monaco, San Marino)
- Southern Asia – 9 countries instead of 8 (added Iran from Western Asia)
- Western Asia – 12 countries instead of 14 (removed Iran and Occupied Palestinian Territory)
- Oceania – 14 countries instead of 9 (added Cook Islands, Nauru, Niue, Palau, and Tuvalu)
- Latin America & the Caribbean – 33 countries instead of 32 (Saint Kitts and Nevis)

B.2 Details of vital registration and study input data

B.2.1 ICD to CHERG cause-of-death code conversions

Here, I have included the mapping from the 9th and 10th International Classification of Diseases (ICD-9 and ICD-10) to the 7 programmatically relevant cause categories we used in our work for the VR data (Table B.1). This table has been updated for subsequent work. The most recent version is available on the "Disease burden and mortality estimates" section of the WHO website: http://www.who.int/healthinfo/global_burden_disease/estimates/en/index3.html

Table B.1: Mapping between ICD codes and CHERG cause categories used in neonatal cause-of-death modelling work

	ICD-10 codes	ICD-9 codes
Complications of preterm birth	P01.0-P01.1, P07, P22, P25-P28, P52, P61.2, P77	434.9, 518.1-518.9, 761.0-761.1, 765, 769-770.0, 770.2-770.9, 772.1, 774.2, 776.6, 777.5-777.6, 786.3
Intrapartum-related complications	P01.7-P02.1, P02.4-P02.6, P03, P07, P10-P15, P20-P21, P24, P50, P90-P91	348.1-348.9, 437.1-437.9, 723.4, 761.7-762.1, 762.4-762.6, 763, 767-768, 770.1, 772.2, 779.0-779.2
Congenital disorders	D55-D68.9, E01-E07, E70-E84, G10-G99, H, I, K, L, M, N, P35, P76, Q	056, 240-243, 245-259, 272-277, 279.3-286, 288.2, 303, 330-348.0, 349-426, 429-434.0, 435-437.0, 438-451, 520-723.0, 724-728, 731-759, 775.2, 777.0, 795.2
Sepsis and other severe infections	A00-A35, A38-A99, B, G00-G09, P36-P39	000-031, 034-055, 057-134, 136-139, 320-326, 491, 730, 771, 780.6, 785.4
Pneumonia	A36-A37, J, P23	032-033, 460-490, 492-518.0
Injuries	S, V, W, X, Y	800-999
Other	C, D00-D54.9, D69-D99, E00, E08-E69, E85-E99, P00, P01.2-P01.6, P02.2-P02.3, P02.7-P02.9, P04-P06, P08, P29, P51, P53-P61.1, P61.3-P74, P78, P80-P83, P93-P94	135, 140-239, 244, 260-271, 278-279.2, 287-288.1, 288.3-289, 427, 452-459, 760, 761.2-761.6, 762.2-762.3, 762.7-762.9, 764, 766, 772.0, 772.3-774.1, 774.3-775.1, 775.3-776.5, 776.7-776.9, 778.0, 779.5-779.6

We excluded the following ICD codes as non-specific/ill-defined:

1) ICD-10 – F, O, P92, P95-96, R

2) ICD-9 – 295.4, 305.6, 205.9, 308.9, 311.0, 317.0, 319.0, 779.3, 779.8-799 (except for 780.6, 785.4, 786.3, and 795.2 as mentioned above)

B.2.2 Missing vital registration data years

Table B.2 describes the missing data years for the high-quality VR countries used in the low mortality model.

Table B.2: List of years with missing data for high-quality vital registration countries

Country	Years with missing data	Country	Years with missing data
Antigua / Barbuda	2010-2012	**Latvia**	2011-2012
Argentina	2011-2012	**Lithuania**	2011-2012
Australia	2005, 2012	**Luxembourg**	2012
Austria	2012	**Macedonia**	2001-2004, 2011-2012
Bahamas	2009-2012	**Malta**	2012
Bahrain	2010-2012	**Mauritius**	2012
Barbados	2009-2012	**Mexico**	2011-2012
Belgium	2000-2002; 2010-2012	**Montenegro**	2000-2004, 2010-2012
Belize	2008-2012	**Netherlands**	2012
Brazil	2011-2012	**New Zealand**	2010-2012
Bulgaria	2012	**Norway**	2012
Chile	2010-2012	**Panama**	2010-2012
Colombia	2010-2012	**Poland**	2012
Costa Rica	2010-2012	**Republic of Korea**	2012
Croatia	2012	**Republic of Moldova**	none
Cuba	2011-2012	**Romania**	2012
Czech Republic	2012	**Saint Kitts / Nevis**	2009-2012
Denmark	2000, 2010-2012	**Saint Lucia**	2000-2001, 2007, 2009-2012
Dominica	2011-2012	**Saint Vincent / Grenadines**	2011-2012
Estonia	2009, 2012	**Serbia**	2012
Finland	2012	**Singapore**	2012
France	2010-2012	**Slovakia**	2011-2012
Germany	2012	**Slovenia**	2011-2012
Greece	2012	**South Africa**	2000-2005, 2008, 2010-2012
Grenada	2000, 2011-2012	**Spain**	2012
Guyana	2000, 2009-2012	**Suriname**	2010-2012
Hungary	2012	**Sweden**	2011-2012
Iceland	2010-2012	**Trinidad and Tobago**	2003, 2009-2012
Ireland	2000-2006, 2011-2012	**United Kingdom**	2011-2012
Israel	2011-2012	**United States**	2011-2012
Italy	2004-2005, 2011-2012	**Uruguay**	2002-2003, 2005-2006, 2010-2012
Japan	2012	**Venezuela**	2008-2012
Kuwait	2000, 2012		

B.2.3 Search terms used for the high mortality setting cause-of-death literature review

Below is a list of the full search terms for the literature review we performed to find neonatal COD studies in high mortality settings.

Our search covered the following areas (including variations, alternatives, and MESH for these terms):

Cause-specific (one or more of these): haemorrhage, jaundice, abnormality, malformation, neural tube defect, sudden infant death syndrome, congenital malformations, congenital abnormalities, necrotic, respiratory distress, prematurity, preterm birth, asphyxia, tetanus, sepsis, birth injury, intrapartum, birth, cause, foetal alcohol syndrome, rubella, diarrhoea, dysentery, cholera, gastroenteritis, digestive tract infection, pneumonia, respiratory infections, bronchitis, croup, meningitis, encephalitis, meningococcal

AND

Age group (one or more of these): infant/newborn, neonatal, perinatal

AND

Mortality terms (one or more of these): death, mortality, fatality

AND

High mortality countries/regions: Argentina, Bolivia, Brazil, Brazil, Chile, Colombia, Ecuador, French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela, Mexico, Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, Puerto Rico, Panama, West Indies, Antigua, Bahamas, Barbados, Cuba, Dominica, Dominican Republic, Grenada, Guadeloupe, Haiti, Jamaica, Martinique, Antilles, Anguilla, Saint Kitts, St Kitts, Saint Lucia, St Lucia, Saint Vincent, St Vincent, Trinidad, Tobago, Virgin Islands, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan, Borneo, Brunei, Cambodia, East Timor, Indonesia, Laos, Malaysia, Mekong Valley, Myanmar, Burma, Philippines, Singapore, Thailand, Vietnam, Bangladesh, Bhutan, India, Nepal, Pakistan, Sri Lanka, China, Korea, Macao, Mongolia, Taiwan, Afghanistan, Bahrain, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen, Fiji, New Caledonia, Papua New Guinea, Vanuatu, Micronesia, Melanesia, Guam, Palau, Polynesia, Samoa, Tonga, Armenia, Azerbaijan, Georgia, Albania, Estonia, Latvia, Lithuania, Bosnia, Herzegovina, Serbia, Bulgaria, Belarus, Croatia, Czech Republic, Hungary, Macedonia, Moldova, Montenegro, Poland, Romania, Russia, Bashkiria, Dagestan, Slovakia, Slovenia, Ukraine, Cameroon, Central African Republic, Chad, Congo, "Democratic Republic of the Congo", Equatorial Guinea, Gabon, Burundi, Djibouti, Eritrea, Ethiopia, Kenya, Rwanda, Somalia, Sudan, Tanzania, Uganda, Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa,

Swaziland, Zambia, Zimbabwe, Benin, Burkina Faso, Cote d'Ivoire, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, Togo, Algeria, Egypt, Libya, Morocco, Tunisia, Comoros, Madagascar, Mauritius, Reunion, Seychelles, developing country, third world country, less developed, sub-Saharan, Caribbean, Pacific Islands, Mexico, Latin America, South America, Indian Ocean Islands, Central America, Asia, Africa, Far East

AND

NOT "case reports", editorial, comment, practice guideline

B.2.4 Details of study data used as inputs into the high mortality model

Below is a table describing the studies and surveys included in the high mortality model input dataset (Table B.3). New studies have an asterisk (*) in front of the author name. While most studies were included in the dataset as a single observation, those stratified by neonatal period or in other ways (e.g. setting location) were included as multiple observations. The full dataset can be found at the WHO Global Health Observatory (www.who.int/gho).

Table B.3: Details of the high mortality model input dataset

First author	Year published	Country	Additional strata	Median data year	# of causes reported	# of deaths used in analysis
Adazu [239]	2005	Kenya		2002	5	75
Aguilar [240]	1998	Bolivia		1995	8	79
***Akgun [241]**	2012	Turkey		2003	4	34
Aleman [242]	1998	Nicaragua		1993	6	72
Anand [243]	2000	India		1993	8	50
Asling-Monemi [244]	2003	Nicaragua		1995	6	56
Awasthi [245]	1998	India		1994	7	286
Baiden [246]	2006	Ghana		1999	4	1001
Balci [247]	2008	Turkey	early/late	2006	5	68
Bang [248]	1999	India		1994	6	36
***Bapat [249]**	2012	India	early/late	2006	5	102
Baqui [132]	2006	India	early/late	not given	7	477
Barros [250]	1987	Brazil		1982	7	113
Barros [251]	2008	Brazil		2004	8	54
Bassani [252]	2010	India		2002	7	10892
Bezzaoucha [253]	2010	Algeria	early/late	2003	5	2167
Bhatia [254]	1989	Bangladesh		1982	7	513
Bhutta [255]	-	Pakistan		2001	7	152
Campbell [256]	2004	Egypt		2000	7	103

First author	Year published	Country	Additional strata	Median data year	# of causes reported	# of deaths used in analysis
Chowdhury [257]	1996	Bangladesh		1983	5	474
Chowdhury [258]	2005	Bangladesh	early/late	1992	5	43
Chowdhury [146]	2010	Bangladesh		2004	8	332
Darmstadt [259]	2010	Bangladesh	time period	2002	7	56
Darmstadt [259]	2010	Bangladesh	time period	2005	7	125
Datta [260]	1988	India		not given	8	163
Deribew [261]	2007	Ethiopia		2005	8	45
***DHS [262]**	2005	Bangladesh		2002	8	302
***DHS [263]**	2006	Honduras		2004	7	142
***DHS [264]**	2007	Nepal		2004	8	220
***DHS [265]**	2008	Pakistan		2006	8	1484
***DHS [266]**	2009	Mozambique		2007	5	718
Djaja [267]	2003	Indonesia		2001	8	178
Dommisse [268]	1991	South Africa	early/late	1988	7	276
Edmond [269]	2008	Ghana	early/late	2004	5	582
Ekanem [270]	1994	Nigeria		1991	7	24
El-Zibdeh [271]	1988	Saudi Arabia		1983	7	78
Fantahun [272]	1998	Ethiopia		1992	7	47
Fauveau [273]	1990	Bangladesh		1982	8	201
Fikree [274]	2002	Pakistan		1992	8	497
Fonseka [275]	1994	Sri Lanka		1987	7	253
Garenne [276]	2007	Morocco	time period	1987	8	109
Garenne [276]	2007	Morocco	time period	1987	8	329
***Gill [277]**	2011	Zambia		2007	7	58
Gomes [278]	1997	Brazil	early/late	1991	6	138
Greenwood [279]	1987	Gambia		1982	8	32
Halder [280]	2009	Bangladesh		2007	8	49
Hinderaker [281]	2003	Tanzania	early/late	1995	5	71
Jehan [282]	2009	Pakistan	early/late	2004	5	49
Karabulut [283]	2009	Turkey	early/late	2006	6	178
Khalique [284]	1993	India	early/late	1990	8	21
Khan [285]	1993	Pakistan		1986	7	80
***Khanal [264]**	2011	Nepal	early/late	2006	5	183
Khanjanasthiti [286]	1984	Thailand		not given	8	22
Kishan [287]	1988	Libya		1984	6	245
***Krishnan [288]**	2012	India	time period	1983	5	154
***Krishnan [288]**	2012	India	time period	2003	5	106
Leach [133]	1999	Gambia		1992	8	130
Liu [289]	1985	China		1980	7	956
Lucas [290]	1996	Sri Lanka		1993	6	120
Manandhar [291]	2010	Nepal		2007	5	640
***Matendo [292]**	2011	Congo		2005	6	56

First author	Year published	Country	Additional strata	Median data year	# of causes reported	# of deaths used in analysis
Matijasevich [293]	2008	Brazil	time period	1982	4	115
Matijasevich [293]	2008	Brazil	time period	1993	4	88
Matijasevich [293]	2008	Brazil	time period	2004	4	47
Mendieta [294]	2001	Paraguay		1996	6	3638
Nandan [295]	2005	India		2001	8	299
***Nga [296]**	2012	Viet Nam		2009	7	225
Pattison [297]	2013	South Africa		2006	7	45848
Perry [298]	2003	Bangladesh		1997	8	102
Perry [298]	2003	Haiti		1997	8	28
Phukan [299]	1998	India		1994	7	101
Pison [300]	1993	Senegal		1987	6	26
Pratinidhi [301]	1986	India		1982	7	135
Rahman [302]	1989	Bangladesh	early/late	1985	7	69
Rajindrajith [303]	2009	Sri Lanka		1999	6	17946
***RHS [263]**	1997	Honduras		1994	7	121
***RHS [263]**	2002	Honduras		1999	7	143
Samms-Vaughan [304]	1990	Jamaica		1986	7	885
Schumacher [305]	2002	Guinea		1998	8	94
Settel [306]	2004	Tanzania	location	2000	8	75
Settel [306]	2004	Tanzania	location	2000	8	119
Settel [306]	2004	Tanzania	location	2000	8	124
Shah [307]	2010	India		2006	8	22
Sharifzadeh [308]	2008	Iran		2005	7	87
Shrivastava [309]	2001	India	early/late	1995	8	895
Singhal [310]	1990	India	early/late	1984	7	50
Sivagnanasundram [311]	1985	Sri Lanka	early/late	1982	8	46
Tikmani [312]	2010	Pakistan		2005	7	41
***Turnbull [313]**	2011	Zambia		2008	6	35
Vaid [314]	2007	India	early/late	2000	8	102
Waiswa [315]	2010	Uganda		2007	7	58
Walraven [316]	2003	Gambia		2000	6	70
Woods [317]	2001	South Africa		2001	7	253
Yassin [318]	2000	Egypt		1994	8	39

B.3 Key methodological differences between current and previous estimates

Three previous rounds of estimation, from 2005 to 2012, were done to estimate neonatal causes of death. A summary of key methodology differences between these different estimation

rounds is presented in section 4.2. A detailed list of the differences is included below in Table B.4.

Table B.4: Methodological differences between the current and previous CHERG estimates

Publication year	2005	2010	2012	Current
Estimation year(s)	2000	2008	2000-2010	2000-2013
Goal of estimation	*Neonatal cause-of-death distribution for 193 countries in 2000*	Neonatal cause-of-death distribution for 193 countries in *2008*	Neonatal cause-of-death distribution for 193 countries from *2000-2010*	Neonatal cause-of-death distribution for 194 countries from *2000-2013 by early and late periods*
Countries in each model	VR: 45 Low mortality: 37 High mortality: 111	VR: 73 Low mortality: 37 Low/high averaged: 22 High mortality: 61	VR: 61 Low mortality: 51 High mortality: 81	VR: 65 Low mortality: 49 High mortality: 80
Thresholds for classifying countries into high-quality VR, low mortality model, or high mortality model	- High quality VR = >90% coverage - Low mort model = countries without high-quality VR data and NMR<10 (or NMR<15 in WHO EURO/AMRO regions) - High mort model = remaining countries	- High quality VR = >80% coverage - Low mort model = countries without high-quality VR and NMR<15 - averaged model = countries without high-quality VR and NMR between 15-20: low/high mortality models averaged - High mort model = countries without high-quality VR data and NMR>20	- High quality VR = >80% coverage + quality criteria (e.g. limited non-specific/implausible codes) - Low mort model = countries without adequate VR and with U5MR≤35 from 2000-2010. - High mort model = countries without adequate VR and with U5MR>35 from 2000-2010.	- High quality VR = >80% coverage + quality criteria (e.g. limited non-specific/implausible codes) - Low mort model = countries without adequate VR and with U5MR≤35 from 2000-2013. - High mort model = countries without adequate VR and with U5MR>35 from 2000-2013.
Input data	High mortality model: 13685 deaths Low mortality model: 96,797 deaths	High mortality model: 23,220 deaths Low mortality model: 1,005,478 deaths	High mortality model: 56,890 deaths Low mortality model: 1,013,599 deaths	High mortality model: 98,222 deaths Low mortality model:1,267,404 deaths

Publication year	2005	2010	2012	Current
Covariate selection	-High mortality model: expert opinion on which covariates may be associated with outcomes -Low mortality model: forward stepwise with 5% sig. level	Same covariates used as previous round	-High and low mortality models: jackknife procedure to minimize out-of-sample prediction error. - Allowed relationship between covariate and outcome to be linear or quadratic	-High and low mortality models: jackknife procedure to minimize out-of-sample prediction error. - Allowed relationship between covariate and outcome to be linear, quadratic, or spline
Multinomial model	All causes in multinomial	Infections in multinomial with sepsis/pneumonia split done after; tetanus as single-cause	Infections in multinomial with sepsis/pneumonia split done after; tetanus as single-cause	All causes included in multinomial
Causes	Preterm, Intrapartum, Congenital, Infection, Diarrhoea, Tetanus, Other	Preterm, Intrapartum, Congenital, Sepsis, Pneumonia, Diarrhoea, Tetanus, Other	Preterm, Intrapartum, Congenital, Sepsis, Pneumonia, Diarrhoea, Tetanus, Other	Preterm, Intrapartum, Congenital, Sepsis, Pneumonia, Diarrhoea, Tetanus, Other (+ injuries for VR and low mortality model countries)
Uncertainty	Jackknife	Jackknife	Bootstrap	Bootstrap

B.4 Additional results

This section includes additional results for the work presented in chapter 4. These results correspond to sections 4.3 (main results) and 4.4 (sensitivity analyses) of the main thesis.

B.4.1 Comparison of input data and prediction covariate values

Table B.5 is similar to Table 4.5 in section 4.3, but the prediction data are now for Indian states. In the analysis, we predicted the proportional cause-of-death distribution at the state-level for India and aggregated the estimates to develop national estimates.

Table B.5: Comparison of high mortality model input data and covariates for Indian states

	High mortality model input data			Indian state-level prediction data[1]		
	Mean (SD)[2]	Median (IQR)[3]	Range (min-max)	Mean (SD)	Median (IQR)	Range (min-max)
NMR[4]	33.0 (15.9)	30.2 (18.8-47.2)	10.5-70.1	26.8 (11.6)	26.8 (18.1-33.9)	**6.3**-59.9
IMR	62.1 (30.0)	58.7 (35.7-81.6)	14.7-142.0	40.2 (17.3)	38.8 (27.1-52.2)	**9.2**-95.5
U5MR	88.6 (45.8)	89.1 (52.9-125.4)	17.1-227.0	54.9 (25.2)	51.0 (36.1-72.5)	**11.8**-122.0
LBW	18.9 (11.3)	15.9 (10.6-27.6)	2.5-50.0	20.4 (8.4)	21.0 (15.0-26.0)	**0.0**-43.0
GFR	0.127 (0.043)	0.118 (0.092-0.158)	0.057-0.235	0.089 (0.007)	0.088 (0.083-0.096)	0.080-0.103
ANC	67.5 (26.5)	73.9 (50.0-92.0)	5.0-98.3	77.8 (13.2)	75.2 (70.1-88.8)	33.1-**99.5**
DPT	67.9 (24.6)	73.5 (60.5-83.5)	0.0-99.0	68.9 (14.4)	70.6 (61.2-78.1)	30.2-96.9
BCG	78.8 (23.9)	87.0 (73.0-93.0)	0.0-100.0	87.2 (10.9)	89.7 (85.0-96.1)	45.4-99.5
PAB	63.3 (25.0)	68.0 (51.4-83.5)	0.0-98.5	72.6 (15.5)	76.9 (60.6-85.4)	31.3-97.1
FLR	51.9 (24.7)	48.7 (34.6-77.3)	4.0-94.0	62.9 (14.7)	63.8 (52.7-70.3)	35.4-94.0
SBA	48.7 (33.1)	45.3 (18.9-83.7)	0.0-100.0	59.8 (16.8)	59.4 (48.1-70.0)	25.5-99.6

[1] bolded values are those outside the input data range; [2] SD = standard deviation; [3] IQR = interquartile range; [4] see Table 4.2 for covariate acronym definitions, the region covariate was "Southern Asia" for all Indian states

B.4.2 Regression coefficients for the low and high mortality models

The following table includes the regression coefficients from the multinomial regressions for the two models used in this work.

Table B.6: Regression coefficients for the low and high mortality models

	Early neonatal period	Late neonatal period
	Covariates with regression coefficients[1]	**Covariate with regression coefficients**
Low mortality model[1]		
Intrapartum: Preterm	FLR^{L} (-0.018); const (0.572)	FLR^{S1} (-0.030); FLR^{S2} (-0.025); FLR^{S2} (0.134); DPT^{S1} (0.002); DPT^{S2} (0.006); DPT^{S2} (0.064); const (1.173)
Congenital: Preterm	$GINI^{L}$ (-0.008); DPT^{L} (0.015); FLR^{L} (0.004); IMR^{S1} (-1.669); IMR^{S2} (11.454); IMR^{S2} (-13.670); $U5MR^{S1}$ (1.417); $U5MR^{S2}$ (-9.443); $U5MR^{S2}$ (13.523); LBW^{S1} (0.185); LBW^{S2} (-1.133); LBW^{S2} (2.726); const (-0.807)	NMR^{L} (-0.039); DPT^{L} (0.007); const (-0.307)
Sepsis: Preterm	GNI^{L} (-2.6×10^{-5}); $GINI^{L}$ (0.025); ANC^{L} (0.029); IMR^{S1} (-0.274); IMR^{S2} (2.647); IMR^{S2} (-4.009); DPT^{S1} (-0.013); DPT^{S2} (0.034); DPT^{S2} (-0.555); const (-3.267)	FLR^{L} (-0.009); $GINI^{L}$ (0.031); DPT^{S1} (-0.026); DPT^{S2} (0.062); DPT^{S2} (-1.125); IMR (0.049); IMR^{Q} (-0.001); const (0.643)
Pneumonia: Preterm	GNI^{L} (-6.7×10^{-5}); const (-2.078)	GNI^{L} (-4.8×10^{-5}); ANC^{S1} (0.072); ANC^{S2} (-0.205); ANC^{S2} (2.250); const (-6.979)
Injuries: Preterm	GFR^{L} (-104.414); GFR^{Q} (763.714); const (-0.903)	const (-2.942)
Other: Preterm	GFR^{L} (103.863); GFR^{Q} (-908.192); const (-4.231)	LBW^{L} (0.029); NMR^{S1} (-0.212); NMR^{S2} (1.117); NMR^{S2} (-1.514); DPT^{L} (-0.090); DPT^{Q} (0.001); const (2.433)
High mortality model[1]		
Preterm: Intrapartum	BCG^{L} (0.007); PAB^{L} (-0.007); SBA^{L} (0.010); DPT^{L} (-0.004); LBW^{S1} (0.028); LBW^{S2} (-0.006); GFR^{S1} (-11.980); GFR^{S2} (14.787); const (0.602)	LBW^{S1} (0.038); LBW^{S2} (-0.054); PAB^{S1} (-0.011); PAB^{S2} (0.000); GFR^{S1} (-21.739); GFR^{S2} (19.502); SSA (0.305); const (2.764)
Congenital: Intrapartum	LBW^{L} (0.009); NMR^{L} (-0.073); NMR^{Q} (0.001); $U5MR^{L}$ (-0.022); $U5MR^{Q}$ (0.000); BCG^{S1} (0.004); BCG^{S2} (0.003); early (0.223); const (0.364)	SBA^{L} (-0.021); SBA^{Q} (0.000); $U5MR^{L}$ (-0.021); $U5MR^{Q}$ (0.000); late (0.739); SSA (-0.186); const (-0.070)
Sepsis: Intrapartum	LBW^{L} (0.013); BCG^{L} (0.022); BCG^{Q} (0.000); early (-0.765); SA (0.253); const (-2.303)	PAB^{L} (0.019); PAB^{Q} (0.000); LBW^{S1} (0.023); LBW^{S2} (-0.018); late (1.510); const (-1.708)
Pneumonia: Intrapartum	$U5MR^{L}$ (0.006); LBW^{L} (0.009); early (-0.674); const (-1.713)	PAB^{L} (-0.029); PAB^{Q} (0.000); const (-0.092)
Diarrhoea: Intrapartum	DPT^{L} (-0.005); GFR^{L} (12.725); NMR^{L} (0.282); NMR^{Q} (-0.003); early (-1.725); SA (-0.226); SSA (-1.491); const (-8.746)	DPT^{L} (-0.003); BCG^{L} (-0.011); GFR^{L} (-3.800); FLR^{L} (-0.022); LBW^{S1} (0.139); LBW^{S2} (-0.192); const (-1.332)
Tetanus: Intrapartum	PAB^{L} (-0.011); ANC^{L} (-0.018); NMR^{L} (0.043); early (-0.985); const (-1.642)	NMR^{L} (0.038); IMR^{L} (-0.010); $U5MR^{L}$ (0.011); PAB^{L} (-0.020); late (0.839); const (-2.186)
Other: Intrapartum	GFR^{S1} (-22.942); GFR^{S2} (33.572); SSA (-0.746); const (0.824)	GFR^{S1} (-25.195); GFR^{S2} (25.625); late (0.474); SSA (0.317); const (1.317)

[1] see Table 4.2 for covariate acronym definitions; const = constant; most regression coefficients are rounded to 3rd decimal place but GNI rounded to 6th decimal place because GNI values are on order of thousands instead of the much smaller values of the other coefficients

^{L} linear; ^{Q} quadratic; ^{S1} first restricted cubic spline; ^{S2} second restricted cubic spline; ^{S3} third restricted cubic spline

B.4.3 Comparison of estimated cause-of-death proportions between current and previous estimates

Differences in global estimates between previous CHERG estimates and this round are shown in Table B.7. Since time trends are a recent addition, we have compared the data reported in the earlier studies with the relevant year from our current work. For example, the paper published in 2005 had estimates for 2000, so we have compared the 2000 predictions from that paper and our work.

Table B.7: Comparison of estimated proportions between current and previous estimation rounds

	Percentages from previous work (%)	Percentages from current work (%)
2000 estimates [13]		
Preterm	27.9 (0.19-0.35)	33.1 (23.1-44.0)
Intrapartum[1]	22.8 (0.15-0.27)	24.8 (16.4-32.8)
Infections[2]	26.0 (0.17-0.31)	21.5 (10.5-35.9)
Congenital	7.4 (0.06-0.12)	8.3 (5.0-14.2)
Tetanus	6.5 (0.05-0.20)	3.8 (1.0-9.7)
Diarrhoea	2.8 (0.02-0.10)	0.9 (0.1-4.7)
Other	6.6 (0.05-0.16)	7.6 (3.8-13.2)
2008 estimates [14]		
Preterm	28.9 (20.1-34.0)	34.6 (24.4-46.2)
Intrapartum[1]	22.8 (15.7-27.9)	23.8 (16.8-31.5)
Sepsis	14.6 (10.0-20.6)	16.0 (8.2-24.7)
Pneumonia	10.8 (7.4-15.2)	5.5 (2.8-10.4)
Congenital	7.6 (5.7-10.7)	9.6 (6.2-15.0)
Tetanus	1.7 (0.9-2.3)	2.1 (0.6-5.5)
Diarrhoea	2.2 (1.6-5.9)	0.7 (0.1-3.6)
Other	11.4 (8.9-24.7)	7.7 (3.7-13.2)
2010 estimates [15]		
Preterm	30.2 (25.6-37.1)	35.0 (24.6-46.9)
Intrapartum	20.1 (17.1-24.5)	23.6 (16.8-31.0)
Sepsis	11.0 (7.0-15.4)	16.0 (8.2-24.7)
Pneumonia	9.1 (5.8-13.1)	5.2 (2.7-10.0)
Congenital	7.6 (5.8-10.2)	10.1 (6.4-15.6)
Tetanus	1.6 (0.6-7.7)	1.8 (0.5-4.8)
Diarrhoea	1.4 (0.5-4.2)	0.6 (0.2-3.3)
Other	5.1 (3.2-7.9)	7.6 (3.7-13.2)

[1] Intrapartum-related conditions were previously referred to as "birth asphyxia"; [2] infections include sepsis and other severe infections as well as pneumonia, which in recent years have been estimated separately. To compare with previous estimates, however, these are included in the aggregate infection category for estimates.

Note that these estimates are not necessarily directly comparable. One reason for this is that UN-IGME updates their NMR, U5MR, and IMR time series each year, and so the values estimated for 2008 in one year may be different from those from another year. Since our model includes these as covariates, the covariate values for the same year may be different for these values, which could affect the proportions. But as is seen in Table B.7, the previous and current estimates generally fall within each other's uncertainty bounds.

B.4.4 Comparison of different proportional cause-of-death distributions for China

Comparison of proportional COD distributions for China from the WHO, our low mortality model, and our high mortality model are included in Table B.8.

Table B.8: Proportional cause-of-death distribution estimated for China by the WHO, low mortality model, and high mortality model

	Intra-partum	Preterm	Con-genital	Sepsis	Pneumonia	Injuries	Diarrhoea	Tetanus	Other
Estimates from the WHO									
2000	0.35	0.25	0.08	0.07	0.09	0.04	0.02	0.01	0.09
2001	0.34	0.24	0.09	0.06	0.09	0.04	0.02	0	0.11
2002	0.33	0.24	0.09	0.06	0.09	0.04	0.02	0	0.13
2003	0.32	0.24	0.09	0.06	0.09	0.04	0.02	0.01	0.14
2004	0.31	0.23	0.1	0.05	0.08	0.04	0.02	0	0.15
2005	0.3	0.23	0.1	0.05	0.08	0.04	0.02	0	0.17
2006	0.29	0.23	0.11	0.05	0.08	0.04	0.02	0	0.18
2007	0.28	0.23	0.12	0.05	0.07	0.04	0.02	0	0.18
2008	0.28	0.23	0.12	0.04	0.07	0.04	0.02	0	0.19
2009	0.27	0.23	0.13	0.04	0.07	0.04	0.02	0	0.2
2010	0.26	0.23	0.14	0.04	0.07	0.04	0.02	0	0.2
2011	0.25	0.23	0.14	0.04	0.07	0.04	0.02	0	0.21
2012	0.25	0.23	0.15	0.04	0.07	0.04	0.02	0	0.21
2013	0.25	0.23	0.15	0.04	0.07	0.04	0.02	0	0.21
Estimates using the low mortality model[1]									
2000	0.15	0.42	0.15	0.08	0.07	0.01	---	---	0.12
2001	0.15	0.41	0.15	0.09	0.08	0.01	---	---	0.11
2002	0.15	0.42	0.15	0.09	0.07	0.01	---	---	0.11
2003	0.15	0.42	0.16	0.09	0.07	0.01	---	---	0.11
2004	0.14	0.41	0.17	0.09	0.07	0.01	---	---	0.10
2005	0.14	0.41	0.18	0.09	0.07	0.01	---	---	0.10
2006	0.14	0.40	0.20	0.10	0.06	0.01	---	---	0.10
2007	0.14	0.40	0.20	0.10	0.06	0.01	---	---	0.10
2008	0.13	0.38	0.21	0.11	0.06	0.01	---	---	0.10
2009	0.13	0.38	0.23	0.10	0.06	0.01	---	---	0.10
2010	0.13	0.38	0.23	0.10	0.06	0.01	---	---	0.10

	Intra-partum	Preterm	Con-genital	Sepsis	Pneumonia	Injuries	Diarrhoea	Tetanus	Other
2011	0.13	0.38	0.23	0.09	0.05	0.01	---	---	0.10
2012	0.13	0.38	0.24	0.09	0.05	0.01	---	---	0.10
2013	0.13	0.39	0.24	0.09	0.05	0.01	---	---	0.10
Estimates using the high mortality model[2]									
2000	0.18	0.42	0.13	0.09	0.03	---	0	0.01	0.13
2001	0.18	0.41	0.14	0.1	0.03	---	0	0.01	0.13
2002	0.19	0.4	0.15	0.1	0.03	---	0	0.01	0.14
2003	0.19	0.39	0.16	0.1	0.03	---	0	0.01	0.14
2004	0.19	0.37	0.18	0.1	0.03	---	0	0	0.14
2005	0.19	0.36	0.19	0.1	0.03	---	0	0	0.13
2006	0.18	0.35	0.21	0.1	0.03	---	0	0	0.13
2007	0.18	0.35	0.22	0.1	0.03	---	0	0	0.13
2008	0.17	0.34	0.24	0.1	0.02	---	0	0	0.13
2009	0.17	0.33	0.25	0.1	0.02	---	0	0	0.12
2010	0.17	0.33	0.25	0.1	0.02	---	0	0	0.12
2011	0.17	0.33	0.25	0.1	0.02	---	0	0	0.13
2012	0.17	0.33	0.25	0.1	0.02	---	0	0	0.13
2013	0.17	0.32	0.25	0.1	0.02	---	0	0	0.13

[1] Diarrhoea and tetanus are not included in the low mortality model; [2] Injuries are not included in the high mortality model

Appendix C: Further details on issues with the current multinomial models

C.1 Details of additional study input data

Table C.1 describes the studies and surveys which were added in 2015 to the high mortality model input dataset described in Table B.3. Thus, the updated high mortality input dataset includes the studies listed in Table B.3 as well as those listed here in Table C.1. As noted earlier, most studies were included in the dataset as a single observation, but those stratified by neonatal period or in other ways (e.g. setting location) were included as multiple observations.

Table C.1: Details of additional studies/surveys added to the high mortality model input dataset in 2015

First author	Year published	Country	Additional strata	Median data year	# of causes reported	# of deaths used in analysis
Debelew [319]	2014	Ethiopia		2013	6	110
DHS [320]	2011	Afghanistan		2007	7	506
DHS [321]	2013	Bangladesh		2008	8	286
Dogra [322]	2015	India	location	2012	5	46
Dogra [322]	2015	India	location	2012	5	80
Dogra [322]	2015	India	location	2012	5	51
Fottrell [323]	2013	Bangladesh	time period	2008	4	324
Fottrell [323]	2013	Bangladesh	time period	2010	4	249
Jain [324]	2013	India	early/late	2005	6	60
Ma [325]	2014	China	early/late	2011	8	244

Appendix D: Example of guidance document when publishing modelled estimates

Below is an example of a guidance document to aid interpretation of modelled results. This document could be part of or in addition to other necessary documentation (e.g. detailed methods).

Introduction

The purpose of this document is to help guide the interpretation and future use of the neonatal cause-of-death estimates and models from the Maternal Child Epidemiology Estimation Group (MCEE). Potential input data and modelling limitations are included here to give context to the produced estimates. Caution is also advised if trying to adapt the models for further analyses.

Publicly available files

This document should be read in conjunction with the documentation at: https://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html. This site includes detailed descriptions of the modelling methods, result files, and full statistical code for the MCEE neonatal cause-of-death models.

Comments about the input data and interpretation of results

The validity of our estimates relies on the quality of the input/prediction data, and our modelling techniques. The majority of cause-of-death input data came from verbal autopsy (VA) studies for the high mortality model and vital registration (VR) data for the low mortality model. VA data are known to be of variable quality, and the reported cause distributions depend heavily on factors like the causal hierarchies and case definitions used to attribute deaths. Lack of both standardized VA methods and full reporting of methods can make comparisons between studies challenging. Accurate cause attribution using VA is especially problematic for causes that are difficult to distinguish between, such as neonatal sepsis and pneumonia, or difficult to identify, such as congenital disorders without external signs. VA for neonatal deaths has the added complication that the sick baby is unable to communicate symptoms to a caretaker.

High-quality VR data also have limitations. ICD-10 codes are not ideal for neonatal causes, particularly because several programmatically relevant causes are relegated to the often-unused fourth digit in the codes. ICD coding practices can also vary between (and within) countries and over time. Such variations reduce our model's ability to predict true variation in

causes of death. Other issues in VR coding include changes during the transition between ICD revisions, differences in relegating certain causes to non-specific/ill-defined cause categories, and the assumption inherent in our exclusion of such codes that the deaths attributed to them are a random sample of all deaths. Finally, the availability and quality of VR data in a country may change over time, especially in countries with newly emerging surveillance systems. Developing consistent time trend estimates given such changes can be challenging.

There is also heterogeneity in data quality for the covariates in our models, both between and within covariate time series. This is because of a combination of what is being measured, how it is measured, and how the time series data are put together. The covariate data in our models are a mixture of data reported by countries, determined from household surveys, modelled, and/or imputed. A particular covariate time series may have a combination of these sources across countries or over time for the same country. Such variation may not always be known or fully understood, and is usually impossible to specifically identify within national covariate time series data. Specific sources for covariate time series data can be found through the link included in the "Publicly available files" section above.

Regression-based models such as ours inherently depend on the relationship between outcome variables and covariates, which should ideally come from the same population and time period. While we sought to include as much local covariate information as possible for the input studies, 52% of the total covariate information came from national data instead of from local/regional data. Additionally, when re-categorizing reported VA causes of death, we had to make choices, for example placing deaths reported as being due to "very low birth weight" into the preterm complication category. This may introduce a degree of misclassification as some "very low birth weight" deaths may be attributable to congenital abnormalities. We made those choices that we believed would introduce the least misclassification, but until VA methods improve, this will continue to be a challenge. Similar issues exist in ICD coding, but are more common in VA studies because of the limited and lower quality information collected.

Comments about interpretation of the results

When interpreting the estimates from the neonatal cause-of-death models, it is important to remember that the results (e.g. cause-of-death proportions by country and year) are outputs of a complex statistical machinery with several different components (e.g. range of input data of differing quality, various statistical methods).

As with all such modelling exercises, our estimates should be viewed as an interim measure to help policymakers, particularly in settings with little or no data currently. It is important to distinguish *estimates* from *data*. For most countries, our cause-specific estimates are not based on empirical data from that country, but from a model bringing together data from many countries. The model then predicts the cause-of-death distribution, and changes in the cause-of-death distribution, in individual countries based on covariate values for the individual country. Some countries contribute little or no input data to the modelling process. Our estimates should not therefore be interpreted as "tracking" changes in causes of death for the majority of countries, but rather as predictions of what might be occurring in countries. To track changes in burden due to specific causes of death requires each country to collect representative and consistent data on cause of death on a continuing basis. Our estimates are not a panacea for actual data collection.

Comments about adaption of these models for other analyses

Careful consideration should be taken if trying to extend these models beyond their original purpose. An assessment must first be made about whether the existing models can support the desired extension. For example, it may be tempting to apply the estimated covariate coefficients to subnational prediction data. However, if the range of the proposed prediction data is outside the range of the model input data, such an analysis is beyond the scope of these models. This is one example; careful consideration is needed before any desired adaptation of the models. If the model is applied to or adapted for alternative analyses, rigorous checks and validation exercises should be performed to assess suitability and model performance.

Appendix E: Relevant publications

Below are publications that I have been involved with which pertain to the work presented in this thesis. All open-access articles have web links provided.

Liu L, Chu Y, **Oza S**, Hogan D, Perin J, Bassani D, Ram U, Fadel S, Pandey A, Dhingra N, Sahu D, Kumar P, Cibulsis R, Wahl B, Shet A, Mathers C, Lawn J, Jha P, Kumar R, Black R, Cousens S. National, regional, and state-level all-cause and cause-specific under-5 mortality in India in 2000-15: a systematic analysis with implications for the Sustainable Development Goals. *Lancet Global Health.* 2019; 7(6): e721-e734. Available at: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(19)30080-4

Sadoo S, Blencowe H, **Oza S**, Lawn J. Introduction to newborn health. In *Global Health of Women, Newborns, Children, and Adolescents*, Eds. Delan D et al., Oxford University Press, USA. 2018.

Liu L, **Oza S**, Hogan D, Chu Y, Perin J, Zhu J, Lawn J, Cousens S, Mathers C, Black R. Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet*. 2017;388(10063):3027-35. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)31593-8

Liu L, Hill K, **Oza S**, Hogan D, Chu Y, Cousens S, Mathers C, Stanton C, Lawn J, Black R. Chapter 4: Levels and Causes of Mortality under Age Five Years. In: Black RE, Laxminarayan R, Temmerman M, et al., editors. Reproductive, Maternal, Newborn, and Child Health: Disease Control Priorities, 3rd ed (Vol 2). Washington DC: The International Bank for Reconstruction and Development/World Bank; 2016. Available at: https://www.ncbi.nlm.nih.gov/books/NBK361908/

Oza S, Lawn J, Hogan D, Mathers C, Cousens S. Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000-2013. *Bulletin of the WHO*. 2015; 93(1): 19-28. Available at: https://www.who.int/bulletin/volumes/93/1/14-139790/en/

Oza S, Cousens S, Lawn J. Estimation of daily risk of neonatal death, including the day of birth, in 186 countries in 2013: a vital-registration and modelling-based study. *Lancet Global Health.* 2014; 2(11): e635-e644. Available at: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(14)70309-2

Liu L, **Oza S**, Hogan D, Perin J, Rudan I, Lawn J, Cousens S, Mathers C, Black R. Global, regional, and national causes of child mortality in 2000-2013, with projections to inform post-2015 priorities: an updated systematic analysis. *Lancet*. 2014; 385(9966): 430-440. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)61698-6

Lawn J, Blencowe H, **Oza S**, You D, Lee AC, Waiswa P, Lalli M, Bhutta Z, Barros AJD, Christian P, Mathers D, Cousens S, Lancet Every Newborn Study Group. Every Newborn: progress, priorities, and potential beyond survival. *Lancet*. 2014; 384(9938): 189-205. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)60496-7

Save the Children. Surviving the first day: state of the world's mothers report 2013. Available at: https://resourcecentre.savethechildren.net/library/surviving-first-day-state-worlds-mothers-2013