The GATHER Statement: Explanation and Elaboration

The GATHER Working Group

June 2016
Acknowledgments

The GATHER Explanation and Elaboration was written by Margaret C. Hogan, Gretchen A. Stevens, and Daniel R. Hogan with inputs and assistance from other Gather Working Group members: Leontine Alkema, Robert E. Black, J. Ties Boerma, Gary S. Collins, Majid Ezzati, John T. Grove, Richard Horton, Joy E. Lawn, Ana Marušić, Colin D. Mathers, Christopher J.L. Murray, Igor Rudan, Joshua Salomon, Paul J. Simpson, Theo Vos, and Vivian Welch. The GATHER checklist and GATHER statement were prepared by the GATHER working group. We received financial support from the Bill & Melinda Gates Foundation. We gratefully acknowledge comments from Simon I. Hay and Christopher Paciorek.

This document is updated periodically. The current version of the document, together with the GATHER statement and the GATHER checklist can be found on the GATHER website: www.gather-statement.org

For queries or suggestions, please contact gather@who.int

Please cite this document as follows:


© World Health Organization

This work is licensed under the Creative Commons Attribution 3.0 IGO License. To view a copy of this license, visit https://creativecommons.org/licenses/by/3.0/igo/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers’ products does not imply that they are endorsed or recommended by the World Health Organization in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by the World Health Organization to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall the World Health Organization be liable for damages arising from its use. This publication contains the collective views of the GATHER working group and does not necessarily represent the decisions or policies of the World Health Organization.
# Contents

Acknowledgments ................................................................................................................................. 2  
Introduction ............................................................................................................................................... 4  
Background .............................................................................................................................................. 4  
Scope ........................................................................................................................................................ 5  
How to use this document ....................................................................................................................... 6  
The GATHER checklist .............................................................................................................................. 8  
  Objectives and funding ............................................................................................................................ 9  
  Data inputs .............................................................................................................................................. 10  
    Data inputs that are synthesized in the study ..................................................................................... 10  
    Data inputs that were used without modification .............................................................................. 18  
  All data inputs ....................................................................................................................................... 19  
Data analysis ........................................................................................................................................... 20  
Results and discussion ............................................................................................................................. 30  
References ............................................................................................................................................... 36
Introduction

Global, regional, national, and subnational estimates of population health indicators are needed for monitoring health and for guiding resource allocation. The international community’s focus on monitoring progress toward the Millennium Development Goals has underscored the need for this information, as well as the limitations in the data that are available.\(^1\) Even as the MDG era draws to a close, demand for timely, comparable, and reliable global health estimates is increasing, from governments, non-governmental organizations, funding agencies and researchers.\(^2,3\)

However, even basic population health indicators are rarely available for every population and year, and in many cases there are discrepancies in available measurements.\(^4,5\) Drawing comparisons between populations or over time can be complicated by differences in measurement methods and variation in data availability.\(^6\)

These data gaps and measurement challenges have led to the use of a variety of analytic procedures to produce estimates of health indicators. These methods often require assumptions that may affect the interpretation of the estimates they produce, for example, whether and how a particular estimation approach allows the health indicator to vary over time. The need for comparable health indicators has led, in recent years, to statistical models of increasing flexibility and sophistication. These methods may adjust for differences in types or quality of available health data, allowing for comparison across populations, or may allow for the calculation of more comprehensive uncertainty intervals. At the same time, these new methods can lead to confusion about why estimates have changed,\(^7-10\) e.g., whether changes were a result of measured epidemiological changes versus a new analytical method.

Accurate interpretation and responsible use of health estimates requires understanding the input data on which estimates were based, including information on quality, and a clear explanation of the methods used to derive estimates from the input data.\(^11-13\) The need for guidelines for reporting health estimates was a key conclusion of WHO expert meetings in February and December 2013, which were the impetus for the present set of guidelines.\(^14\)

Background

The GATHER working group was convened in 2014, with the aim to define and promote good practice in reporting global health estimates. The working group’s approach was based on published guidance for developing reporting guidelines.\(^15\) All members of WHO’s Reference Group on Global Health Statistics were invited to join the working group; other experts and journal editors with complementary expertise were sought and invited to join. The working group consists of practitioners, including statisticians, from academia and the World Health Organization (WHO), journal editors, representatives of the EQUATOR network,\(^16\) and members of existing guideline steering groups. The working group reviewed existing reporting guidelines for relevance to global health estimates and sought guidance from experts who had previously developed reporting guidelines. The group determined that existing reporting guidelines would not ensure adequate reporting of global health estimates.

Based on the review of existing guidance and reporting guidelines\(^17-22\) and on input from working group members, we generated a comprehensive list of potential reporting items. We subsequently sought feedback from a broader community of researchers and users via an on-line survey between January and February 2015. Working group members distributed the survey to their respective networks. We received 118 responses. The responses were compiled, summarized and presented at a 2-day consensus meeting held in London, United Kingdom, in February 2015.

The primary objective of the working group consensus meeting was to agree on the list of items to be reported whenever health estimates are published. During the meeting, reporting items were evaluated in light of the responses to the online survey and working group members agreed to retain, omit, or combine items to generate the checklist in Table 1.
Our working group and the responses to our online survey, both drawn from our networks of collaborators, were dominated by residents of high-income countries. We therefore sought additional feedback from a geographically diverse group of stakeholders by sharing an earlier version of this statement prior to publication. The statement was revised based on this feedback.

**Scope**

GATHER defines best practices for documenting studies that report global health estimates. Global health estimates include all quantitative population-level estimates (including global, regional, national, or subnational estimates) of health indicators, including indicators of health status such as estimates of total and cause-specific mortality, incidence and prevalence of diseases, injuries, and disability and functioning; and indicators of health determinants, including health behaviours and health exposures (Box 1).

GATHER aims to define best practices for reporting studies that synthesize evidence from multiple sources to quantitatively describe past and current population health and its determinants. These studies include comparisons among multiple populations, over time or by place of residence. GATHER covers reporting of studies that disaggregate disease and injuries by underlying cause as defined by a classification system such as the International Classification of Disease [ICD] as well as those that attribute disease and injury to their determinants, for example, estimating the number of deaths attributable to tobacco smoking.

Health determinants can range from proximal determinants of health, such as behaviours like tobacco smoking that have a direct effect on incidence of disease and mortality, to intermediate determinants of health, such as availability of essential medicines, to distal determinants of population health, such as wealth inequality in a population. Of the universe of health determinants, these reporting guidelines were developed for estimates of health behaviours and health exposures. They were not designed for health systems indicators, such as those related to health financing, essential medicines, or health workforce, nor for service coverage indicators. The guidelines were also not designed for estimates of distal determinants of health, such as average educational attainment or wealth inequality.

GATHER is intended to provide guidelines for the reporting of studies that synthesize multiple data sources in order to report comparisons among populations, whether these comparisons are across geographic areas or in the form of time trends. As such, the GATHER checklist is not designed for reports of a health indicator from a single study or data source, such as a health survey or health service records for a single time period. For example, a study reporting results from one health interview survey or deaths from one vital registration system do not fall within the GATHER scope. To fall within the GATHER scope a study must use multiple data sources to draw comparisons across populations. An example of a simple study that falls within GATHER’s scope would be a study that uses the latest round of the DHS to draw comparisons between the populations those surveys represent.

Authors of studies not included in the GATHER scope may still find these reporting recommendations useful. In particular, a commitment to documenting all data inputs and analysis methods should be a universal feature of published reports providing estimates designed for policy planning.

**Box 1: Definitions of technical terms.**

**Health indicator:** A measurable quantity that may be used to describe a population’s health or its determinants. Indicators can be categorized into four domains: health status (e.g., life expectancy, HIV prevalence), risk factors (e.g., childhood stunting, prevalence of smoking), service coverage (e.g., immunization coverage rate), or health systems (e.g., hospital bed density, death registration coverage).23
Health estimates: Quantitative population-level estimates (including global, regional, national, or subnational estimates) of health indicators, including indicators of health status such as estimates of total and cause-specific mortality, incidence and prevalence of diseases, injuries, and disability and functioning; and indicators of health determinants, including health behaviors and health exposures. Examples of health indicators that fall within the scope of GATHER include life expectancy, disability-adjusted life years by cause, under five mortality rate, maternal mortality ratio, mortality rate from road traffic injuries, HIV prevalence, prevalence of stunting in children under 5 years of age, prevalence of current tobacco use, prevalence of obesity in adults, condom use among sex workers, and percentage of population using a safely managed drinking water source.

Data inputs: All numerical inputs to mathematical or statistical models that are used to generate global health estimates. Model inputs may include raw health data, processed health data, covariates, and/or other parameters, depending on the analytic approach. Raw health data are measures derived from primary data collection with no adjustments or corrections. Processed health data are health statistics that have been calculated from raw health data, but which are not the result of synthesizing multiple data sources. Examples of processing raw health data include cleaning data by removing implausible values; calculating an indicator using an algorithm; or adjusting a statistic for bias.

Covariates: Data, including non-health data, which are used in a statistical model to improve the estimation of the health indicator of interest. These data are population-specific and are available for every population included in the analysis. A common covariate is gross domestic product per capita.

Statistical model: Statistical models describe associations between variables, and use empirical data to estimate parameters.

Mathematical model: Mathematical models represent theorized causal pathways and mechanisms between exposures and health outcomes, and use estimated parameters to make predictions.17

Population health: The aggregate health of a defined group, including health status (including health outcomes, both fatal and non-fatal) and health determinants (including health exposures and health behaviors).

Cross validation: An analytic approach used to assess model performance. It is carried out by withholding subsets of data, to create the appearance of populations or time points without data. The model is then refit to the remaining data, and predictions are made for the withheld data. The predictions are then compared to the known-but-withheld data. Two metrics are often used to evaluate these predictions: root mean square error (RMSE) and average relative error (ARE).

How to use this document

The GATHER checklist is intended as a tool for authors, reviewers, journal editors, and other editorial professionals to promote best practices in reporting global health estimates. It is recommended that this E&E document be read in conjunction with the GATHER statement. For each item on the GATHER checklist, the E&E provides a description or definition of the item, and the rationale for the inclusion of the item, including relevant evidence from the literature. It also provides a published example (or examples) of good reporting for the item.

It is important to note that, given the wide variety of journals and organizations publishing global health estimates, authors have discretion in deciding where in their reporting this information belongs. The various items may be appropriate to include in the title, the abstract, the main body, or additional technical appendices. This may also be influenced by the requirements of the journal or publishing organizations.
The GATHER checklist is not intended to give guidance on study design or the methodological approach to estimation of global health indicators. Instead, it is meant to ensure that reporting is transparent and complete. This will allow informed users to assess the quality of the estimates, compare them with other estimates, and decide on their appropriate use.

In some cases, authors may not be able to comply with all GATHER guidelines due to particular circumstances. In this situation, authors should state why compliance was not possible (for example, in the case of copyright restrictions, it may not always be possible to provide access to statistical code as in item 14).
The GATHER checklist of information that should be included in new reports of global health estimates

<table>
<thead>
<tr>
<th>Item #</th>
<th>Checklist item</th>
<th>Reported on page #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objectives and funding</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Define the indicator(s), populations (including age, sex, and geographic entities), and time period(s) for which estimates were made.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>List the funding sources for the work.</td>
<td></td>
</tr>
<tr>
<td><strong>Data Inputs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Describe how the data were identified and how the data were accessed.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Specify the inclusion and exclusion criteria. Identify all ad-hoc exclusions.</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Provide information on all included data sources and their main characteristics. For each data source used, report reference information or contact name/institution, population represented, data collection method, year(s) of data collection, sex and age range, diagnostic criteria or measurement method, and sample size, as relevant.</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Identify and describe any categories of input data that have potentially important biases (e.g., based on characteristics listed in item 5).</td>
<td></td>
</tr>
<tr>
<td><strong>For data inputs that contribute to the analysis but were not synthesized as part of the study:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Describe and give sources for any other data inputs.</td>
<td></td>
</tr>
<tr>
<td><strong>For all data inputs:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Provide all data inputs in a file format from which data can be efficiently extracted (e.g., a spreadsheet rather than a PDF), including all relevant meta-data listed in item 5. For any data inputs that cannot be shared because of ethical or legal reasons, such as third-party ownership, provide a contact name or the name of the institution that retains the right to the data.</td>
<td></td>
</tr>
<tr>
<td><strong>Data analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Provide a conceptual overview of the data analysis method. A diagram may be helpful.</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Provide a detailed description of all steps of the analysis, including mathematical formulae. This description should cover, as relevant, data cleaning, data pre-processing, data adjustments and weighting of data sources, and mathematical or statistical model(s).</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Describe how candidate models were evaluated and how the final model(s) were selected.</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Provide the results of an evaluation of model performance, if done, as well as the results of any relevant sensitivity analysis.</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Describe methods for calculating uncertainty of the estimates. State which sources of uncertainty were, and were not, accounted for in the uncertainty analysis.</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>State how analytic or statistical source code used to generate estimates can be accessed.</td>
<td></td>
</tr>
<tr>
<td><strong>Results and Discussion</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Provide published estimates in a file format from which data can be efficiently extracted.</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Report a quantitative measure of the uncertainty of the estimates (e.g. uncertainty intervals).</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Interpret results in light of existing evidence. If updating a previous set of estimates, describe the reasons for changes in estimates.</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Discuss limitations of the estimates. Include a discussion of any modelling assumptions or data limitations that affect interpretation of the estimates.</td>
<td></td>
</tr>
</tbody>
</table>

This checklist should be used in conjunction with the GATHER statement and Explanation and Elaboration document, found on gather-statement.org
Objectives and funding

1 Define the indicator(s), populations (including age, sex, and geographic entities), and time period(s) for which estimates were made.

Explanation and rationale: The key components of a study’s scope should be clearly defined: what indicator or indicators are estimated, for which populations (including age, sex, and geographic region), and for which time period(s).

A variety of indicators may be appropriate for a given study, and readers should be informed as to the specific indicator, or indicators, estimated. For example, it would be insufficient to describe that a study examines maternal mortality; instead, a more specific description would identify the indicator as (for example) the maternal mortality ratio, defined as the number of maternal deaths per 100,000 live births. More detail would also include a clear definition of what constitutes a maternal death, such as the provision of applicable ICD codes. In some cases, diagnostic criteria should be reported to ensure that the specific definitions applied to the data are clear. For example, a study reporting prevalence of current tobacco smoking would need to define current tobacco smoking in terms of frequency and/or duration of tobacco smoking.

In some cases, the data, or some subset of the data, used to estimate a particular indicator may not exactly match the target indicator. For example, a study reporting estimates of current tobacco smoking may include data that provides information on daily tobacco smoking. Reporting discrepant measurement methods is discussed further under items 5 and 10.

A clear description of the age, sex, and geographic region of the population(s) for which estimates are made allows readers to determine if the study is relevant to their own population of interest. In some cases, age and sex constraints are part of the indicator definition, such as the MDG indicators for children under five, or women of reproductive age (15–49 years). The study should make these definitions explicit. Any age groups used should be specified. There may be inconsistencies between the input data and reported estimates; for example, a study may make use of any data available for women of reproductive age, but the estimates may be formally reported for women aged 15–49. Geographic regions should also be defined, for example with a reference to a country classification or a table of countries (or administrative units where subnational) in each region. Where subnational information is provided a map of the administrative sub-divisions used should be referenced. Finally, the time period for which estimates are made should be clearly defined.

Example: “We estimated 1995–2011 trends in distributions of haemoglobin concentration for children aged 6–59 months and for women of reproductive age (15–49 years), by pregnancy status, in 190 countries and territories organised into 11 regions.”

2 List the funding sources for the work.

Explanation and rationale: Given the critical and growing importance of global health estimates to decision-makers, as well as the growing field of actors in the measurement of global health, authors should be transparent in reporting any and all funding sources for a given study. Trust in scientific publication depends on transparency, and clear identification of all funding sources is essential. In other health fields, such as economic evaluation and randomized controlled trials of pharmaceuticals, studies have shown that authors’ financial relationships have been associated with their findings.

We are not aware of any evidence for financial relationships leading to biases in global health estimates. However, in the global health field, advocates – which may be at the personal or institutional level – for a particular condition or
disease may have non-scientific incentives to arrive at a certain set of results. To help readers evaluate the potential for conflicts of interest, they should always be made aware of the studies’ funding. The International Committee of Medical Journal Editors (ICMJE) set of recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals recommends, among other things, full disclosure of the “sources of support for the work, including sponsor names along with explanations of the role of those sources if any in study design; collection, analysis, and interpretation of data; writing of the report; the decision to submit the report for publication; or a statement declaring that the supporting source had no such involvement.” Most peer-reviewed medical journals already require authors to report this information (for example, PLoS Medicine, The Lancet, BMJ, Bulletin of the WHO, JAMA), and this should be adopted as a universal component of any reporting on global health estimates, including publications and reports from governments, inter-governmental agencies, think tanks, academic institutions, non-governmental organizations, or any other type of organization.

Example: “Funding: This research was supported by a cooperative agreement from the Centers for Disease Control and Prevention (CDC) through the Association of Schools of Public Health (ASPH) (Grant No. U36/CCU300430-23). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of CDC or ASPH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.”

Example: “Finally, WHO gratefully acknowledges the financial support of the Government of Norway for the development and production of this report.”

Example: “P.W.G. is a Career Development Fellow (no. K00669X) jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and receives support from the Bill and Melinda Gates Foundation (BMGF; nos OPP1068048, OPP1106023). These grants also support E.C., S.B., B.M., U.D., D.J.W., D.B. and A.H. The Swiss TPH component was supported through the project no. OPP1032350 funded by the BMGF. D.L.S. is funded by the BMGF (OPP1110495). S.I.H. is funded by a Senior Research Fellowship from the Wellcome Trust (no. 095066), which also supports K.E.B., and grants from the BMGF (nos. OPP1119467, OPP1106023 and OPP1093011). S.I.H. and D.L.S. also acknowledge funding support from the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. J.T.G. is funded by an MRC Fellowship (no. G1002284). E.A.W. and P.A.E. are funded by the Global Good Fund.”

Data inputs

Estimation strategies for global health estimates often rely on a variety of data inputs. These inputs can be organized into two broad categories: first, input data that are synthesized by the researchers for the purpose of the study, and second, input data that are taken from existing sources and used as reported. The first category would include, for example, survey data that are compiled and combined using a statistical model to arrive at an estimate of diabetes prevalence, or vital registration data from several countries that are adjusted for underreporting and then combined to calculate regional mortality rates. It would also include datasets that are combined to generate a new regression covariate. The second category would include, for example, World Bank estimates of GDP used in a regression as a covariate, or UN Population Division estimates of population as the denominator in a reported metric. These two data types have different reporting requirements, and the checklist separates them for clarity.

Data inputs that are synthesized in the study
3 Describe how the data were identified and how the data were accessed.

**Explanation:** In the medical field, a systematic data search usually includes a review of the published scientific literature. However, in the global health field, a literature review may be only a small component of the data identification strategy, or not used at all. Health data are often found in the grey literature, including health surveys, government reports, and databases maintained by international organizations and academic groups.\(^3\)\(^,\)\(^1\)\(^2\) Because of these disparate sources for global health data, all the methods used to identify relevant data should be clearly specified. This information will allow readers to assess the comprehensiveness and completeness of the data search employed for reported estimates. Researchers should specify all databases or websites searched, with search terms if relevant. In the event that published literature was reviewed, as in a meta-analysis or systematic review, the PRISMA guidelines should be followed for reporting the search procedure.\(^1\)\(^8\) In those guidelines, item 8 requires “Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.” Authors should report any additional search strategies, such as the use of professional networks for unpublished data or contacting authors for additional information. Authors may also wish to note any missing data that were sought but not successfully accessed or used. Most data searches are limited in some way, whether by resources, database accessibility, language limitations, or some other constraint. If relevant, authors should indicate these constraints.

This comprehensive reporting of data identification will help the reader understand the universe of potential data sources considered in the study. In addition to reporting the data identification strategy, authors should report how the identified data were accessed. In some cases, microdata may be downloaded or otherwise accessed for re-analysis. In other cases, hard copies of reports or papers may be obtained and the data manually extracted. Some survey programs may make available pre-analysed indicators through tools, such as StatCompiler on the Demographic and Health Surveys website.\(^3\)^\(^6\) Authors should report the general access strategies for the data used, for example by stating that individual-level data was obtained if possible, and if not, estimates were extracted from reports. This allows readers to understand the initial form of data inputs, and is also important for replication.

Authors may also wish to report their data extraction and data management systems. Depending on the project, data extraction from published reports may be a detailed process that involves forms, duplicate extraction, and other formal procedures common in systematic reviews. Where applicable, authors should provide sufficient detail that readers can understand the full process. In such cases, authors may find further guidance from the PRISMA guidelines, particularly item 10, to be useful in their reporting.\(^3\)^\(^7\) Data management systems refer to the software used to manage the data, and this may be a spreadsheet, statistical software, or specialised data management tool.

**Example:**

“We obtained anonymised individual-level data from health-examination surveys and household surveys with haemoglobin measurements. Most of these sources were multi-country surveys including the Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), Reproductive Health Surveys (RHS) and the Malaria Indicator Surveys (MIS). We also searched websites that archive survey data for health examination surveys that met our inclusion criteria; the websites included the Inter-University Consortium for Political and Social Research (ICPSR), the Institute for Health Metrics and Evaluation’s Global Health Data Exchange (GHDx), and the RAND corporation. Finally, we used health examination survey data available to the authors through the Global Burden of Metabolic Risk Factors of Chronic Diseases Collaborating Group. […]”

WHO maintains a database on anaemia prevalence. Data are identified via periodic MEDLINE searches and an international network of collaborators, who uncover data sources not reported in routine databases. The search is limited to humans, and the following search terms are used:
Studies are included in the WHO database if there is a defined population-based sampling frame; a probabilistic sampling procedure is used; and sample size is at least 100 individuals. We accessed [...] these summary data [...] We also provided a list of data sources in the WHO anaemia database to WHO country offices and to WHO’s network of international collaborators, and requested that additional data to be provided. We extracted additional data from reports obtained through this data request, from preliminary DHS reports for surveys that were not yet available as individual-level records, and from reports archived on the Malaria Indicator Surveys.”

Source: 24

4 Specify the inclusion and exclusion criteria. Identify all ad-hoc exclusions.

Explanation: Consistent application of defined inclusion/exclusion criteria is a key characteristic of a systematic review.37 Researchers may apply a range of inclusion and exclusion criteria to data sources. Information on these criteria helps the reader understand the completeness of data used, flags potentially informative data that were excluded, and indicates when potentially lower-quality data were used. The authors should specify any time frames, languages, study designs, or demographic groupings that restricted the data or publications included, as relevant.

The quality of data available for global health estimates varies widely.6,11,38,39 In some cases, researchers may apply a statistical rule (i.e. a data source falls more than 2 SD from the model’s mean) to identify outlying datasets that should be excluded or treated differently40; any such statistical rules should be identified. In other cases, researchers rely on expert opinion to identify data sources or estimates from data sources that appear implausible or that are believed to be unreliable.10,41,42 If any data sources that met inclusion criteria were eventually excluded as a result of expert review, such ad-hoc exclusions should be explicitly identified. This will allow other researchers to replicate the study if desired. If there were “rules of thumb” or particular red flags that the researchers relied on for their ad-hoc exclusions, those should be clarified where possible.

Example:

“Eligible data were WHO VR data or data derived from MEASURE demographic and health surveys (DHS) (appendix). We reviewed the latest publicly available WHO VR data from the years 2006–10, except for Canada, for which the data from Statistics Canada were used because they were more recent. [...] We used the VR data to generate national risk estimates if the country had VR coverage of adult mortality of at least 80%,10 and the data on neonatal deaths were categorised in the time periods day 0, days 1–6, and days 7–27 (appendix), which is the most detailed breakdown of data that WHO provides. For countries with more than 50 neonatal deaths recorded in the latest year with available data, we used those data. For countries with fewer than 50 neonatal deaths in the latest year with available data, we combined deaths from the previous 2–5 years until the total number of neonatal deaths was at least 50, to avoid instability because of small numbers. We took reports of 20% or less of deaths occurring on day 0 or no deaths occurring on days 1–6 to indicate poor data quality and excluded data from these countries (appendix).” 43

Example (ad-hoc):
“In some cases, entire surveys appear to have been adversely affected by non-sampling errors, and all the observations from such surveys are consequently excluded from the fitting process. As a rule of thumb, any survey in which all the observations lie below (or in rare cases above) all the observations from other surveys for corresponding time points is excluded from analysis. For a very small number of countries, the UN IGME partners consider the estimates from several surveys to lack face validity given what is known about the country, in which case data from those surveys are also excluded. Such countries have generally experienced a lengthy period of civil unrest or war, with adverse effects both on child mortality and on high-quality data collection.”

<table>
<thead>
<tr>
<th>Country Name</th>
<th>USMR Series</th>
<th>USMR Included</th>
<th>IMR Series</th>
<th>IMR Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Afghanistan Mortality Survey (AMS) 2010, Direct</td>
<td>0</td>
<td>Afghanistan Mortality Survey (AMS) 2010, Single year</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Demographic Health Survey 2007, Indirect</td>
<td>1</td>
<td>Demographic Health Survey 2007, Indirect</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Multiple Indicator Cluster Survey 2008, Indirect</td>
<td>1</td>
<td>Multiple Indicator Cluster Survey 2008, Indirect</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>WHO Vital Registration Data 2011 version, VDR (Single year)</td>
<td>1</td>
<td>WHO Vital Registration Data 2011 version, VDR (Single year)</td>
<td>1</td>
</tr>
<tr>
<td>Albania</td>
<td>Demographic and Health Survey 2009, Indirect</td>
<td>1</td>
<td>Demographic and Health Survey 2009, Indirect</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Reproductive Health Survey 2012, Indirect</td>
<td>1</td>
<td>Reproductive Health Survey 2012, Indirect</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Multiple Indicator Cluster Survey 2000, Indirect</td>
<td>1</td>
<td>Multiple Indicator Cluster Survey 2000, Indirect</td>
<td>1</td>
</tr>
</tbody>
</table>

Source: 44, including Table S1 truncated.

Example (ad-hoc):

“Sixth, we excluded outliers based on four criteria. (1) Studies with biologically implausible values, such as 100% of mortality from a single rare cause. (2) Studies with results that were greatly inconsistent with other studies for the same country. (3) Studies that were greatly inconsistent with studies from other countries with similar sociodemographic profiles within the same region. (4) Studies that, if included, led to abrupt changes in model-estimated time trends that could not be explained by contextual changes or policy initiatives. Outliers (0-89% of database entries) are shown in the online data visualisation of the cause of death database.” 42

5 Provide information on all included data sources and their main characteristics. For each data source used, report reference information or contact name/institution, population represented, data collection method, year(s) of data collection, sex and age range, diagnostic criteria or measurement method, and sample size, as relevant.

Explanation: Provision of information about the key characteristics of synthesized data sources is essential. Compliance with this item allows the reader to understand the input data upon which estimates are based, as well as to make an informed assessment of the completeness of the data amassed for the study. 22 It allows the user to assess how similar the data sources are, to easily identify gaps and limitations in the data included, and to make some limited assessments about data quality. For example, in a multi-country study reporting trends in cardiovascular disease, some countries may have multiple, overlapping surveys with very large sample sizes that use standard measurements, while others may rely on a very small number of data points of using diverse diagnostic criteria or representing specific subpopulations. 45 This information is critical in the interpretation of the estimates and should be readily available to the reader. In particular, in the case that there are major known sources of bias that affect multiple data sources, and for which analytic methods were developed, those affected data sources should be identified as well. See item 6 for more information on this. Explicit identification of all synthesized input data is also important for study replication.

An ideal report would include, for each data source, reference information or contact name/institution, sampling strategy, population represented, year of data collection, sex and age range, diagnostic criteria or measurement method, and sample size. Further elaboration follows.
Clearly identifying the reference information or the contact name/institution for every data source is essential basic information. Without this information, readers cannot assess the input data sources, nor is replication possible. In some rare cases, it may not be possible to identify a particular reference out of genuine concern for the safety of those who collected or maintain the data. In this situation, authors should provide as much information as safely possible.

The ideal report will allow readers to identify how the population represented in the data sources compares to the population for which an indicator is reported. A global health indicator for a given population may be estimated using data from a range of data sources that are more or less representative of the target population for which estimates are made. It is therefore important that the population represented by each data source is identified. Depending on the inclusion and exclusion criteria for the study, the categories for this item may be, for example, national or sub-national; urban, rural or mixed; general population, clinic attendees, or employment-based (e.g. farmers); or some combination (e.g. subnational urban). Depending on the report, there may be other categories that are appropriate. Authors should endeavor to use categories that identify data on any subpopulations that differ from the target population.

A brief description of the data collection method for each data source provides the reader with information on how subjects were selected from the population represented. This characteristic would identify broad categories of data collection methods, such as survey (typically with random selection), administrative data source (typically with passive data collection, a type of convenience sample), or census. It would not require (nor preclude) inclusion of detailed information on sample design.

The year(s) of data collection for each data source should be identified. In some cases, the data may be collected in one year, but actually represent a different time period. For example, a sample survey may collect birth history information from women of reproductive age. Depending on how child mortality is calculated, the time period for which estimates are produced will be different from the year in which the data is collected (such as covering the five years prior to the survey date). In this case, it may be useful for the authors to report this five-year period in addition to the year the survey was conducted. In the case of some medical or vital record reviews, the year the study was conducted may not be relevant, and the year to which the health measurements pertain should be reported instead.

The sex and age range of the data used should be identified. This is particularly relevant when an indicator is defined with a particular age range, but the input data is only available for a different range. For example, a study of maternal mortality may make use of data sources that differentially define “reproductive age” (i.e., 15-49, 13-49, or 15-45). This characteristic would allow readers to identify where the age ranges were inconsistent with the target indicator. In cases where authors sought data disaggregated by age and sex, authors may also report whether data were available by age and sex.

A study may include input data that identify cases using multiple diagnostic criteria or measurement methods. These should be identified and reported if they affect the accuracy or precision of the input data, and should always be reported in cases where analytic methods are developed to adjust for different measurement methods (see item 6). For example, some studies may use data inputs that are based on both self-reported status and data inputs that are based on a clinical diagnosis or physical measurement. Sometimes input data based on clinical diagnoses use different diagnostic criteria to identify cases, for example, anaemia cases are identified using different thresholds of blood haemoglobin concentration.

Providing the sample size for each data source allows for a basic assessment of sampling error. While there are many components to sampling error, sample size more consistently available than other measures of sampling error. In the event that a data source uses a complex analysis method or a complex survey design, the sample size may not adequately represent the sampling error, and it may be useful to report the effective sample size (ESS, the sample size adjusted for design effects), standard errors or confidence intervals if these are consistently available. In cases where age- and sex-specific data inputs were used, authors may wish to report these measures for each age/sex group.
In addition to the characteristics discussed above, for some studies, it may be appropriate to include additional information, such as case detection method (active vs. passive),\textsuperscript{53} point vs. period prevalence,\textsuperscript{54} or response rate.

In some cases, a report may include multiple indicators. In this situation, the data sources and their main characteristics should be identified for each indicator. This may require multiple tables or additional columns; the presentation is up to the researchers’ discretion, but the information should be available for each indicator estimated.

To avoid redundancy, it would not be necessary to include characteristics that form part of the inclusion or exclusion criteria and are therefore the same for all data sources. For example, if a study of depression only included survey data that used the WHO Composite International Diagnostic Interview (CIDI) scale\textsuperscript{55}, this information would not be necessary to include in the table, as it would self-evident from other sections of the report.

Example:

<table>
<thead>
<tr>
<th>Country</th>
<th>Coverage</th>
<th>Years (NS = not specified)</th>
<th>Age range</th>
<th>Total examined</th>
<th>Urban / rural</th>
<th>Cause data used</th>
<th>Presenting or best-corrected</th>
<th>Visual acuity levels used</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia local</td>
<td>1992-1994</td>
<td>49-99</td>
<td>urban</td>
<td>glaucoma, cataracts, macular degeneration, diabetic retinopathy</td>
<td>Both</td>
<td>blind, MSVI</td>
<td>BLUE Mountains Sydney Microdata</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Table truncated. Note that in this example, all input data were surveys with random selection of the population, and therefore the data collection method is not reported for each data source.

In some cases, the list of data sources is lengthy. In addition to the table recommended above, a graphical representation of data availability can illuminate whether there are geographical areas or time periods where data are particularly sparse. This is desirable in situations where there are a large number of sources as it highlights limitations in data availability.
Other presentations of data availability may be useful and are encouraged.

**Example:** “We computed an index of the geographical and temporal representativeness of the data sources available for non-fatal health outcomes for each cause or impairment—the data representativeness index (DRI). The overall DRI simply counts the fraction of countries that have any incidence, prevalence, remission, or excess mortality data available for causes that are prevalent in that country. We did not count cause of death data in this measure, even if it was used in the estimation of incidence or prevalence. We computed the same measure for three periods: before 1998, 1998–2005, and 2006 onwards.”

**6 Identify and describe any categories of input data that have potentially important biases (e.g., based on characteristics listed in item 5).**

**Explanation:** In some cases, certain types of input data may provide a biased estimate of the health indicator of interest. There are several biases in health data that should be considered when reporting; see Box 2 for some of the most frequently observed.

For example, case definitions or diagnostic criteria may vary between data sources; one source may rely on self-reported diagnosis (for example, using a question like, “have you ever been told by a health professional that you have diabetes mellitus?”), while another may take a physical measurement (such as an A1C test) and then apply a cut-off to make a diagnosis. Both of these sources are attempting to measure the same quantity, but both have a set of potential biases relative to the gold-standard definition of the reported indicator that should be noted. As another example, health indicators may vary by demographic group. It is recognized that surveillance of HIV seroprevalence in antenatal clinics is often undertaken in clinics that have higher prevalence than the national average.
Any biases in the input data that the researcher determined to be important for the interpretation of the results should be clearly identified and described. Researchers know of many of these biases when they begin working with data, and they often apply existing methods or develop new ones to help correct for them, making the input data more reliable. Any adjustments undertaken to correct for these biases should be clearly explained, as outlined in item 10 (detailed description of analytical steps). In addition, if there are known biases, and researchers did not make adjustments for them, that decision should also be explained, as outlined in item 18 (discussion of study limitations).

Box 2: Common sources of bias in input data.

Health data may provide a biased estimate of the health indicator of interest. A summary of important sources of bias follows.

**Inconsistent case definitions or diagnostic criteria:** Health data usually identify persons who test positive for a particular case definition. Case definitions may vary by data source, limiting their comparability. Assessor’s qualifications may vary, which may lead to differences in ascertained prevalence. In addition, laboratory protocols may change over time, reducing comparability even when case definitions have not changed. Changes in sensitivity or specificity of detection methods may have an important effect on case identification, as will decisions about whether to adjust for sensitivity/specificity.

**Self-report biases:** With some survey instruments, systematic biases can arise from difficulties in obtaining accurate responses from survey respondents. Examples of self-report biases include recall bias or social desirability bias. Self-reports of prior diagnosis are often an underestimate of true incidence or prevalence since some cases do not interact with the health system or are not diagnosed. These biases may vary systematically by populations and over time.

**Incomplete population-based surveillance:** Surveillance systems designed to capture all cases of a disease in a population via health service contact or notification are often incomplete. In most cases, the sub-population excluded from the system is not a random sub-group, but may be, for example, a more vulnerable population.

**Non-representative population bias:** Some data types are collected for a subset of the general population by design, e.g. when data are collected from clinic attendees or samples of volunteers, or when data pertain to urban or rural groups only. Health status and health determinants may differ systematically between these selected populations and the general population.

Example: “The sources of data described in Box 1 are often affected by known biases. Where this occurs, UN IGME recalculates or adjusts the data to allow for these biases. More information on these processes is given in other papers in this collection, but here we briefly consider three specific examples of recalculation and adjustment that were applied during the calculation of the latest child mortality estimates.

Full birth history (FBH) surveys (see Box 1) represent the largest source of data on child mortality for low- and middle-income countries, but they involve complex data collection and extensive interviewer training. Samples therefore tend to be quite small, typically 5,000–20,000 households. With a carefully designed sample, such numbers are adequate to produce national estimates, but do not permit extensive disaggregation. A paper in this collection by Pedersen and Liu shows that survey reports typically give estimates of child mortality measures for five-year periods before the survey, but UN IGME recently recalculated child mortality estimates from such data from the Demographic and Health Surveys (DHS) program for periods defined in calendar years [4]. On the basis of sampling precision, estimates for single calendar years can often be used for periods shortly before the survey, with the interval width gradually increasing further in the past. The criterion used to determine the interval width is the coefficient of variation (a measure of sampling uncertainty) of the USMR estimates, specifically, adjusting period length to keep the coefficient of variation below 10% [4]. Importantly, this variable period approach addresses “birth...
transference,” an issue often encountered in FBH surveys whereby births early in a recent window for which extra information is collected are shifted backwards out of the window [5]. Child mortality estimates from FBHs are typically calculated for periods up to 25 years before the survey date, even though such estimates are increasingly affected by selection bias because earlier recorded births increasingly represent those to mothers young at the time of the birth who survived to the survey.

In another example, UN IGME adjusts the data used to estimate child mortality in eastern European countries because of concerns about the low levels of early neonatal death recorded in the civil registration systems of these countries compared to western European countries as a result of differences in the definition of live births [6]. In a regression analysis [7] of the ratio of early neonatal (under seven days) deaths to total neonatal deaths, UN IGME found that this ratio was significantly below the western European average of 0.77 for ten eastern European countries (Belarus, Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Romania, Russian Federation, and Slovakia) and for Spain and Greece, where the ratio is less than 0.70. Overall, eastern European countries, Spain, and Greece fall into two strata based on the adjustment to the USMR required to match their average ratio of early to total neonatal deaths to the western European average. Consequently, UN IGME applies a 10% upward adjustment to the USMR for Belarus, Hungary, and Lithuania, and a 20% upward adjustment for the other countries, including Spain and Greece. Importantly, Estonia adopted the WHO definition of a live birth in 1992, and the civil registration data from Estonia show no sign of incompleteness after this adoption, so the adjustment of 20% is made to data from Estonia only prior to 1992. The remaining countries show no time trends in the incompleteness of early neonatal death data.

Finally, UN IGME also recalculates FBH child mortality estimates for countries with generalized HIV epidemics (defined for the purposes of CME as an HIV prevalence at any point in time since 1980 exceeding 5%). The assumption underlying the estimation of child mortality (whether directly or indirectly) from reports of mothers is that there is no correlation between the mortality risks of mothers and their children. HIV clearly breaks this assumption: both mothers and their children have higher risk, leading to a downward selection bias in child mortality estimates. As described by Walker and colleagues in a paper in this collection [8], in the latest UN IGME child mortality estimates, this bias is estimated by projecting both children and mothers forward from birth to a given survey date, separating mothers into HIV-negative and HIV-positive streams, and births into HIV-negative children born to HIV-negative mothers, HIV-negative children born to HIV-positive mothers, and HIV-positive children born to HIV-positive mothers; bias is estimated on the basis of child deaths that would not be reported because the mother has died.”

Source: [44]

**Data inputs that were used without modification**

In addition to the synthesized data considered in items 3-6, many studies will use other types of data without modification. Item 7 is relevant for these types of fixed quantities used as data inputs.

7 *Describe and give sources for any other data inputs.*

**Explanation:** Global health indicators are often calculated using the results of published meta-analyses or published sets of estimates. For example, health indicators are frequently reported as rates or ratios, and so population estimates, or another demographic quantity (such as live births), are used as denominators. There may be various sources available for these quantities, and there may be important differences between them. [62,63] Another example of this category of input data may include covariates for regression models (such as level of education, GDP per capita, or HIV prevalence). [61,64-66] All of these quantities are the result of another estimation procedure, which may
not be documented as recommended by GATHER. Model parameters in a natural history model are another example of this type of data. For each data input of this type, the report should specify the source used. The report may also describe important known limitations of that source if the authors wish to include this information.

**Example (covariates):** “We used a time-varying covariate to inform our estimates, namely, maternal education (average years of schooling for women of reproductive age) [22].”

**All data inputs**

**8 Provide all data inputs in a file format from which data can be efficiently extracted (e.g., a spreadsheet as opposed to a PDF), including all relevant meta-data listed in item 5. For any data inputs that cannot be shared due to ethical or legal reasons, such as third-party ownership, provide a contact name or the name of the institution that retains the right to the data.**

**Explanation:** Reproducibility is a key component of the scientific method. It increases trust in estimates and also promotes improvement in methods. In order to facilitate replication of results, researchers need access to the input data. Therefore, all data inputs should be made available in electronic, machine-readable format such as a spreadsheet, database or CSV file upon publication of the estimates. The format should allow for easy import and use, and not require difficult extraction procedures or data re-entry, as these steps may result in errors and hamper replication efforts. Electronic, machine-readable form does not include formats like electronic PDF files, Word documents, or visualization tools, which require manual labour to extract the data into a useable format. Input data may be shared as an appendix to the journal article, on the researchers’ website, or in an external archiving system such as the Center for Open Science (http://centerforopencience.org/).

In some cases, researchers cannot share input data because they do not have the right to publish input data or because of confidentiality concerns. In these cases, a contact name or the institution that retains the rights to the data should be provided. For example, a study’s data inputs may include individual-level data from the Demographic and Health Surveys program. Users must register with DHS to access the microdata from these surveys, and under the terms of use, they are not permitted to republish these data (nor would it necessarily be desirable for secondary users to republish it). Instead, the authors should direct readers to the DHS website and clearly state that the data is available from that institution.

Data owners have increasingly developed workarounds for situations in which data cannot be made freely available, such as remote secure access or allowing on-site access for reanalysis purposes (i.e.). These approaches may allow for replication of the study with the full dataset by other interested researchers. However, these workarounds may have substantial associated costs and are rarely provided for by research funders.

In the case where some data are restricted access, the researchers may choose to include a comparison of the results between the full set of data and the openly available data. This allows other analysts to understand the importance of the restricted datasets to the results, even if they are not able to access all the data themselves.

In some studies, key intermediate datasets are calculated as part of the study, which are used as inputs to subsequent steps of the analysis. For example, several sources of data may be combined to generate a covariate dataset. In these cases, best practices would be to also make the intermediate dataset available.

It is important to note that some donors and journals are increasingly calling for and requiring public access to data. For example, The Bill and Melinda Gates Foundation updated their Open Access policy as of January 1st, 2015, and specifically require that for any research funded by that foundation, “data underlying published research results will be accessible and open immediately.” In the *PloS Medicine* submission requirements, they require that,
“All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository, unless already provided as part of the submitted articles."78,79

Example: “The empirical dataset (original and with USMRs and NMRs rescaled) can be found on the following Web site (www.who.int/healthinfo/statistics/mortality/en/).”

Source:80

Data analysis

9 Provide a conceptual overview of the data analysis method. A diagram may be helpful.

Explanation: Global health indicators are intended for a wide audience, including both expert and non-specialist users. In some cases, the details of the analysis may be very complicated and use advanced statistical methods that are not accessible to many readers.8 However, regardless of the level of complexity, an ideal report would allow all users, including non-specialists, to understand, conceptually, the basic analytic steps and key model assumptions. This requires the use of plain language and the avoidance of jargon to bridge the “gulf of misunderstanding” between data users and data producers.7

An ideal conceptual overview would allow readers to use the description provided to contrast with other methodological approaches, as well as consider limitations in the methodological approach presented. In particular, there may be assumptions embedded in the analytic approach which are important for interpreting and using results, and these should be made clear to all users. For example, if a study projects a health indicator based on a linearity assumption, this should be made clear.81 Some assumptions may be particularly relevant for out-of-sample estimates. For example, in many studies, estimates are reported for countries (or another geographical region) for which there are no available data. The conceptual overview should provide information on which data inform these out-of-sample estimates. The importance of highlighting these analytic assumptions is discussed in more detail in item 18.

Studies often involve a series of analytic steps, including, for example, data identification and collection, application of exclusion criteria, data adjustments, statistical modelling, and production of final estimates. These steps may be very complicated to describe in the text, particularly in an accessible way. In some cases, a diagram can lend clarity and improve understanding, particularly for non-specialist users.

It may be useful for researchers to have a non-specialist colleague review the manuscript before submission or publication to assess whether the conceptual overview is accessible to a broader audience. If available, communications specialists are another source of useful advice on making text broadly accessible.

Example: “We used a Bayesian hierarchical mixture model to estimate distributions of height-for-age and weight-for-age Z scores by rural and urban place of residence for all 141 countries for every year from 1985 to 2011. The statistical model is an extension of the model used and described in detail previously.5 […] In the hierarchical model, estimates for each country-year were informed by data from that country-year itself, if available, and by data from other years in the same country and in other countries, especially those in the same region with data for similar time periods. The hierarchical model shares information to a greater extent when data are non-existent or weakly informative (eg, because they have a small sample size), and to a lesser extent in data-rich countries and regions.”82
Example:

![Diagram of analysis strategy]

**Figure 1: Analysis strategy**
VR=vital registration. DHS=demographic and health surveys. IGME=Inter-agency group for Child Mortality Estimation.

Source: 43

---

10 **Provide a detailed description of all steps of the analysis, including mathematical formulae. This description should cover, as relevant, data cleaning, data pre-processing, data adjustments and weighting of data sources, and mathematical or statistical model(s).**

**Explanation:** A detailed, step-by-step analytical description should be included, in addition to an accessible overview as described in Item 9. Most analyses are composed of numerous steps, each of which is essential to the final, reported estimates. Researchers should be thorough in describing all analytic steps. For example, in some cases, the first stage of an analysis often involves cleaning, processing, and adjusting the input data. After this stage, the analyst may assign weights, such as quality weights, to the input data. This may be followed by the application of a statistical or mathematical model. Often, regression modelling is used at some stage in the estimation of global health indicators, and any estimated coefficients should be reported even if they are not central to the reported conclusions. Other model parameters should also be reported as appropriate. Analysis undertaken to generate secondary model inputs, such as covariates or parameters in a natural history model, should also be fully described.
Not all analyses will have all of these steps. For example, microdata is not always used; input estimates may be extracted from published papers or other sources, and therefore data cleaning and processing may not have been undertaken. It is incumbent on the authors to identify the analytical steps and describe them in detail. This makes it possible for technical readers to have a thorough understanding of all components of the analysis. Ideally, there would be sufficient detail that another informed analyst could replicate the analysis; for this to be possible, each of these stages must be thoroughly described, as each is dependent on the previous.

Data cleaning. Data cleaning is the first stage of most analyses that use individual-level measurements, as these data often include implausible values due to data entry errors, equipment malfunctions, or some other cause. Even for well-established and often used data sources, there may be cleaning that researchers undertake before proceeding with their analysis. This may include, for example, the identification of implausible or duplicate observations or other errors. Any criteria or implausibility limits applied to the raw, individual-level data by the researchers should be reported. For example, in a study on child mortality, birth history survey data may be examined for plausible reported ages. If a mother self-reports her own age as 20 but her eldest child as 15, researchers may elect to drop that observation from the analysis data set as unreliable. In another example, in a study of obesity, researchers may elect to apply an implausibility limit on reported BMI measurements by recoding reported extreme values. Where implausibility limits require justification or evidence, this should be provided. This kind of screening for data errors, and how any errors were handled, should be noted in the report.

Data pre-processing. After initial data cleaning, each input data set is processed to produce the indicator of interest. In many cases, this can be the most time-consuming portion of the research study. Pre-processing procedures may vary depending on the type of data source, and, depending on the study, there may be several different procedures applied to different input data sets or data types. These should all be described with sufficient detail for replication. This may include information on the methodological approach for estimating the population-level indicator of interest from individual-level data, how standard errors were calculated, how complex survey design was incorporated into the analysis, whether and how sample weights were applied, or how missing observations were handled. For example, consider a study of child mortality. For this indicator, a variety of types of data may be included, such as survey data (with either direct or indirect birth histories) and vital registration data. Each of these data types should be addressed separately in the pre-processing reporting. For example, for the survey data, the ideal report would include detail on how complex survey design was incorporated into the analysis, and the methodological approach used to calculate the child mortality rate, for both direct and indirect birth histories (as appropriate). In the case that an existing, standard methodology was used (for example the synthetic cohort life table approach for direct birth histories), it would be appropriate to make reference to the method, along with sufficient information about any adjustments, assumptions, or other decisions required for replication. In the event that a new methodological approach was developed, this should be thoroughly described. For the vital registration data in this example, any redistribution of deaths with ill-defined causes or adjustments for completeness of death registration made to the data should be reported. Essentially, for each data source, sufficient information should be provided so that another analyst could process the data and arrive at the same estimate of the indicator(s) from that data source.

Data adjustments and weighting of data sources. After the data cleaning and pre-processing described above, in many analyses, the indicators estimated from the input data sources are adjusted before being included in the next stage of the analysis. This may be carried out as part of the pre-processing or as part of the mathematical or statistical model, but conceptually is usually a separate procedure that should be addressed in the report. Input data may need to be adjusted because of differences in definitions of an indicator across data sources, different availability of age groupings, or due to coverage of a particular data source. For example, a study of diabetes mellitus prevalence may include some data sources that identify diabetes using HbA1c measurements and others that measure fasting plasma glucose; researchers should report any methods used for cross-walking between these two measurements. As another example, in some cases, reported age groups are too broad or do not match the target
age groups in a study, and so researchers will undertake some method for age splitting. These adjustments are made to improve consistency and comparability between the included data sources, so that they can be combined in the next stages of the analysis. In addition to these adjustments, in some studies, analyst-chosen quality weights are applied to input data to identify high or low quality sources (or more or less representative sources). Such weights should be explicitly identified and justification for their values provided.

**Mathematical or statistical models.** After cleaning, pre-processing, and any adjustments or weights are applied to the input data, the next stage of an analysis is usually the development and application of statistical or mathematical models to synthesize the input data and estimate the health indicator of interest. Items 11 and 12 discuss the development and evaluation of the performance of these models in more detail. Modelling in global health has become increasingly complex, with researchers drawing from diverse technical fields to develop more flexible and, hopefully, more appropriate models for the indicators under study. As a result, it is increasingly important for researchers to carefully report their statistical or mathematical modelling methods and assumptions, with sufficient detail that an informed analyst could replicate the modelling procedures. Any mathematical or statistical formulae should be reported. This allows other researchers to examine the mathematical or statistical assumptions used in the analysis, as well as replicate the precise structure of the model. Any parameter assumptions or estimated coefficients should be reported, even if they are not central to the reported conclusions.

Researchers will prefer different statistical or mathematical software packages for their analyses. In some cases, an analysis may be completed with multiple packages, perhaps because different analysts contributed to the report or because different packages were considered best suited for different tasks. Researchers should name the software packages, and versions of those packages, that are used in the analysis, i.e. These different packages may use slightly different assumptions or estimation procedures to implement particular models. In addition, in some cases, an analysis may make use of user-written, non-standard “add-on” packages, and these should also be identified and made available to allow other researchers to make use of them (see item 14).

This item identifies a lengthy list of reporting components on the analytical approach taken in a study. Not all components will be relevant to every analysis, and there may be additional components in some analyses that authors wish to describe. In many cases, much of the technical detail outlined in this item may belong in a technical appendix, rather than the main body of the report or paper. Nevertheless, comprehensive reporting on all methods used is necessary to ensure appropriate documentation of the data analysis.
Example:


\[
\begin{align*}
\text{Equation 1:} \\
&= \sum_{t=0}^{27} \frac{a}{(by)^{t+1}} \left(1 \leq t \leq 27 \right) \\
&= \text{probability of dying on day } t \text{ conditional on surviving until that day,}
\end{align*}
\]

where \( h \) is the probability of dying on day \( t \) conditional on survival until that day. The unconditional probability of dying on day \( t \) of the neonatal period, \( p \), can be derived from the multinomial distribution (appendix). The likelihood of observing \( n_0, ..., n_0 \) deaths in the neonatal period conditional on \( N \) livebirths and the proportion surviving the neonatal period, \( p \), can be expressed as:


\[
\text{Equation 2:} \\
\begin{align*}
&= p_1^{n_1} \times p_2^{n_2} \times \ldots \times p_N^{n_N} \times p_{N+1}^{X} = p_1^{N-1}N \times \prod_{i=2}^{N} p_i^{n_i} \\
\end{align*}
\]

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


\[
\text{Equation 3:} \\
\begin{align*}
&= \frac{(p_n + p_1)^{n_0}}{\prod_{i=2}^{N} p_i^{n_i}} \times \prod_{i=2}^{N} p_i^{n_i} \\
\end{align*}
\]

We used maximum likelihood to estimate the parameters \( a, \beta, \) and \( \gamma \) (Appendix). This model allowed 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. With use of these estimates, we calculated the expected proportion of neonatal deaths on a given day (Appendix) and during the time periods days 1-6, days 7-27, and weeks 1 (days 0-6). We initially applied the model to the aggregated DHS data, followed by fitting the model to subsets of the data (neonatal mortality rates, national income category and geographic region [appendix], and survey period) to investigate whether these affected the proportional distribution of deaths.

We compared our postulated model with a simpler two-parameter model that assumes \( \gamma = 1 \), by use of a likelihood ratio test (appendix). To enable us to correct for misreporting between days 0 and 1, we fitted a model in which the relation between day 1 deaths and those on subsequent days was constrained.

Source: 43

Example: “For surveys and censuses in which only summary birth histories were obtained, we calculated sex-specific estimates of USMR using the Brass method,28 either from microdata or from published tabulations of number of children ever born and children living (we chose the USMR because it is more robust to the choice of model life table than the IMR, which can vary substantially according to the model selected). When microdata or tabulations were not available, we used estimates from published surveys or census reports.” Source: 97

11 Describe how candidate models were evaluated and how the final model(s) were selected.

Explanation: The process of model building can be complex and time-consuming. A broad range of model forms may be appropriate for a given outcome. Distributional assumptions will narrow the field of appropriate models, but there is usually a wide array of candidate models that could feasibly be chosen.98,99 Models may vary based on distributional form, covariates, weighting schemes, flexibility, level of borrowing strength, and other parameters. Despite the growing complexity of models in global health, the selection of model form receives scant attention in the global health literature.
Model selection depends on the question being asked and the nature of the data to which it is being applied. This implies that there is no standard metric for the evaluation of model performance. For example, a study may seek to estimate whether countries will achieve particular indicator targets in the future, which requires the model to be particularly good at out-of-time predictions. In another case, a study may be based on a dataset in which many countries have missing data; this requires the model to handle out-of-sample predictions well.

Depending on the nature of the study question, researchers should report how candidate models were evaluated, and how the final model(s) were selected. In some cases, dozens of models may have been considered, with iterative changes between them. It is not necessary to provide the statistical equation for every possible model considered, but readers should understand from the report the types of models considered, the covariates considered, and how the analyst selected among different models. In some cases, limitations to computing power may also be a factor in model selection.

Borrowing from other disciplines, in some analyses, an ensemble modelling approach is taken, wherein no single model is chosen. This may be the preferred approach particularly where there is no clear single “winner” amongst candidate models in the evaluation stage. In this case, the universe of candidate models should be outlined, and those included in the final combinatorial model should be identified.

**Example:** “Following the principles laid out by Hogan et al, we chose the best model for maternal mortality estimation using strict out-of-sample predictive validity tests. To capture the rich debate on model specification that followed our publication in 2010, we substantially expanded the set of models that we evaluate. We have explicitly attempted to capture the type of mixed effect linear models used by Wilmoth et al in this assessment. We have further drawn on advances in the literature in other fields such as meteorology, the Netflix Challenge, and verbal autopsy to expand the pool of models that we assess. Model assessment follows in three stages.

Based on the published literature, we first identified a range of plausible covariates for maternal mortality. These are shown in Web Table 4. We divide these covariates into three groups based on the strength of epidemiological evidence: class 1 covariates for which there is strong evidence and a biologically plausible pathway, class 2 covariates with some evidence but with a less direct causal pathway, and class 3 covariates where there is general correlation evidence for a relationship as observed in previous time-series or cross-sectional studies.

The table also shows the expected direction of the relationship for each covariate. In the first step, we run regressions for all possible combinations of category 1 covariates. We run regressions for models where the dependent variable is the rate in logarithmic scale by age and models where the dependent variable is logit cause fraction by age. Counting both types of models, we assess a possible universe of 3,840 models. All models, where the signs for all covariates in that model are in the expected direction and the coefficient is significant at the p<0.05 level, are retained. At levels two or three, category two and three covariates are added to these models using a forward stepwise technique which is not order dependent. This is achieved by starting the forward stepwise evaluation for each base model over for 7 each category two covariate. Models which are subsets of other models at levels two or three were dropped leaving 98 models for rates and 71 models for cause fractions. Since each set of covariates is run as both a simple mixed effects model and as a spatiotemporal model, this results in a total pool of 338 component models. As noted, we have repeated the entire process for models of rates and cause fractions. There has been extensive debate in the literature on maternal mortality on the advantages and disadvantages of models of rates or cause fractions. 8 We prefer to develop models of both rates and cause fractions for maternal mortality and then let the predictive validity assessment drive the extent to which we choose rates or cause fraction models.”

12 Provide the results of an evaluation of model performance, if done, as well as the results of any relevant sensitivity analysis.

**Explanation:** Readers often seek assurance that the final models chosen perform relatively well. This is particularly relevant given the increasing complexity of models in the global health field; when the mechanics of the analysis are
not evident to, or clearly understood by, the reader, transparent and objective evaluations of model performance can improve validity, trust, and ultimately, policy uptake.

In many analyses, the final model(s) are subjected to performance evaluation. If conducted, this process should be reported. Authors should report whether and how model assumptions were checked. Depending on the type of model used, a range of evaluation techniques may be appropriate, and these should all be described. For example, with linear regression, plots of residuals may be examined to check model assumptions such as homoscedasticity or linearity of the response function. If used, such graphical assessment of model assumptions should be described, and potentially made available. If quantitative measures of goodness-of-fit were calculated, these should also be reported. These may include metrics such as R-squared, AIC or BIC, root mean or median squared errors, and many others, depending on the model(s) used and what the model(s) are tasked with doing.\textsuperscript{98,99} If cross-validation was undertaken to assess how well a prediction model performed in an independent dataset, quantitative measures of fit should be reported.\textsuperscript{107,108} If k-fold cross-validation, or some variation of this approach, was conducted, the procedure should be described in detail, including the types of training and test sets used, and the metrics used for evaluating model performance.\textsuperscript{80,109} If estimates were verified by comparisons between different models, this should be reported as well. Reporting these evaluation procedures will give an indication of the thoroughness of the model-building process. Further, it will offer the reader information on how well the final model performed, hopefully in comparison to other candidate models.

All estimation processes involve assumptions, including inclusion criteria for data and the functional form of a model. Some analysts may choose to use sensitivity analysis to assess the degree to which the final values of the estimates depend on these assumptions. These sensitivity analyses are encouraged and any sensitivity analyses that are relevant for the interpretation of results should be reported.\textsuperscript{43,110,111} If the sensitivity analysis suggests that various analytical approaches produce similar estimates, this lends credibility to the estimates and strengthens the results of the report. If, on the other hand, the sensitivity analysis suggests that the estimates are highly dependent on the modelling approach or data inclusion/exclusion criteria, this encourages readers to examine carefully the analytical assumptions, and may help to inform future research. Authors may wish to explore reasons why different assumptions affect the results.
Example:

2.6 Model validation

Model performance was assessed through an out-of-sample validation. Given the retrospective nature of child mortality data and the occurrence of data in series, the training set was not constructed by running out observations forward, but based on including all available data in some year in the past (3), here 2006 was chosen. To construct the training dataset, all data that were collected in or after 2006 were removed. Fitting the model to the training dataset resulted in point estimates and uncertainty intervals that would have been constructed in 2006 based on the proposed method. To evaluate model performance, we calculated various validation measures (see Table 1) based on the left-out observations and based on the estimates obtained from the full dataset and the estimates obtained from the training dataset.

For the left-out observations, errors are defined as 
\[ e_{ij} = \hat{x}_{ij} - X_{ij}, \]
where \( \hat{x}_{ij} \) denotes the posterior median of the predictive distribution for a left-out observation \( x_{ij} \) based on the training dataset. Coverage is given by
\[ \frac{1}{n} \sum I( e_{ij} < 5\% \text{ prediction interval}) \]
which is the proportion of left-out observations contained in the lower and upper bounds of the 95% prediction intervals for the \( j \)th observation in group \( i \). The validation measures were calculated for 1,000 sets of left-out observations, where each set consisted of a random sample of one left-out observation per country. Reported results include the median of the validation measures based on the outcomes in the 1,000 sets.

"Upward" estimates, denoted by \( \hat{\delta}_{ij,t} \), for country \( i \) in year \( t \), refer to the sex ratio estimates obtained from the training dataset. The error in the estimate based on the training dataset is defined as \( c_{ij,t} = \hat{\delta}_{ij,t} - \hat{\delta}_{ij,t} \).

where \( \hat{\delta}_{ij,t} \) refers to the posterior median estimate based on the training dataset. Coverage was calculated in a similar manner as for the left-out observations, based on lower and upper bound of the 95% uncertainty interval for \( \hat{\delta}_{ij,t} \) obtained from the training dataset.

3.1 Validation results

We let out all observations that were collected in or after the year 2006. 1,085 observations were left out, corresponding to 6.6% of all observations. Table 2 summarizes the results related to the left-out observations for the validation exercise. Median errors were very close to zero for left-out observations in age groups [0,5) and [5,10]. Coverage of 95% prediction intervals was slightly higher than expected for 95% for age group (0,5) and 95% for age group (5,10).

<table>
<thead>
<tr>
<th>Age group</th>
<th>[0,5)</th>
<th>[5,10]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median error</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Median absolute error</td>
<td>0.10</td>
<td>0.16</td>
</tr>
<tr>
<td>% of left-out observations below 95% prediction interval</td>
<td>4.4</td>
<td>3.2</td>
</tr>
<tr>
<td>% of left-out observations above 95% prediction interval</td>
<td>5.6</td>
<td>2.9</td>
</tr>
<tr>
<td>Expected proportions (%)</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2: Validation results for left-out observations by age group. Error is defined as the difference between a left-out observation and the posterior median of the predictive distribution.

Table 3 shows the results for the comparison between estimates obtained based on the full dataset and estimates based on the training set. Median errors and the median absolute errors were close to zero and the proportion of upward estimates that fell outside the uncertainty interval constructed based on the training set was small.

<table>
<thead>
<tr>
<th>Age group</th>
<th>[0,5)</th>
<th>[5,10]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median error</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Median absolute error</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Below 95% uncertainty interval (%)</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Above 95% uncertainty interval (%)</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Expected proportions (%)</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3: Summary of differences in sex ratio estimates in observation years 2000 and 2005 based on training set and full dataset. Errors are defined as the difference between estimates based on the full dataset and the training set. The proportions refer to the proportion (%) of women in which the median sex ratio estimates based on the full data set fell below or above their corresponding 95% uncertainty intervals based on the training set. The results are broken down by age group and observation years.

We also verified that the global relations between total mortality and expected sex ratios and resulting county estimates were not substantially affected by omitting countries by leaving out countries with multiple items that were 0% smaller or greater than one.

Source: 112

Example for sensitivity analysis: See example under item 10 above for methods. The results were described as follows: "Our three-parameter model fitted the observed DHS data better than the two-parameter model (p<0.0001)." 43
Describe methods for calculating uncertainty of the estimates. State which sources of uncertainty were, and were not, accounted for in the uncertainty analysis.

**Explanation:** The reporting of uncertainty along with point estimates of global health indicators is increasingly common, though some studies only document point estimates. The reporting of uncertainty ranges around estimates provides readers with an understanding of the precision of the estimates, and is critical for comparability. However, uncertainty ranges require careful interpretation, and should be clearly defined by the authors. Often uncertainty ranges reflect only a subset of all possible sources of uncertainty in the estimates. For some types of uncertainty, there is no established methodology for estimating that uncertainty, and so analysts may develop their own methods, based on a set of assumptions, or choose to acknowledge this limitation of their analysis. In some cases, this means that more complex analytical processes may appear to have wider uncertainty than simpler approaches, but it may be that the former is capturing more sources of uncertainty. For comparability across sets of estimates, it is essential that readers are able to understand what is contributing to the uncertainty intervals, and, equally important, if there are any major sources of uncertainty that are excluded from the uncertainty intervals. A list of potential sources of uncertainty is given below to guide authors in their reporting. Authors should also specify the methods used for propagating all included sources of uncertainty into the final intervals.

**Uncertainty from errors in input data**

**Sampling error** is the most commonly captured source of uncertainty, but it may not be the largest or most important. Sampling error results from samples that are a subset of the population they are meant to represent, including the way in which samples are identified, as in cluster-based sampling or other complex survey designs.

**Stochastic error** in the input data refers to the difference between the number of events that occur in a population and the number of events that are implied by the true underlying rate or proportion, due to the data-generating process. If the modelled estimates capture the true underlying rate or proportion, stochastic error introduces uncertainty to the estimation process. For example, for two populations with 10,000 live births and identical neonatal mortality risks, hence identical neonatal mortality rates, a difference in the number of neonatal deaths may be observed due to random chance, introducing stochastic errors.

**Non-sampling error** refers to error in the input data that is not associated with random fluctuations due to sampling. It can also be called systematic or non-random error. This type of error is often the result of data collection or data processing procedures, and may result from problems like missing data, coverage problems or errors in the sampling frame (resulting in a sampled population that is not truly representative of the population of interest). It may also be caused by measurement errors at the time of data collection. Box 2 gives examples of measurement challenges that may result in non-sampling error. Non-sampling error is notoriously difficult to quantify.

**Model covariates** or other input data used “as is” may have associated error. This is sometimes referred to as “errors-in-variables.” This includes things like error associated with population estimates or live birth estimates, as well as covariates like education or GDP per capita. These inputs are often the result of complex modelling projects of their own, however, in many cases, these data do not have reported uncertainty, so it may not be possible to incorporate into the uncertainty estimates of the reported indicators. Further, computationally, it may be intractable to include this source of error.
Uncertainty from analytic methods

Parameter uncertainty (or model estimation uncertainty) is uncertainty related to the estimation of the model parameters.\textsuperscript{124} For analyses that include statistical modelling, there is uncertainty in associated the model parameters (such as the beta coefficients), because the relationships between variables are not certain.

Unexplained variation (or fundamental uncertainty) is variation that remains but is not explained by the model. Even if all the model parameters were known perfectly, there would still be error in the predictions, because there are other influences on the outcome not included in the model.\textsuperscript{124} Depending on the model specification, the unexplained variation may also reflect sampling error, stochastic error, and/or non-sampling error mentioned above, but separating these components from fundamental uncertainty is a methodological challenge.

Specification uncertainty arises from the choice of which model or models to use. A different functional form or set of independent variables will result in different predictions; these predictions may vary more than what would be estimated by parameter and fundamental uncertainty, particularly for out-of-sample predictions.

Uncertainty from multiple steps in the analytical process includes a range of potential additional uncertainty, depending on the complexity of the analysis. The pre-processing work outlined in item 10 may have associated uncertainty, especially with regard to the transformation of the raw data into the indicator of interest. For example, if indirect birth histories are used to calculate under-five child mortality, that modelling process will have associated uncertainty that the authors may choose to estimate and propagate through to the final intervals.\textsuperscript{125} Another example would be any cross-walking of data that happen before the main model, as in the conversion between different metrics of excess body weight in a study of obesity\textsuperscript{126} or between diagnostic methods in a diabetes study\textsuperscript{49}. Each of these steps may have specification, parameter and fundamental uncertainty.

Example: “The uncertainties of our estimates incorporated sampling error in each data source; non-sampling error of national data (eg, because of issues with sample design and measurement); additional error associated with subnational data; uncertainty due to conversion from NCHS reference to WHO standards; and uncertainty due to estimates made by country and year for which data were missing altogether, when only summary statistics (rather than individual-level data) were available, or when data were not available separately by place of residence.”\textsuperscript{82}

14 State how analytic or statistical code used to generate estimates can be accessed.

Explanation: Item 8 outlined the importance of making all data inputs available to improve transparency and replicability of global health estimates. Another key component of these goals is allowing access to analytic or statistical code.\textsuperscript{70} Public access to code ensures that results are replicable. Regardless of how thorough the description of methods is (as outlined in item 10), it is generally difficult and often impossible to exactly replicate an analysis based on such a description, as there are simply too many details to be comprehensively described in the text. In most cases, providing access to the analytic or statistical code may be the most appropriate mode for ensuring that an analysis is replicable and the methods fully transparent.

The ideal would be a freely available, fully-documented set of code (or spreadsheets, in some cases) that can be run off-the-shelf.\textsuperscript{127,128} However, in many cases, the production of such a product represents a very substantial investment of human capital. This may not be provided for by funders, and would require substantial trade-offs in terms of other uses of time.\textsuperscript{12} Such a product would not only need to be prepared and documented, but it may also require responses to queries from interested users. While this is the ideal, it may not be possible for many research projects given current funding levels.
At a minimum, undocumented code for key analytic steps should be available on request. Researchers are not responsible for providing technical assistance to potential users, as it is rare that this type of ongoing support is funded. Researchers may find a compromise between the ideal described above and the minimum requirement; it may be possible to provide well-documented code even if it cannot be made to run off-the-shelf. Some analysts may elect to make code available without waiting for requests. One online tool that may be used to share code effectively, with minimal investment, is the GitHub code hosting service.\textsuperscript{129,130} Regardless of how authors elect to share their code, it should be available upon publication of the health estimates.

Example: “Additional details of the input data and estimation methodology, including statistical codes, are online and publicly available through \textit{WHO’s Global Health Observatory} and the \textit{Maternal and Child Epidemiology Estimations website.”\textsuperscript{131}

Results and discussion

15 \textbf{Provide published estimates in a file format from which data can be efficiently extracted.}

Explanation: Publications of reported global health indicators often include numerous tables, perhaps with estimates in several time points for a large number of geographic areas. The extraction of these estimates from papers and reports can be burdensome to other researchers. To maximize the use of published estimates in other applications, the published final estimates should be made available in electronic data format upon publication of the estimates. This includes all point estimates and uncertainty reported. This allows other researchers to easily work with the estimates, and minimizes the introduction of errors. Other researchers may use published estimates for a range of purposes: to make figures or tables, to use as inputs for other modelling exercises, to draw comparisons with other estimates or data sources, and many other uses. Provision of the estimates in electronic format promotes the use of the published estimates. It should be noted that an electronic file may not be necessary if only a few estimates are reported (such as 5-10 world regions and a single time period). The final estimates should be accompanied by documentation that clearly states the indicator(s) definition and a reference for the detailed methods.

In some cases, such as the use of some Bayesian or bootstrap methods, the calculation of uncertainty may produce a set of draws or samples that may be useful to other researchers. For example, an informed analyst could use them to calculate uncertainty intervals for specific populations, or propagate uncertainty in further analyses. Best practice would be to make the draws or samples available as well (either publicly with the estimates themselves, or on request).

Example: See example for item 14.

GBD 2010 data downloads: \url{http://ghdx.healthdata.org/global-burden-disease-study-2010-gbd-2010-data-downloads}

16 \textbf{Report a quantitative measure of the uncertainty of the estimates (e.g. uncertainty intervals).}

Explanation: It is widely agreed that quantifying uncertainty around global health estimates -- typically by calculating and reporting uncertainty intervals -- is a desirable and necessary component of global health reporting.\textsuperscript{68} They provide context and an important reminder of the degree of precision of the estimates. However, despite agreement on the importance of reporting uncertainty intervals, methodology for constructing uncertainty intervals is an area of active research. This is because the calculation of uncertainty in the case of global health indicators is a complex analytical task that requires substantial statistical expertise and computing power.\textsuperscript{8}
Researchers should report a quantitative measure of uncertainty. In most cases, there are several types of uncertainty that may be propagated through the analysis to the final intervals, as outlined in item 13. Best practices are to ensure that the largest source(s) of uncertainty are quantified. It may be a useful exercise to outline the relative scale of the various types of uncertainty in order to determine which sources are the largest.

Uncertainty intervals are commonly reported for global health indicators, but there may be other ways to quantify uncertainty that would be equally appropriate, depending on the research question.

**Examples:** “In 2010, among women 20–44 y of age who were exposed to the risk of pregnancy, 1.9% (95% uncertainty interval 1.7%, 2.2%) were unable to attain a live birth (primary infertility). Out of women who had had at least one live birth and were exposed to the risk of pregnancy, 10.5% (9.5%, 11.7%) were unable to have another child (secondary infertility).”

---

**Figure 2:** The posterior probability that vitamin A deficiency decreased from 1991 to 2013. VAD = vitamin A deficiency.

Source: 132

17 Interpret results in light of existing evidence. If updating previous set of estimates, describe the reasons for changes in estimates.

**Explanation.** The interpretation of the final results is an essential component of reporting on global health indicators. Readers benefit from the authors’ interpretation of the results in light of the existing evidence, as it provides context to the results and should offer some qualitative sense of the quality of the estimates. In the field of global health estimation, data can be sparse and of widely varying quality, and methods are rapidly evolving. There are two categories that authors should consider reporting on: first, the placing of the results in the context of the input data and second, placing the results in the context of other estimates.
**Context of input data**

Final estimates, even with reported uncertainty, can give users a false sense of confidence. As discussed above, reported uncertainty intervals may not reflect all sources of uncertainty. Further, there is a common tendency for users, particularly non-specialists, to focus on central estimates rather than uncertainty intervals. Authors should consider discussing the results in the context of the input data, and there are several suggested best practices for accomplishing this listed below.

**Plot input data with estimates.** One commonly used approach is the use of plots to provide a graphical interpretation of model fit. Such a plot would compare model output to unadjusted and/or adjusted data inputs. This is particularly appealing as such a graphical representation is accessible to all users, regardless of their statistical background. It provides a simple visual for how well the model fits the available data. It also provides easy-to-interpret information on which input data sources were used, and may provide information on how they were adjusted. Such plots can be included in an appendix or in other supplementary materials. Interactive tools can be a very effective way to share this information.

**Categorize estimates by underlying data quality.** Input data for the estimation of global health indicators can vary widely; most comprehensive reports include estimates that are informed by underlying data of varying quantity and quality. For example, population A may have no data available, population B may have one low-quality data source, and population C may have several high quality data sources. A given report may produce estimates for all three populations, but readers should come away understanding the very different data inputs for the three populations. For example, for populations where out-of-sample predictions are relied upon to produce estimates (such as in population A), these estimates may be used to identify likely health priority areas where measured health data are not available, and can provide insight into probable trends in health statistics where trend data are otherwise not available. However, users must recognize that such out-of-sample predictions have serious limitations. For example, they are not a substitute for measurements when evaluating disease programs or health systems reform. Best practice would indicate health estimates based on missing or low-quality input data, in either graphical, map, or tabular form as appropriate. This could be accomplished by using a different colour or font to indicate that particular point estimates should be used and interpreted with caution. It may also be appropriate to provide a rank or score for populations based on the quality of input data available.

**Context of other estimates**

In addition to placing estimates into the context of the underlying data it is also important for authors to place their results in the context of other existing evidence. As new methods and new data come into use, estimates are updated and, hopefully, improved upon. However, it can be confusing for users to be confronted with differing estimates, whether from the same group over time or from different groups. This makes it imperative that new results are placed in context. Whether authors are updating their own methods and results, or if they are presenting what they believe is an improved set of estimates from other researchers or research groups, authors should present their results in light of the totality of existing evidence. Increased attention to differences in estimates for the same indicator from different sources may have the positive effect of drawing attention to critical data gaps.

Estimates may change due to updated data, changes to the methodological approach, and/or real epidemiological change. A best practice is to perform an analysis comparing datasets and methods to identify why estimates have changed. This allows users to understand more fully the difference between two sets of estimates. However, this may be limited by how complete the documentation is for other estimates. It should also be noted that decomposing the change in estimates is a substantial analytical undertaking, and when it has been done in the past it has typically been published as a stand-alone journal article.
Examples (plot input data with estimates):

Causes of Death (COD) Visualization - IHME

Childmortality.org

Example (classify results by input data quality):

Following the group classifications from the 2008 and 2010 rounds, we continued to classify countries into groups A, B, or C (Table 1). For Group A countries, data from civil registration systems were used directly to calculate estimates of MMR. For countries in Groups B and C, a two-part multi-level regression model was developed to estimate MMRs for all target years.

Table 1. Sources of maternal mortality data used in generating the 2013 maternal mortality ratio estimates

<table>
<thead>
<tr>
<th>Group</th>
<th>Source of maternal mortality data</th>
<th>Number of countries/territories</th>
<th>% of countries/territories in each category</th>
<th>% of births in 183 countries/territories covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Civil registration characterized as complete, with good attribution of cause of death(^a)</td>
<td>67</td>
<td>37</td>
<td>17</td>
</tr>
<tr>
<td>B</td>
<td>Incomplete civil registration and/or other types of data</td>
<td>96</td>
<td>52</td>
<td>81</td>
</tr>
<tr>
<td>C</td>
<td>No national data</td>
<td>20</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>183</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

\(^a\) For Bahamas, Belgium, Iceland, Malta, Saint Lucia and Saint Vincent and the Grenadines (0.1% of global births), the statistical multi-level regression model was used to obtain maternal mortality estimates because the scarce number of maternal mortality events resulted in erratic trends.

Source: 138
Example (show cross-analysis when updating):

Figure 4. Decomposition of differences in U5MR for 1990 and 2011 into differences due to estimation method and differences due to data. The gray box represents differences up to 10 deaths per 1,000 live births.

Source: 140 Note: the goal of this paper is an explicit comparison of results after a change in estimation method by the UN-IGME, not the presentation of the new results.

18 Discuss limitations of the estimates. Include a discussion of any modelling assumptions or data limitations that affect interpretation of the estimates.

Explanation: Any scientific report should include a discussion of the limitations of the study. This is particularly relevant in the global health field, as policy-makers, funding agencies and other decision makers may wish to use the reported estimates to inform action. Such users should be made aware of the limitations of the estimates, especially with regard to how they should be appropriately used. Authors should present a detailed, thoughtful description of the limitations of their estimates. That description will depend on the nature of the study and estimation method(s) used, but a set of generally applicable categories is presented below.

Appropriate use. The reporting of global health estimates is often targeted at policy-makers and other decision makers, but there may be important limitations in how the estimates should be used.7,12 The analyst should provide guidance on which uses are appropriate given the analytic methods and the data availability. Appropriate uses may differ by geographic unit (e.g. for the entire study population vs. well-measured populations vs. poorly measured populations). For example, in cases where out-of-sample predictions are used to produce estimates, such as in countries where there are no data available, the prediction model could not reflect recent health system or programme achievements in that country.12,136 As a result, such estimates are not appropriate to use for programme or health system evaluation. A thoughtful and realistic discussion of the limits to the use of new estimates will help ensure appropriate use, as well as guide future data collection to fill in data gaps.

Assumptions. Any important assumptions used in the analysis should be discussed with regard to how they may impact the results. This may include the effects of analyst choices regarding data inclusion, the type of model used to make estimates, or methods used to adjust for potentially biased data. If the effect of changes in assumptions on the results is known, authors may wish to highlight these impacts. In some cases, the analytic assumptions may have important implications for the appropriate use of the estimates, as discussed above. For example, some models may share a single time trend for multiple countries (an analytical assumption) that has important implications for how the resultant estimates are interpreted (i.e. that trends are indicative rather than reflecting country-specific program achievements).67
Unadjusted bias. In some cases, authors may need to identify known sources of bias that were not adjusted for. This may be particularly relevant in the case where a particular type of data input is known to be biased, but adjustments were not made. For example, a study of adult mortality may make use of sibling survival survey data, but not adjust for known survivor bias. Authors should describe the bias, discuss why methods were not developed to adjust for the source of bias, and indicate how it may have affected the global health estimates (if known).

Data gaps. It may be useful for authors to provide commentary on major data gaps and data collection priorities, particularly with regard to how data availability patterns may have had effects on the final estimates.24

Uncertainty. There may be important limitations in the quantification of uncertainty that authors should discuss. As discussed in items 13 and 16, all sources of uncertainty may not be captured, and authors may discuss this as a limitation of their study.

Future research. The discussion of the limitations of a given study is often a natural segue into identifying the need for future research. If there are areas of a study that are particularly weak, or subject to several caveats and assumptions, it may indicate areas for future work. For example, in the case where it was not possible to adjust for a known bias, this may signal the need for additional primary data collection to clarify the nature of the bias, or a new statistical method to adjust for the bias.68 Authors should flag these future research needs.

Example:

“The main limitation of our analysis is that, despite the extensive data search and access, fewer data were available for anaemia and haemoglobin than for other nutritional and physiological indicators, such as anthropometric status and blood pressure in women and children,14,17,53 especially in the early years of our analysis period. This restriction might have arisen from a scarcity of low-cost portable instruments for field measurement of haemoglobin concentrations. As a result, the estimates might not capture the full variation across countries and regions, and could tend to shrink towards global means when data are sparse. This outcome might have especially affected the estimates in high-income and upper-middle income countries, where anaemia prevalence is low and typically addressed in a clinical setting. Because our analysis unit was age-country-year, we could use only covariates for which we had data for every country-year, and could not incorporate potentially important predictors of haemoglobin with scarce data, especially dietary iron and iron supplementation. Similarly, although we adjusted haemoglobin for altitude and used data adjusted for smoking when available, we could not do so for inflammation because most surveys did not obtain information about inflammation and because no standard adjustment exists. The adverse effects of low haemoglobin are largest in pregnant women and young children, but there are also effects in other groups, including adolescents, elderly people, and men.18 We did not include these groups in our analyses because substantially fewer representative data are available; these other groups could account for 45% of all anaemia cases. Finally, our study focused on national-level patterns of haemoglobin and anaemia; it would be ideal to have information about nutritional indicators at the subnational level—eg, by rural versus urban place of residence, province or state,54 or socioeconomic status.55" 24
References


Journals following the ICMJE recommendations. [http://www.icmje.org/journals-following-the icmje-recommendations/].


64. Ram R, Ural S. Comparison of GDP Per Capita Data in Penn World Table and World Development Indicators. 2014;116(2):639-646.
72. Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. Health Serv Res. 2010;45(5 Pt 2):1456-1467.


