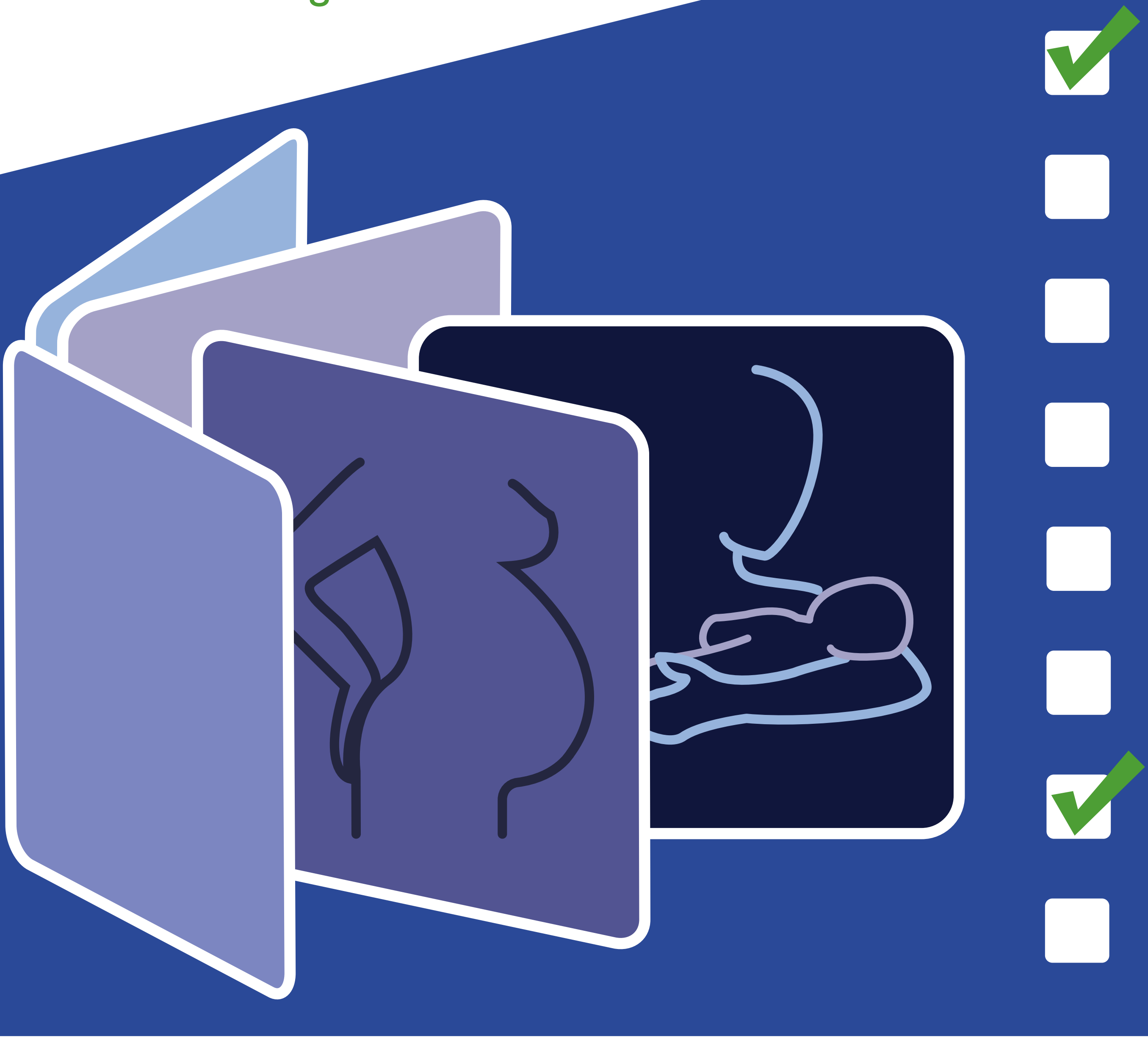
MoNITOR

Monitoring maternal and newborn health

Case studies and recommendations on indicator testing and validation



World Health Organization

Monitoring maternal and newborn health – case studies and recommendations on indicator testing and validation

ISBN XXX-XX-X-XXXXXX-X

Suggested citation. Monitoring maternal and newborn health – case studies and recommendations on indicator testing and validation. Geneva: World Health Organization; 2020. Licence:

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see http://apps.who.int/bookorders. To submit requests for commercial use and queries on rights and licensing, see http://www.who.int/about/licensing.

Third-party materials. If you wish to reuse material from this work that is attributed to a third party, such as tables, figures or images, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.
General disclaimers. 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 WHO 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 WHO 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 WHO 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 WHO be liable for damages arising from its use.

TABLE OF CONTENTS

Acknowledgements v

Abbreviations vi

Introduction 1

Status of maternal and newborn health 1

Role of data, measurement and monitoring 1

Opportunities and challenges at global and country level 1

Purpose 2

Objectives 2

Monitoring maternal and newborn health – Indicator testing and validation 3

Indicators 3

Methods used for indicator assessment 3

Validity testing methodology and application 5

Overview of case studies 5

Case studies – criterion validity 7

Scale-up readiness benchmarks for newborn health *7*

Monitoring childbirth care in primary health facilities in Nigeria *11*

Routine maternal postnatal care indicator validation in Swaziland (Eswatini) and Kenya *15*

Routine newborn health information system indicators *19*

Case studies – convergent validity 22

Tracking facility-based perinatal deaths in United Republic of Tanzania – results from an indicator validation assessment *22*

Randomized comparison of two household survey modules for measuring stillbirths and neonatal deaths *25*

Case studies – construct validity 28

Development of the person-centred maternity care scale *28*

Additional sources 31

Interpretation 32

Interpretation and use of findings from validity studies 32

Disseminating results of a validation study 33

Other considerations 34

Strengths of validity testing approaches included in this report 34

Limitations of validation approaches included in this report 34

Other concepts/issues to consider 34

References 36

ACKNOWLEDGEMENTS

We would like to acknowledge and thank all the individuals who contributed to the conception, writing and execution of this document. Special thanks to Linda Vesel for drafting the document and coordinating with all the case study authors.

MoNITOR advisory group members: Huda Alkitkat, Agbessi Amouzou, Shams El Arifeen, Khalifa Elmusharaf, Mike English, Claudia Hanson, Rajesh Kumar, Alison Morgan, Roger Myrick, Srivanjit Perera, Katherine Semrau, Florina Serbanescu, Lara Vaz and Peter Waiswa.

Case study authors: Patience Afulani, Ann Blanc, Hannah Blencowe, Louise-Tina Day, Tanya Marchant and Melinda Munos.

Technical review group: Lenka Benova, Jo Borghi, Ooona Campbell, Emily Carter, Simon Cousens, Ank Dejonge, Theresa Diaz, Veronique Filippi, Joy Fishel, Asha George, Wendy Graham, Kathleen Hill, Zelee Hill, Debra Jackson, Rima Jolivet, Joy Lawn, Hannah Leslie, Zoe Matthews, Kaveri Mayra, Affette McCaw-Binns, Özge Tunçalp Mingard, Jean-Pierre Monet, Melinda Munos, Moise Muzigaba, Holly Newby, Andrea Nove, Tom Pullum, Jennifer Harris Requejo, Carine Ronsmans, Miriam Sabin, Kuntal Saha, May Sudhinaraset, Jesper Sundewall and Saraswathi Vedam.

World Health Organization Secretariat: Ann-Beth Moller and Allisyn Moran.

Donor: The Bill & Melinda Gates Foundation provided support for this publication.

ABBREVIATIONS

AUC	area under the receiver operating curve
CI	confidence interval
DHIS2	District Health Information Software 2
DHS	Demographic Health Survey
EmONC	emergency obstetric and newborn care
ENAP	Every Newborn Action Plan
EN-BIRTH	Every Newborn – Birth Indicators Research Tracking in Hospitals
EN-INDEPTH	Every Newborn – International Network for the Demographic Evaluation of Populations and Their Health
EPMM	Ending Preventable Maternal Mortality
FBH+	full birth history with additional questions on pregnancy losses
FPH	full pregnancy history
FPM	facility perinatal mortality
HDSS	Health and Demographic Surveillance System
HMIS	health management information system
HSR	health system research
IDEAS	Informed Decisions for Actions in Maternal and Newborn Health
IF	inflation factor
JHPIEGO	Johns Hopkins Program on International Education in Gynecology and Obstetrics
KMO	Kaiser-Meyer-Olkin
MCA	Department of Maternal, Newborn, Child and Adolescent Health
MICS	Multiple Indicator Cluster Survey
MoNITOR	Mother and Newborn Information for Tracking Outcomes and Results
PCMC	person-centred maternity care
PNC	postnatal care
ROC	receiver operating curve
SDG	Sustainable Development Goal
SRH	WHO Department of Sexual and Reproductive Health and Research
USAID	United States Agency for International Development
WHO	World Health Organization

INTRODUCTION

Status of maternal and newborn health

While large gains have been observed in saving lives and improving the health of women and newborns, further progress is needed to achieve the Sustainable Development Goals (SDGs) by 2030 *(1)*. In 2017, approximately 810 women died from complications related to pregnancy and childbirth every 24 hours *(2)*. In 2018, globally an estimated 2.5 million newborns died in the first month of life – approximately 7000 every day *(3,4)*. Most of these deaths are preventable with access to high-quality antenatal, childbirth and newborn care and the targeting of the most vulnerable sick and small newborns *(5–7)*.

A number of key initiatives have been established to facilitate the achievement of the SDGs geared towards reducing maternal and newborn mortality, such as the Global Strategy for Women's, Children's and Adolescent's Health (2016–2030) *(8)*, Ending Preventable Maternal Mortality (EPMM) strategies *(9)*, Every Newborn Action Plans (ENAPs) *(10)* and Countdown to 2030 *(11)*. All of these initiatives highlight the need for valid indicators to monitor progress at global, national and local levels, and to improve the quality of care provision in efforts to enhance maternal and newborn health outcomes *(7–12)*.

Role of data, measurement and monitoring

Measuring different aspects and determinants of the health of women and their newborns is essential to constructing a reliable picture of the state of women's and newborns' health at global, regional, national and subnational levels. It is also instrumental in tracking progress towards achieving the targets of the Global Strategy and the SDGs, as well as other global initiatives *(1,8)*. Accurate measurement enhances our knowledge of whether quality respectful interventions are being provided and received by pregnant, intrapartum and postpartum women and their newborns, and enables the targeting of resources to those most vulnerable. Additionally, it allows policy-makers to prioritize where resources should be directed for maximum impact and to create an enabling environment for the delivery of key interventions.

In order to develop a set of norms and guidance for the measurement and monitoring of maternal and newborn health, the World Health Organization (WHO) convened a group of technical experts well versed in such measurement challenges. They have come together as the Mother and Newborn Information for Tracking Outcomes and Results (MoNITOR) Technical Advisory Group, which acts as an advisory body to the Organization on matters of measurement, metrics and monitoring of maternal and newborn health for the WHO Departments of Maternal, Newborn, Child and Adolescent Health and Ageing (MCA) and Sexual and Reproductive Health and Research (SRH) *(13,14)*. The guidance and norms that they develop will be used to guide improvements in maternal and newborn health metrics and, ultimately, to improve health outcomes *(11)*.

Opportunities and challenges at global and country level

A variety of initiatives are focusing on maternal and newborn health measurement by developing, testing and validating existing and new indicators designed to track progress towards national and global targets *(15–18)*. These initiatives utilize slightly different definitions of validity and methodologies of analysis, thus making the formulation of recommendations both more enriching and more challenging. In the meantime, decision-makers at the country level are working to prioritize indicators and data collection methods to use and invest in while minimizing the burden of data collection and maximizing data quality and the level of impact of the information yielded. Use of indicators with poor validity will not serve as rigorous markers of change and might give a wrong picture of the circumstances surrounding maternal and newborn health. Therefore, countries and global actors need to consider level of validity when choosing which indicators and methods upon which to base their monitoring and policy decisions.

The role of the MoNITOR Technical Advisory Group is to advise WHO on harmonization of validation efforts and to provide standardized guidance and tools for global comparisons and evidence-based decision-making. This document is part of a larger toolkit of resources that MoNITOR developed to facilitate the monitoring of maternal and newborn health at various levels. The toolkit will include guidance on prioritizing indicators for country-level monitoring, indicator reference sheets to provide more detailed information on priority indicators, as well as an online tool (Insert link to toolkit) to facilitate the prioritization of indicators based on a set of standard filters. While the initial focus is on maternal and newborn health indicators (input, process, output, outcome and impact indicators), parts of this document are applicable to the broader metrics community with hopes for future expansion to child and adolescent health indicators.

Purpose

This document provides methodological guidance for stakeholders conducting research on indicator validation and includes recommendations on the following:

1. How to define, design and conduct indicator validation studies for maternal and newborn health indicators.

2. How to interpret and apply the body of evidence from available studies to assess whether an indicator meets a standard of validity.

Objectives

1. To highlight the need for and value of assessing the validity of different types of indicators.

2. To define validity and its application across different indicators.

3. To provide specific recommendations on methods for conducting indicator validation studies.

4. To present recommendations on how results from a single and multiple validation studies can be interpreted and acted upon to make decisions for the prioritization of indicators.

Type of indicators:
This document pertains to maternal and newborn health indicators that are currently being used but for which questions remain around validity. It also includes untested indicators, which we refer to as “aspirational” (for example, policy- and patient-centred indicators).

Target audience:
The document is intended primarily for stakeholders interested in conducting indicator testing. It may also be useful for those individuals tasked with prioritizing the most useful indicators to measure specific strategic outcomes.

Content overview:

- Role of indicators and validity testing.
- Validity testing methodologies, including illustrative case study examples.
- Interpreting the body of evidence from available studies of validity and making decisions on further testing or indicator use.
- Other considerations related to indicator validation, including strengths and limitations of the indicator.

MONITORING MATERNAL AND NEWBORN HEALTH – INDICATOR TESTING AND VALIDATION

Indicators

Before delving into indicator validation, one must define the indicator in question and what it is intending to measure (the concept) using a rigorous scientific process *(19,20)*. The following should be delineated:

1. ***What*** is being measured – detailing the numerator and denominator or the components that make up a composite indicator.

2. ***Who*** it is being measured for – defining the target population(s) for which this indicator will be measured.

3. ***Why*** it is being measured – determining the concept the indicator is attempting to capture; the underlying relationship between the indicator and the concept; and how the estimated levels of the indicator are thought to relate to progress in maternal and newborn health.

4. ***Where*** it is being measured – describing the context or setting in which the indicator is useful.

5. ***How*** it will be measured – determining the data source(s) that will be used to measure the indicator, and the intended frequency of measurement. Examples of common data sources include civil registration and vital statistics systems, censuses, health management information systems (HMIS), population-level surveys, health facility surveys and records, administrative databases capturing financial and human resource data, key informants and policy documents, among others *(16,21)*.

This document groups indicators into five types *(15, 16)*:

1. ***Input:*** Infrastructure, policies, commodities, equipment, human resources and finances that make a programme or intervention possible.

2. ***Process:*** Specific tasks and their execution that are mechanisms for achieving the goals of a programme.

3. ***Output:*** Results of a programme related to service provision, including access, availability, quality and safety.

4. ***Outcome:*** Intermediate results of a programme – namely, coverage of services.

5. ***Impact:*** Long-term outcomes of a programme which it aims to change, such as morbidity, fertility and mortality.

Methods used for indicator assessment

While the focus of this document is on validity testing and three key methods, other kinds of indicator testing are used in the development and implementation of indicators – namely, *reliability* and *accuracy*. Reliability is the ability to obtain the same results repeatedly. Accuracy is the proximity of the result to the exact/true value. Validity testing is a process for ensuring that the indicators being used to monitor maternal and newborn health are measuring what they intend to measure in order to provide accurate evidence to inform national and global programmes *(22)*. The use of indicators with high levels of validity, under appropriate supporting conditions, allows for the collection of prioritized data for informed decision-making, planning and resource allocation.

When conducting validity testing, there are a number of important attributes and considerations to keep in mind. Every validation study is time and place specific. Some validation studies evaluate the validity of using a specific data source or methodology to derive estimates for a set of indicators, while others assess the validity of a specific indicator irrespective of data source or methodology used (see case studies in this document for further information).

Assessing indicator validity is an ongoing process of considering the usefulness of an indicator, and there are no objective validity levels or cut-off points. The main question to evaluate is whether the indicator can be measured well enough to be useful for a specific purpose. The key is to follow and document a rigorous methodology and provide justification for the decisions made on whether and how to use an indicator. When evaluating the validity of using a particular indicator, one must consider both the numerator and denominator. There are some denominators that include the entire population (for example, interventions for all women/newborns), while there are other indicators that require a denominator for a subset of the population. For example, an indicator on newborn resuscitation would only include those newborns that required resuscitation. These types of indicators that look at a subset of the population require particular attention. Validity testing can be conducted as part of a larger intervention or observational study; it does not need to be conducted separately and can be less costly to execute if connected to another study.

VALIDITY TESTING METHODOLOGY AND APPLICATION

Key Methodologies for Validity Testing

There are various approaches to assess validity, of which three – criterion validity, convergent validity and construct validity – are highlighted in this document. The application of these three approaches will be illustrated through a number of case studies covering a spectrum of maternal and newborn health indicators. The three methodologies suit different needs for validation assessment, and sometimes all three can be employed in a single study *(23)*. The following describes each of the approaches:

- **Criterion validity** – comparing two different methods of obtaining an estimate for a specific indicator, one of which is the gold standard, in order to assess the utility of the indicator in producing the intended or true result. This type of validity is also referred to as diagnostic validation. For example, interviewing a mother about a particular service or behaviour compared to observing the service provision or behaviour (gold standard). Criterion validity is mainly used for outcome indicators that measure coverage, but it has also been used to assess impact indicators such as morbidity. When looking at input and output indicators – such as existence of policies and number of services provided – this type of validity assessment might be referred to as verification.
- **Convergent validity** – comparing and exploring various ways of measuring an indicator and determining which is the more appropriate method or if multiple methods need to be triangulated to produce the best estimate. This type of validity is also referred to as triangulation. This approach is utilized when a gold standard does not exist. Convergent validity is often used for impact indicators.
- **Construct validity** – understanding the purpose of an indicator and the larger concept that it represents. Construct validity answers the question of why an indicator is being measured. All validity assessment studies should start with at least a short consideration of construct validity. Construct validity is often used for process indicators. This approach also includes producing an estimate for a more complex indicator through the use of proxies or for qualitative indicators that are often harder ideas to capture, such as quality of care provided *(23)*.

Overview of case studies

The seven case studies included in this document have been written by individual research teams based on existing research studies. The authors of this document reached out to researchers with experience and expertise in validity testing to source case studies illustrating the execution, challenges and strengths of each methodology for various types of indicators. These case studies are not exhaustive examples but cover a wide range of indicators and provide a real-life glance into the application of validity testing methodologies for existing indicators and those that are under development or aspirational. Two case studies address multiple indicators as they were part of the same study.

There are four case studies describing criterion validity, two for convergent validity and one for construct validity (see Table 1). The case studies include indicators that are currently being used in health surveys, as well as aspirational indicators. While the validation process may be different for existing versus new indicators, this document aims to provide a high-level overview of validation. Each case study has a link to a full-text publication (where applicable) and captures the type of indicator tested, geographical location of application of the study, purpose or aim of the study, process (methods and data sources, data collection, data management, quality assurance and statistical methods), summary and interpretation of the results, dissemination of the results and actions based on these results, and finally, lessons learned, such as strengths and limitations of the indicator.

Table 1. Case study indicators by methodology and type

	Criterion validity	Construct validity	Convergent validity
Indicator (Indicator type)	1. Scale-up readiness (output) for newborn health 2. Blood pressure taken at initial client assessment and baby weight at birth (process) 3. Content of postpartum care (outcome) 4. Breastfeeding and neonatal resuscitation (outcome)	5. Facility perinatal mortality (impact) 6. Stillbirth (impact)	7. Experience of care (outcome)

Case study 1

Criterion validity

Scale-up readiness benchmarks for newborn health

Case study authors: Lara Vaz[1] and Deborah Sitrin[2]

Affiliations: [1]Save the Children; [2]Johns Hopkins Program on International Education in Gynecology and Obstetrics (JHPIEGO)

Publication: Moran AC, Kerber K, Pfitzer A, Morrissey CS, Marsh DR, Oot DA, et al. Benchmarks to measure readiness to integrate and scale up newborn survival interventions. Health Policy Plan. 2012;27(3):iii29–39

Validity testing methodology Criterion validity (gold standard)

Type of indicator Output

Specific indicator of interest Scale-up readiness

- This was a composite indicator (for more details, see Moran *et al.* and Afulani *et al.* [Case study 7]) measured on a scale of 0 to 27, depending on how many benchmarks of scale-up readiness are in place per a tool used to measure this indicator.

- The purpose of the indicator is to measure the degree to which the policy, health system and programmes of a given country are prepared to deliver newborn care interventions, or packages, at scale.

- The indicator can also be disaggregated into separate composite indicators of stages in the public policy process – agenda setting, policy formulation and early implementation – to measure scale-up readiness for each stage. Agenda setting is measured on a scale of 0–6, policy formulation on a scale of 0–13, and policy implementation on a scale of 0–8.

Location of study Data collection initially took place in nine countries with diverse contexts: Bangladesh, the Plurinational State of Bolivia, Ethiopia, Malawi, Mali, Nepal, Pakistan, Uganda and United Republic of Tanzania. Countries were selected for analysis based on: (1) high burden of neonatal mortality; (2) capacity of in-country staff; (3) availability of funding; and (4) health system structure around facility- and community-based interventions for newborn survival. Additionally, these were the countries in which Saving Newborn Lives was working. Data collection was subsequently done in two additional countries – Liberia and Zambia – to test the relevance of the indicator in countries that were not part of a global project to support national-level readiness for scale-up of newborn interventions.

Purpose / aim of study

- To validate a new, semi-quantitative approach using readiness benchmarks to assess scale-up readiness, in order to be able to use the benchmarks as proxies to assess progress for global health initiatives in complex systems over time. This is a critical step towards increasing coverage and reducing deaths.

- To describe a methodology that uses a set of benchmarks to measure scale-up readiness – that is, the degree to which the policy, health system and programmes of a given country are prepared to deliver newborn care interventions, or packages, at scale.

Process

Methods and data sources

Benchmarks were identified from an extensive literature review, input from technical and policy experts, and refined following field testing. A total of 27 core benchmarks were selected. Standard definitions were developed for each of the benchmarks and incorporated into an Excel-based tool (called the benchmark achievement tool). The policy benchmarks are coded as “achieved”, “partially achieved”, or “not achieved” in the tool, with additional columns to record the year when the benchmark was accomplished and a reference document (for example, national policy or training manual) as evidence.

Data collection

1. The benchmark achievement tool was completed, as well as a policy and programme timeline.

2. A stakeholder meeting (or series of meetings) and one-on-one interviews were convened to review and modify the findings using a standard process and list of questions. National stakeholders included in-country experts (such as government leaders and development partners), as well as technical experts in policy, advocacy, and maternal, newborn and child health.

3. Benchmark scores were then refined, based on the outcomes of this process.

4. Supporting documents for each benchmark were collected and reviewed for consistency with benchmark scores. Any inconsistencies were resolved through facilitated discussions.

Analysis/statistical methods

Several analyses were conducted.

1. Changes in benchmarks over time for all nine countries were assessed.

2. The 27 benchmarks were divided into three categories using the following pieces of the stages heuristic policy process: (1) agenda setting, (2) policy formulation and (3) policy implementation.

3. The evolution of newborn health policies over time was analysed to address the major causes of newborn death – for example, birth asphyxia, sepsis and low birthweight/prematurity.

4. Benchmarks were also analysed for frequency of attainment as well as year of achievement. Each benchmark was categorized as "not achieved", "partially achieved" or "achieved", and validated based on review of documentation plus consensus generation in each country. A summary for each country was generated by adding the number of benchmarks within each category.

Summary of results and interpretation

At the end of this process there was a wide consensus on the core set of benchmarks and agreement that they were valid and reliable measures of readiness to implement newborn programmes. For the nine countries assessed, achievement of benchmarks increased over time, with evidence of rapid change in some countries, especially after 2005. In 2000, the nine countries had achieved between 0% and 15% of the benchmarks. By 2005, progress had been made in Bangladesh, the Plurinational State of Bolivia and Mali, with about half of the benchmarks completed. Between 2005 and 2010, Malawi, Nepal, Pakistan, Uganda and United Republic of Tanzania made substantial progress, and by 2010, three countries (the Plurinational State of Bolivia, Nepal and Pakistan) had achieved about 80% of the benchmarks, while other countries, with the exception of Ethiopia, had achieved over half of the benchmarks. While Bangladesh and Mali made progress between 2000 and 2005, less progress was made between 2005 and 2010, which differs from the pattern seen in other countries where progress accelerated after 2005. By 2010, among the three categories, countries had made the most progress in agenda setting.

In all countries, a national needs assessment was conducted, and a convening mechanism was established to advance newborn health policy and programmes, either as a group that focused solely on newborn health or as a wider body that also addressed safe motherhood and/or child health. Most of these nine countries were generating and disseminating local evidence for newborn health. All had a national policy for newborn health, either as a stand-alone document or integrated within maternal and/or child health policies. Countries were at various stages for each of the remaining benchmarks.

Dissemination of results and actions based on results

- Follow-up was carried out in the nine countries to continue efforts towards scale-up, leading to early establishment of national Every Newborn Action Plans (ENAPs) in all but three countries. Subsequent review of the benchmarks with full stakeholder engagement was carried out in 2016 in Mali and led to the development of Mali's ENAP by 2017.

- Study results were published.

- The exercise was replicated in 2012 in Liberia and Zambia, with recommendations made for changes/updates.

- Some of the benchmarks, including the existence of a costed plan, are included in the global tracking for the ENAP.

Dissemination of results and actions based on results

This benchmarking methodology allows for an assessment of policy, system and programme readiness to scale up newborn health packages over time and among countries. In the nine countries where the methodology was applied, remarkable progress has been made in readiness to scale up newborn health programmes in the last decade, which continued in subsequent years.

Strengths/achievements

This benchmarking methodology allows for a valid assessment of policy, system and programme readiness to scale up newborn health packages over time and among countries.

Limitations/challenges

- Varying epidemiological and health systems contexts within and between countries.

- Variability over time.

- Data were collected retrospectively, so they may be subject to recall bias.

- Time required by country teams and by support staff to verify documents.

- A few countries have yet to implement newborn interventions or packages of interventions at scale, so it is difficult to assess this indicator's validity in predicting scale-up.

- The benchmarks around newborn health policies and integration into other programmes were sometimes difficult to ascertain, with overlap between policies and programmes. The collection and verification of the benchmarks was a time-consuming process.

- Review of and revisions to the benchmarks will be required to ensure that they remain aligned with global guidance for national newborn health programmes.

Case study 2

Criterion validity

Monitoring childbirth care in primary health facilities in Nigeria

Case study authors: Antoinette Alas Bhattacharya and Tanya Marchant

Affiliations: London School of Hygiene and Tropical Medicine

Publication: Bhattacharya AA, Allen E, Umar N, Usman AU, Felix H, Audu A, et al. Monitoring childbirth care in primary health facilities: a validity study in Gombe State, northeastern Nigeria. J Glob Health. 2019;9(2):020411

Validity testing methodology

Criterion validity (gold standard)

Type of indicator

Process

Specific indicator of interest

- Blood pressure taken – initial client assessment during childbirth
- Baby weighed at birth

Location of study

Ten primary health facilities in Gombe State, northeastern Nigeria. After randomly selecting 107 primary health facilities to review their birth records, 10 facilities with the highest number of births were selected for birth observations.

Purpose / aim of study

To validate the responses of women at different recall periods and the documentation of health-care workers in the maternity register for childbirth events in primary health facilities.

Process

Methods and data sources

Using birth observations as a gold standard, we validated women's recall of childbirth events: (1) before exit from a facility after childbirth and (2) at follow-up 9–22 months postpartum. For a subset of indicators, we also validated health worker documentation of the childbirth events in the maternity registers. All women attending the facility for delivery were invited to participate, excluding women admitted for monitoring before the onset of labour. Four data sources were used to validate childbirth care indicators: (1) birth observations (gold standard), (2) facility exit interviews, (3) household follow-up interviews and (4) facility maternity registers.

1. *Birth observations (gold standard):* Trained observers – local midwives who were not employees of the assigned facility – stayed in the same room to continuously document labour and delivery

processes through the first hour after birth, using a structured checklist. Each facility was assigned two observers and one clinical supervisor to work in shifts and cover all deliveries.

2. *Facility exit interviews:* Each observed woman leaving the facility with a live newborn (usually within 24 hours of delivery) was invited to participate in an exit interview. The exit interview covered information recorded during the observation. We requested permission from the women to conduct a follow-up interview and documented identifying information for this purpose.

3. *Household follow-up interviews (9–22 months after childbirth):* To understand the validity of women's recall in the context of household surveys, we conducted household-level follow-up interviews with a subset of the observed women. The women were asked the same questions as in the facility exit interview. To represent a range of recall periods that may be encountered during a household survey, we selected approximately 150 women from each of the first three rounds of birth observations, which occurred in June 2016 (22 months recall), March 2017 (15 months recall) and August 2017 (9 months recall); this selection was done by a simple random sample of a de-identified list of women observed per round. The follow-up interviews were conducted in March 2018 and the women were asked the same questions as in the exit interview. Up to one week before the household interview, the women were contacted to verify participation in the follow-up interview.

4. *Facility maternity registers:* Following the birth observation, regardless of newborn outcome, the observer extracted data about the woman from the maternity register (Nigeria HMIS, version 2013). Data extraction took place on the same day as the observed birth after the first hour of birth.

Data were collected as part of the Informed Decisions for Actions in maternal and newborn health (IDEAS) project.

Data collection

For our study, 1889 women were observed across the five rounds of birth observations. The following questions were asked in the observation checklist, surveys and maternity registers:

Blood pressure taken – initial client assessment

- Birth observation (gold standard): Was blood pressure taken? (Yes/No/Don't know)

- Women's recall during facility exit interviews and household follow-up interviews: When you were [at the facility], did anyone check your blood pressure, put a strap around your upper arm and take a measurement? (Yes/No/Don't know)

Baby weighed at birth

- Birth observation (gold standard): Was the newborn weighed? (Yes/No/Don’t know)
- Women’s recall during facility exit interviews and household follow-up interviews: Was your baby weighed at birth? (Yes/No/Don’t know)
- Health worker documentation in facility register: Birthweight > 2500 grams: Yes/No

Analysis/statistical methods

We constructed two-by-two tables for each indicator that compared the birth observation to the comparison data-recording method (exit interview, follow-up interview, maternity register). During validation analyses, “don’t know” responses were excluded as they were neither positive nor negative affirmations that the event had occurred.

For two-by-two tables which had five or more counts per cell, we calculated the sensitivity, specificity, area under the receiver operating characteristic curve (AUC) for individual-level reporting accuracy, and the inflation factor (IF) for population-level bias. An AUC $\geqslant 0.70$ was the chosen criteria for high individual-level reporting accuracy and $0.75 > \text{IF} > 1.25$ was the chosen criteria for low population-level bias.

Summary of results and interpretation

During exit interviews, women’s reports of clinical care received had high overall validity (AUC > 0.7 and $0.75 < \text{IF} < 1.25$) for having blood pressure taken before delivery. During follow-up, the indicator met no-validity criteria, with a notable decrease in specificity of the women’s recall. For whether the baby was weighed at birth, 9% of women during exit interviews and follow-up interviews responded “don’t know”, indicating that the extent of recall was insufficient to proceed with validation analyses. However, health worker documentation of whether the baby was weighed at birth met criteria for low population-level bias.

Dissemination of results and actions based on results

- Study results were published.
- Findings were shared in measurement and monitoring forums.

Lessons learned

Strengths/achievements

- Health worker documentation may be a reasonable data source to estimate the occurrence of childbirth-related events, particularly those that do not require the mother’s direct involvement. For our study, maternity registers were able to provide a valid estimate of whether a baby was weighed at birth.

- The study learned from and complemented the findings of previous studies for the study design. Previous criterion validity studies noted the challenges in obtaining valid estimates when asking women to recall the timing and duration of childbirth-related events. In our study, we did not specify any time period during which blood pressure was taken. Without asking women to consider a specific time period, the women's responses provided a valid estimate of the indicator during exit interviews, but not during follow-up interviews.

Limitations/challenges

- Our study reflected a specific context. It reflected the documentation of health workers and the responses of relatively healthy women in a rural primary health facility setting.

- Client management and facility layout may affect a woman's recall of a childbirth event. Services that take place outside of the mother's view or without explanation do not offer the opportunity for a mother to recall an event. While the mother–baby pair were usually kept together immediately after delivery, more than 5% of the mothers in our study did not know if their baby had been weighed at birth. It is important to document the layout of the facility and client flow to contextualize the results.

- Facility setting may point to differences in assumed quality of care delivered (positive facility reporting bias). We validated the responses of women accessing care in primary health facilities only, where sensitivity and specificity were relatively higher compared to previous criterion studies that had included validation for blood pressure taken at initial client assessment. As noted in these previous studies, the results may indicate a positive facility reporting bias, where respondents assume a higher quality of care in hospitals or other referral facilities.

- Observed prevalence can affect estimates of IF. Higher birth observation prevalence can mask a high false-positive rate among the small number of clients that did not have their blood pressure taken during the initial client assessment. As a result, the IF would be largely unaffected. Careful interpretation of the IF is needed to ensure that population-level biases are reflected.

Case study 3

Criterion validity

Routine maternal postnatal care indicator validation in Swaziland (Eswatini) and Kenya

Case study authors: Katharine J. McCarthy and Ann K. Blanc

Affiliations: Population Council

Publication: McCarthy KJ, Blanc AK, Warren CE, Mdawida B. Women's recall of maternal and newborn interventions received in the postnatal period: a validity study in Kenya and Swaziland. J Glob Health. 2018;8(1):010605.

Validity testing methodology

Criterion validity (gold standard)

Type of indicator

Outcome (coverage indicators measured in household surveys)

Specific indicator of interest

Aspirational maternal postnatal care (PNC) indicators not measured in core questionnaires:

- Whether the provider performed a breast exam during the maternal postnatal consultation: "At this postnatal health check, did the provider examine your breasts?"
- Whether the provider performed an abdominal exam during the maternal postnatal consultation: "At this postnatal health check, did the provider examine your abdomen?"

Location of study

Swaziland (Eswatini) and Kenya

Purpose / aim of study

This study sought to improve monitoring coverage of the content of postnatal visits by identifying indicators that women are able to recall with accuracy and which can be practically applied in population-based surveys.

Process

Methods and data sources

We conducted secondary analysis of previously collected, de-identified facility-based data to compare women's reports of PNC received against observations by trained third-party observers using a structured checklist in health facilities located in Eswatini and Kenya. The observations are considered the gold standard measure. Women's reports of care received were collected via an exit interview conducted prior to their leaving the health facility following a PNC visit. Data were initially collected as part of the Integra Initiative, a sexual and reproductive health (SRH)/HIV integration intervention.

Data collection

Client exit interviews and observations of PNC were conducted in 20 public health facilities located in three regions (Lubombo, Manzini and Shiselweni) in Eswatini (n=8) and in Central and Eastern districts in Kenya (n=12). All facilities had high client loads (> 50 infants/ month receiving their first immunizations at six weeks at the PNC-HIV clinics); a minimum of two providers qualified in and currently delivering family planning services; and provided a range of services, including counselling and provision of family planning, voluntary counselling and testing, sexually transmitted infections treatment, and interventions related to the prevention of mother-to-child HIV transmission. Eligible participants were postnatal clients aged 18 years and older attending a consultation on the day of the research team's visit to the facility. If the client was willing to participate, her written informed consent to be interviewed and observed was obtained. Each observed client was interviewed immediately after her consultation to measure her perceptions and recollections of the services received. Observations of the provision of PNC were conducted by a trained third party using a structured checklist. Observations included both client–provider interactions (that is, how clients were treated and whether they actively participated) and the technical content of care. All health-care providers who provided PNC services in the study facilities were invited to participate. If the providers agreed, their informed consent was obtained prior to observation. To reduce the risk of biasing client–provider interactions, more than one day of observations were conducted at each facility to normalize the presence of the observer.

Analysis/statistical methods

Indicator validity was assessed by constructing two-by-two contingency tables to estimate the sensitivity (the true-positive rate) and specificity (the true-negative rate) of each indicator. Receiver operating curve (ROC) analysis, which plots the trade-off between sensitivity against its false-positive rate (or 1-specificity) was used as a summary measure of individual reporting accuracy (Hanley and McNeil, 1982; Macaskill et al., 2010). Quantifying the AUC represents the "average accuracy of a diagnostic test" and is interpreted as "the average sensitivity across all possible specificities" (Macaskill et al., 2010). AUC values can range from 0 to 1, with an AUC of 0.5 indicating an uninformative test (equivalent to a random guess) and 1.0 representing perfect diagnostic accuracy (Hanley and McNeil, 1982). An AUC value of 0.7 or higher was the a priori benchmark for high individual-level reporting accuracy.

To assess the population-based validity of an indicator, we estimated the prevalence (Pr) that would be obtained in a survey given its sensitivity and specificity. Each indicator's estimated sensitivity (SE) and specificity (SP) was applied to its true prevalence (P) (that is, observed prevalence) using the following equation:

$$Pr = P * (SE+SP - 1) + (1 - SP)$$

The ratio of the estimated survey-based prevalence to its true population prevalence (observer report) represents the degree to which each indicator would be over- or under-estimated if assessed using a population-based survey. The a priori acceptability benchmark set for IF was between 0.75 and 1.25.

Summary of results and interpretation

In both Eswatini and Kenya, whether the provider performed a breast exam or an abdominal exam of the mother demonstrated relatively high individual-level accuracy (AUC > 0.70). For the population-level criteria, whether a breast exam was performed met the criteria for low bias in both countries (0.75 < IF < 1.25). However, the indicator of whether an abdominal exam was performed met the low-bias benchmark in Eswatini only. This indicator tended to be overestimated by women in Kenya (IF 1.36).

These findings inform the recommendation of indicators for tracking progress of PNC interventions received by mothers. In contrast to earlier validation research that examined women's reporting accuracy on maternal and newborn health interventions received in the intrapartum and immediate postnatal period (within the first hour of birth) in low-resource settings, results from this study suggest that women are more able to accurately report on aspects of routine physical examination during a return PNC visit (from 24 hours to 10 weeks after birth).

Both the Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) currently collect data on the occurrence of PNC health visits for the mother and newborn. No questions on the content of care mothers received during their PNC visit were included in the DHS or MICS until the eighth round of the DHS (launched in 2019), which includes three questions on the content of the visit. That only one of the two indicators assessed met both validation criteria in both countries suggests that global measures of validity (AUC, IF) should be examined alongside sensitivity, specificity and intervention coverage for a more complete picture of diagnostic accuracy.

Dissemination of results and actions based on results

- Study results were published in a Journal of Global Health Supplement on “Improving coverage measurement in low-resource settings”.
- Findings were presented at the Global Symposium on Health Systems Research in Liverpool in the United Kingdom in October 2018, and during a webinar hosted by MoNITOR.
- The results from this and other validation studies were used to provide input into recommendations for the core questionnaires to be used in the eighth round of the DHS.

Lessons learned

Strengths/achievements

- This study takes advantage of existing data to validate outcome indicators with the potential to be included in large household survey programmes at relatively low cost.
- The study draws on a gold standard of direct observation by a trained third-party observer for indicator validation.
- With an increasing body of work on validation of coverage indicators, some broad patterns are becoming more evident. For example, we are learning that, compared to antenatal and routine postnatal visits, women are less able to report accurately on labour, delivery and immediate PNC indicators. This is at least partially attributable to women experiencing pain, fatigue and high levels of emotion during that time. In addition, questions that require precise recall of the timing of events (such as immediate breastfeeding) and questions that include the use of medical terms (such as the names of specific drugs) tend to have low validity.

Limitations/challenges

- The survey questions tested did not exactly replicate those asked in large household survey programmes such as the DHS (rounds 1–7) and MICS. A study designed explicitly for indicator validation purposes would be able to test existing survey questions as well as test new ones.
- If recall bias affects the validity of women's responses, this study is likely to underestimate it; the analysis is based on an exit interview that took place immediately after the PNC visit, whereas the DHS and MICS are conducted in households and ask about visits that took place months or years in the past. It is worth noting, however, that recall bias is only one of several possible types of biases in surveys and, to the extent that it has been examined, does not appear to be a significant factor for questions with high initial recall accuracy.
- The information recorded by external observers is used here as the gold-standard measure but is nevertheless likely to reflect some level of error or bias. Even well-trained observers may miss some component of a visit or mis-record it.
- The methodology used in the validation analysis (2x2 contingency tables comparing observations to exit interviews) means that it is difficult to validate very-low-prevalence or very-high-prevalence indicators. This is because some cells of the table may have small sample size and low precision, which limits the ability to provide reasonable estimates of sensitivity and specificity. We caution users against generalizing results of this study – particularly population-based findings – to other settings, depending on the prevalence of the intervention.

Case study 4

Routine newborn health information system indicators

Case study author: Louise Tina-Day

Affiliations: London School of Hygiene & Tropical Medicine

Criterion validity

Publication: Day LT, Ruysen H, Gordeev VS, Gore-Langton GR, Boggs D, Cousens S, et al. "Every Newborn-BIRTH" protocol: observational study validating indicators for coverage and quality of maternal and newborn health care in Bangladesh, Nepal and Tanzania. J Glob Health. 2019;9(1):010902.

Validity testing methodology

Criterion validity (gold standard)

Type of indicator

Outcome

Specific indicator of interest

- Immediate breastfeeding measurement in facility "labour and delivery registers" for routine information systems.
- Neonatal bag mask ventilation measurement in facility "labour and delivery registers" for routine information systems.

Location of study

Every Newborn-Birth Indicators Research Tracking in Hospitals (EN-BIRTH) study sites – five comprehensive emergency obstetric and newborn care (EmONC) facilities in Bangladesh, Nepal and United Republic of Tanzania.

Purpose / aim of study

The EN-BIRTH study aims to test the validity of selected newborn and maternal care health intervention indicators (coverage/quality aspects and/or safety) in facilities. This study, as part of the Every Newborn Measurement Improvement Roadmap, and working closely with EPMM, aims to increase the evidence base to inform selection and use of maternal and newborn indicators in national HMIS (particularly District Health Information Software 2 [DHIS2]) and global tracking. Immediate breastfeeding and bag mask ventilation were two of the selected indicators.

Process

Methods and data sources

This was a mixed-methods study. For criterion validation the EN-BIRTH external gold standard of direct observation was compared to data extracted from routine "labour and delivery register" records (and a maternal survey).

Register design categorized for how the indicator is documented included:

- Specific column for breastfeeding/bag mask ventilation
- Nonspecific column (with other data elements)
- No column to record breastfeeding/bag mask ventilation.

Additionally, qualitative research via in-depth interviews and focus group discussions explored barriers and enablers to routine recording, including when it comes specifically to breastfeeding recording.

Data collection

Babies were observed after birth, by trained researchers, for timing of initiation of breastfeeding and bag mask ventilation. Routine “labour and delivery register” records of immediate breastfeeding and bag mask ventilation were extracted by other trained researchers. Both observation and extraction data were collected using customized tablet-based software application with time-stamping. Data were synchronized, uploaded on an in-country central server and regularly backed up. Raw data were encrypted and anonymised before datasets were pooled. Data quality assurance included standardized training, software consistency checks, online data dashboards and biweekly all-site calls to promote collaborative quality improvement initiatives. For approximately 5% of cases, simultaneous supervisor observation and duplicate data verification and extraction were conducted, and variability between individual data collectors estimated by calculating inter-rater reliability using Cohen’s kappa coefficient. Qualitative data were digitally recorded, transcribed and translated into English and managed using NVIVO 12 software.

Analysis/statistical methods

- Observed coverage and register-recorded coverage comparisons were conducted, including ratios (register-recorded:observed and survey-reported:observed).
- Results were stratified by register design.
- Metrics of individual-level validity were assessed.

Summary of results and interpretation

- Immediate breastfeeding: Analysis is ongoing but preliminary results show variable accuracy across the five sites. In one facility, the indicator was not captured in the register, and in other sites all had specific columns. Immediate breastfeeding was slightly to very overestimated compared to observation.

- Bag mask ventilation: Analysis is ongoing but preliminary results show accurate recording of bag mask ventilation in sites, with specific columns to document neonatal resuscitation.

Dissemination of results and actions based on results

- Study results were published.

- Used to inform selection and use of maternal and newborn indicators in national HMIS (particularly DHIS2), and global tracking.

- Used to inform standardization and implementation of routine register design, including feedback to health workers to increase source data accuracy.

Lessons learned

Strengths/achievements

- Multi-site study enabled validation across five contexts.

- Used observation as the external gold standard.

- Large cohort (> 22 000 births) observed.

- Time-stamped capacity of customized tablet app.

- Linkage of validation with quality cascades.

- Qualitative methodology to understand the validation results.

- EN-BIRTH study also validated from survey of maternal report, allowing a direct comparison for the same observation by both survey-reported and register-recorded measurement.

Limitations/challenges

- Length of observation varied by site – validation included babies observed for one full hour and all babies for quality cascades.

- For coverage, the denominator “clinical need” is hard to capture and was not the purpose of this study.

- Uncommon outcomes need large sample sizes, but even then it is not possible to measure specificity and negative predictive value without a true-negative measure.

Case study 5

Convergent validity

Tracking facility-based perinatal deaths in United Republic of Tanzania – results from an indicator validation assessment

Case study authors: Marya Plotkin[1], Dunstan Bishanga[2], Hussein Kidanto[3], Mary Carol Jennings[4], Jim Ricca[1], Amasha Mwanamsangu[2], Gaudiosa Tibaijuka[2], Ruth Lemwayi[2], Benny Ngereza[2], Mary Drake[2], Jeremie Zougrana[2], Neena Khadka[5], Jim Litch[6], Barbara Rawlins[1]

Affiliations: [1]JHPIEGO Baltimore, Baltimore, MD, USA; [2]JHPIEGO Tanzania, Dar es Salaam, Tanzania; [3] Ministry of Health, Community Development, Gender, Elderly and Children, Dar es Salaam, Tanzania; [4] Johns Hopkins Bloomberg School of Public Health, Department of International Health, Baltimore, MD, USA; [5]Save the Children, Washington, DC, USA; [6]Global Alliance to Prevent Prematurity and Stillbirth, Lynnwood, WA, USA.

Publication: Plotkin M, Bishanga D, Kidanto H, Jennings MC, Ricca J, Mwanamsangu A, et al. Tracking facility-based perinatal deaths in Tanzania: Results from an indicator validation assessment. PLoS One. 2018;13(7):e0201238.

Validity testing methodology

Convergent validity

Type of indicator

Impact

Specific indicator of interest

"Facility perinatal mortality" (FPM) indicator – a measure of perinatal deaths (intrapartum stillbirths and very early newborn deaths) out of admissions of pregnant women to the facility in which a fetal heart rate was detected and recorded in the labour and delivery register.

- Numerator (perinatal deaths): Fresh stillbirths plus newborn deaths before discharge from the facility
- Denominator: Admissions to the facility in which fetal heart rate was detected and recorded

Location of study

Kagera Region, United Republic of Tanzania

Purpose / aim of study

- To validate perinatal outcomes as recorded in the national HMIS labour and delivery register at health facilities for use in calculating the FPM indicator.
- To provide health facility staff with tools and examples for calculating the FPM indicator to be linked with quality-of-care improvement efforts.

Process

Methods and data sources

This was a prospective indicator validation study that assessed sensitivity and specificity of perinatal death outcomes recorded in the labour and delivery register (HMIS) compared with a perinatal death audit to verify whether the timing and cause of death were accurately recorded in the register. In addition to assessment of the veracity of perinatal mortality as recorded in the register, the register data were used to calculate the FPM indicator.

Data collection

The study was conducted from November 2016 to April 2017 in 10 high-delivery-volume health facilities in the Kagera Region in United Republic of Tanzania. Providers at participating health facilities were provided with Doppler devices and oriented on their use (some facilities already had Doppler devices while others were using Pinard). During the study period, perinatal deaths that occurred were recorded in the labour and delivery register, “improved” perinatal death audits were conducted, and information was recorded by both health facility and study staff (the latter attended the health facility perinatal death audit meetings and used an enhanced perinatal death audit form to record information).

Analysis/statistical methods

A total of 128 register–audit pairs were examined to look at agreement between the register and the audit form. We calculated positive and negative predictive values of the labour and delivery register birth outcomes to predict gold-standard perinatal death audit outcomes. We also calculated the FPM indicator by health facility by month and overall for the six-month study period.

Summary of results and interpretation

Out of 128 register–audit pairs, in only one case did the audit data differ from the register data (a death was registered as a fresh stillbirth while the audit found that the death was a newborn death). This resulted in a very high sensitivity and specificity (range 95.7–100%) of the audits in predicting type of adverse perinatal outcome in the register. All outcomes (fresh stillbirth, macerated stillbirth and newborn death) had high sensitivity and specificity values. The sensitivity (probability of stillbirth or newborn death in the register given that it was classified as such in the audit) was 95.7%, 100% and 97.8% for fresh stillbirth, macerated stillbirth and newborn death, respectively. The specificity (probability of not stillbirth or not newborn death in the register given that it was classified as such in the audit) was 98.8%, 100% and 97.7%, for fresh stillbirth, macerated stillbirth and newborn death, respectively. Given the high accuracy of the register, the FPM indicator was calculated. Results showed FPM rates that corresponded with levels of health facility, with the regional hospital having a rate of 4.2% of all admissions in which a fetal heart rate was detected experiencing a perinatal death, and district hospitals having an average rate of 2.4%.

Dissemination of results and actions based on results

- Study results were published.
- Following completion of the study, a participatory skill-building exercise on how to calculate the indicator was held in the country for health-care providers and district officials from the participating study facilities and other facilities. The participatory workshop was well received and health-care providers were eager to calculate FPM indicator rates in their facilities, to have reference points for initiatives to improve quality of care.
- Findings were presented to the United States Agency for International Development (USAID)/Washington and to global monitoring and evaluation working groups, including the ENAP/EPMM core monitoring and evaluation working group, and at the first annual Africa Regional Forum on Quality and Safety in Healthcare in Durban, South Africa in February 2018.

Lessons learned

Strengths/achievements

- The FPM indicator is a valuable metric that health facility staff can use to track potentially preventable perinatal deaths that occur after admission to the health facility.
- The indicator can be used to track trends over time and relate to quality-of-care initiatives. To be scaled up, “fetal heart rate upon admission” must be added to health facility labour and delivery registers, and providers should receive orientation on how to record the information, calculate and chart the indicator.

Limitations/challenges

- The indicator does not drop low birthweight or congenital malformations from the numerator/denominator; also, the definition of newborn death is death before discharge rather than death within 24 hours after birth. This is for convenience and accuracy of calculation based on how the information is recorded in the register.
- It is recognized that the indicator may not produce a stable value in health facilities with very low numbers of perinatal deaths. The authors recommend that this indicator might be best calculated at quarterly, biannual or even annual intervals in facilities with fewer than 20 perinatal deaths per month. It may also be useful to calculate at a district level, disaggregating by facility type (health centre versus hospital).

Case study 6

Convergent validity

Randomized comparison of two household survey modules for measuring stillbirths and neonatal deaths

Case study authors: Joseph Akuze[1,2]; Hannah Blencowe[1] and Joy E Lawn[1] on behalf of the Every Newborn-INDEPTH study Collaborative Group

Affiliations: [1]London School of Hygiene & Tropical Medicine; [2]Makerere School of Public Health

Publication: Baschieri A, Gordeev AS, Akuze J, Kwesiga D, Blencowe H, Cousens S, et al. "Every Newborn - INDEPTH" (EN-INDEPTH) study protocol for a randomised comparison of household survey modules for measuring stillbirths and neonatal deaths in five Health and Demographic Surveillance Sites. J Glob Health. 2019;9(1):010901.

Validity testing methodology

Convergent validity

Type of indicator

Impact

Specific indicator of interest

Stillbirth rate in household surveys

Location of study

The Every Newborn– International Network for the Demographic Evaluation of Populations and Their Health (EN-INDEPTH) study was undertaken in five Health and Demographic Surveillance Systems (HDSS) sites: Matlab in Bangladesh, Dabat in Ethiopia, Kintampo in Ghana, Bandim in Guinea-Bissau, and Iganga-Mayuge in Uganda, and with coordination by the London School of Hygiene & Tropical Medicine in partnership with Makerere University, Uganda.

Purpose / aim of study

The EN-INDEPTH study was a cross-sectional multi-site study undertaken in five HDSS sites that were part of the INDEPTH network, aiming to inform improvements in measurement of pregnancy outcomes through population-based household surveys. The primary objective was to randomly compare two methods of retrospective recording of pregnancy outcomes:

- Full birth history with additional questions on pregnancy losses (FBH+), as per the current standard in the seventh wave of the DHS (DHS-7).
- Full pregnancy history (FPH).

The study also investigated the performance of existing/modified survey questions for other pregnancy-related outcomes (including existing/modified questions on pregnancy intendedness, gestational age, birthweight, termination of pregnancy, and birth and death certification), and undertook qualitative research regarding barriers and enablers to reporting of these pregnancy outcomes.

Process

Methods and data sources

A population-based survey of women of reproductive age was undertaken (July 2017 – August 2018). Stillbirth rate was the key pregnancy outcome indicator on which the study was powered.

Data collection

Tablet-based data collection was used at all sites, including accurate measures of timing for the survey sections.

- Randomized comparison of the two survey modules through a population-based survey of 69 150 women of reproductive age, of which 34 779 were randomized to FBH+ and 34 371 were randomized to FPH.
- Qualitative work – 34 focus group discussions were undertaken with women and data collectors to seek to improve understanding of barriers and enablers to capturing pregnancy outcomes (including stillbirth) in household surveys.

Analysis/statistical methods

The difference between time to administer a FPH and a FBH+ were calculated together with 95% confidence intervals (CIs).

Summary of results and interpretation

- *Timing:* Minimal differences between time to administer a FPH (10.5 minutes, 95% CI: 10.4–10.6) compared to a FBH+ (9.1 minutes, 95% CI: 9.0–9.3).
- *Stillbirths:* Overall, the capture of stillbirths was 21% higher with the FPH approach compared to FBH+ (95% CI: 10–62%). There was a high level of between-site heterogeneity in the results. Contributing factors include variations in training and survey implementation.
- *Implications:* FPH took an average of 1.4 minutes more than FBH+, yet has potential to capture information on more stillbirths from countries with the highest burden.

FPH is therefore recommended over FBH+, but standardization of interviewer training and consistent implementation of surveys will be important to maximize data quality.

Dissemination of results and actions based on results

- Preliminary results have been submitted to DHS as part of the consultation process for the eighth round of DHS surveys (DHS-8).
- Results are to be published in a peer-reviewed publication; also planned are a series of more detailed papers on the other pregnancy outcomes.

Lessons learned

Strengths/achievements

- First direct randomized comparison of FPH and FBH+ undertaken across five contexts in sub-Saharan Africa and South Asia.
- First study to collect detailed information on time to administer questions.
- Innovative tablet-based data collection.

Limitations/challenges

- Differences in training and survey implementation, particularly at one site (2–3 hours of training on FPH compared to 2–3 days of training), are likely to have contributed to between-site heterogeneity.

Case study 7

Construct validity

Development of the person-centred maternity care scale

Case study authors: Patience A. Afulani[1] and May Sudhinaraset[2]

Affiliations: [1]University of California, San Francisco; [2]University of California, Los Angeles

Publication: Afulani PA, Diamond-Smith N, Golub G, Sudhinaraset M. Development of a tool to measure person-centered maternity care in developing settings: a validation in a rural and urban Kenyan population. Reprod Health. 2017;14(118).

Validity testing methodology Construct validity

Type of indicator Outcome

Specific indicator of interest Experience-of-care measure

Location of study Kenya and India

Purpose / aim of study The purpose of the study was to develop and validate a multidimensional scale to measure person-centred maternity care (PCMC) to summarize women’s experiences of care. Despite growing attention on women’s experiences of care during childbirth, there is a lack of validated indicators. Experience of care is a complex construct with multiple domains, thus it cannot be measured with a single question.

Process *Methods and data sources*

We used the following standard procedure for scale development and validation:

- We defined PCMC as “providing maternity care that is respectful and responsive to individual women and their families’ preferences, needs and values, and ensuring that their values guide all clinical decisions." The 10 domains of PCMC are: (1) dignity and respect, (2) autonomy, (3) privacy and confidentiality, (4) communication, (5) social support, (6) supportive care, (7) predictability and transparency of payments, (8) trust, (9) stigma and discrimination, and (10) health facility environment.

- Item generation – we developed an item pool with questions capturing each of the domains rated on a 5-point scale ranging from 1: "strongly agree" to 5: "strongly disagree".

- Expert reviews – the domains and items were then evaluated through expert reviews and focus group discussions.

- Cognitive interviews (see Annex 1 for definition) – we used cognitive interviews to: assess whether the questions were being interpreted as intended; evaluate problems with the wording of questions; evaluate whether questions were context appropriate and salient; and finally, assess appropriate length of the tool. This exercise reduced the number of items to 38, with each question containing a 4-point response scale: "no, never", "yes, a few times", "yes, most of the time" and "yes, all the time".

- Pre-testing – Revised items were pretested with the full questionnaire among a convenience sample of 39 women in the participating facilities.

Data collection

The final set of items was administered as part of two separate surveys in Kenya:

- Rural sample: In Migori County, women who had delivered in the nine weeks preceding the survey recruited from health facilities and in their homes (n=1052).

- Urban sample: In Nairobi and Kiambu Counties, women who had delivered within a week of the survey (n=531).

Analysis/statistical methods

We conducted psychometric analysis to assess the validity and reliability of the tool, and assured content validity through the literature and expert reviews. We then used factor analysis to reduce the number of items and to assess construct validity. Following extraction of the final items, we regressed the scores from the main scale and subscales on women's ratings of their satisfaction with the services, their perception of the quality of care they received during childbirth, and whether they would give birth in the same facility if they were to have another baby. Finally, we assessed the internal consistency reliability using Cronbach's alpha.

Summary of results and interpretation

The psychometric analysis yielded a valid and reliable 30-item scale with three subscales for “dignity and respect”, “communication and autonomy” and “supportive care”. The scale has high content validity based on our literature and expert reviews. The exploratory factor analysis suggests high construct validity – the items measure an underlying construct, which we believe to be PCMC, based on the content validity. It also has high criterion validity, being strongly correlated with global measures of satisfaction and quality of maternity care. In addition, it has high internal reliability, with an alpha value well above the recommended level of 0.7. These subscales also have good content, construct and criterion validity, with reliability within acceptable ranges of 0.6 to 0.8.

Dissemination of results and actions based on results

- Results have been published in two publications and presented at several meetings.
- A shorter 13-item version using data from the three countries as well as expert input has been developed.
- Data from Ghana, India and Kenya were used to examine PCMC across different settings and to highlight the key areas that are lacking across the settings.
- Future studies can validate this tool to assess its appropriateness for the setting it is to be used. It can be administered through exit interviews as well as through community interviews.

Lessons learned

Strengths/achievements

- The PCMC scale is among the first validated tools for measuring women’s experiences of care during childbirth in a low-resource setting.
- The PCMC score allows for assessing women’s experiences along a continuum rather than as a binary or other categorical variable. The scale can also be used for needs assessments as well as for monitoring and evaluation of the interventions.

Limitations/challenges

- The study samples have not been based on nationally representative samples, which affects their generalizability.
- Validation has been in only three countries, so it is unclear how the indicator will perform in other settings.
- Some may consider 30 items to be too many; the 13-item version helps to address this limitation.

Additional sources

In addition to the papers referenced in the case studies above, please find below a number of additional, useful resources on validity testing. Please note that a comprehensive systematic review of all literature on validity testing was not conducted for the purpose of this guidance document.

- Afulani PA, Diamond-Smith N, Phillips B, Singhal S, Sudhinaraset M. Validation of the person-centered maternity care scale in India. *Reprod Health*. 2018;15:147.
- Afulani PA, Feeser K, Sudhinaraset M, Aborigo R, Montagu D, Chakraborty N. Toward the development of a short multi-country person-centered maternity care scale. *Int J Gynecol Obstet*. 2019;146(1):80–7.
- Blanc AK, Warren KJ, Kimani J, Ndiwiga C, RaoRao S. Assessing the validity of indicators of the quality of maternal and newborn health care in Kenya. *J Glob Health*. 2016;6(1):010405.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
- Macaskill A, Taylor E. The development of a brief measure of learner autonomy in university students. *Stud High Educ*. 2010;35(3):351–9.
- McCarthy KJ, Blanc AK, Warren CE, Kiwani J, Mdawida B, Ndwidga C. Can surveys of women accurately track indicators of maternal and newborn care? A validity and reliability study in Kenya. *J Global Health*. 2016;6(2):020502.
- Stanton CK, Rawlins B, Drake M, dos Anjos M, Cantor D, Chongo L, et al. Measuring coverage in MNCH: Testing the validity of women’s self-report of key maternal and newborn health interventions during the peripartum period in Mozambique. *PLoS One*. 2013;8(5):e60694.

INTERPRETATION

Interpretation and use of findings from validity studies

The case studies presented earlier provide examples of how validity testing is carried out and how results are interpreted, used and disseminated. Based on interviews conducted with key informants and experts working in the field of indicator validation, it was found that there is no specific cut-off point for an “acceptable” level of validity. Respondents shared the view that the level of validity is contingent on the envisioned use of the indicator itself.

Below, please find a number of elements to consider when interpreting results and preparing for dissemination *(13,16,22,24,25)*.

- Usefulness of the indicator in terms of its potential to contribute to improvements and changes in maternal and newborn health.
- Relevance of the indicator to the area of interest or measurement.
- Importance of the indicator for decision-making.
- Endorsement of the indicator by organizations that serve as experts/authorities in the field.
- Study design and statistical methods used (including choice of gold standard).
- Considering evidence from multiple studies, accounting for the methodologies used in each of the studies and the respective levels of evidence they yielded, and determining how the findings align with each other.
- Generalizability of the indicator to different populations and adaptability of the indicator in the presence of contextual changes.
- Contextual factors and quality of documentation are accounted for when interpreting the results of validity studies.
- Feasibility or ease of use of the indicator.
- Repeated use of the indicator yields the same result.
- Indicator can be used over time, is supported politically, and there is technical capacity to use it.
- Sufficient sample size is included in validity testing to be able to assess specificity.
- Harmonization of the indicator with existing portfolios of indicators, where relevant, to adequately monitor a programme.
- Distinctiveness of the indicator to avoid duplication.

Disseminating results of a validation study

As seen in the case studies above, there are several ways to summarize and disseminate the evidence from validity testing studies, including publications, presentations, systematic reviews, reports and policy briefs. Experts in validity testing have shared a number of tips around the dissemination of results, which are summarized below:

- It is often more appropriate and useful to share the results for an indicator under consideration in the context of the overall body of evidence available for that particular indicator (for example, if validity testing has been done elsewhere or if the indicator is part of a large suite of indicators monitoring one process or outcome).
- A single study should not be used for the basis of decisions such as discontinuing the use of an indicator.
- Disseminating validation study results one by one may confuse country-level policy-makers.
- Finally, it is essential to engage the researchers, policy-makers and individuals responsible for data collection upon whom the indicator relies when framing and disseminating findings.

Please see the case studies for various examples of how dissemination was carried out.

OTHER CONSIDERATIONS

Strengths of validity testing approaches included in this report

Validity testing provides the means to assess policy, system and programme readiness to scale up maternal and newborn health packages over time and in various contexts. It allows for the identification of appropriate sources for accurate data collection and the identification of a gold standard against which to determine validity. Validity testing also enables the evaluation and introduction of new indicators, such as the PCMC scale for measuring women's experiences of care during childbirth in a low-resource setting. Each case study shared above highlights the strengths of the validity testing approach used.

Limitations of validation approaches included in this report

While this document serves to provide overarching guidance and aims to standardize pieces of the indicator validation process, it should be emphasized that this guidance document is not all-encompassing, and that validity testing has its limitations. Please find below a number of potential limitations to keep in mind and to try to address when conducting and interpreting the results of validity testing. Please also refer to the case studies for examples of potential limitations by type of indicator and methodology.

Context and resources:

- Context in which validity testing is done may not be typical of most environments where the indicator will be used.
- Systems in place and the quality and availability of the data used to measure indicators (such as epidemiological data, health system, structure, strength of routine HMIS, data literacy) may vary widely.
- Level of effort required by country teams and by support staff to verify documents should be considered.
- It is difficult to validate very-low-prevalence or very-high-prevalence indicators, and the ability to extrapolate results to other settings depends on the prevalence of the intervention.
- The validation samples should be based on nationally representative samples in order to facilitate generalizability.

Methodology and potential biases:

- Hawthorne effect that questions whether results obtained during validation studies will hold true during implementation.
- Recall bias if data were collected retrospectively.
- The information recorded by external observers is used here as the gold-standard measure but is nevertheless likely to reflect some level of error or bias. Even well-trained observers may miss some component of a visit or mis-record it.
- Variability over time.
- Uncommon outcomes need large sample sizes, but even then, it is not possible to measure specificity and negative predictive value without a true-negative measure.
- Observed prevalence can affect estimates of IF.

Other concepts/issues to consider

This guidance document is meant to provide a foundation for validity testing by outlining the importance and utility of validity studies, defining the key terms and methods and providing real-life examples of their application. As mentioned earlier, validity testing is not a black-and-white process and there are many grey areas to keep in mind. While not all indicators can be validated, it is important to understand validity and what is needed. Poor-quality data despite valid indicators, as well as good-quality data with poor indicator validity, will not provide accurate or useful results. In certain circumstances, no matter the results of an indicator validity study, the use of the indicator depends on a larger system with sensitivity to timing of introduction. An indicator may be valid in theory but not necessarily so under certain country circumstances. We hope that this guidance document can be useful for those developing new indicators to be used in the future by laying out best practices to employ and common challenges to avoid as part of a larger capacity-building effort to implement new indicators. Finally, although this document is targeted towards stakeholders interested in conducting research on indicator testing or validation, it may also be helpful for those tasked with indicator prioritization and related policy formulation.

REFERENCES

1. The Sustainable Development Goals report 2019. New York: United Nations; 2019 (https://unstats.un.org/sdgs/report/2019/The-Sustainable-Development-Goals-Report-2019.pdf, accessed 14 August 2020).

2. Trends in maternal mortality 2000 to 2017: estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division. Geneva: World Health Organization; 2019 (https://www.unfpa.org/featured-publication/trends-maternal-mortality-2000-2017, accessed 14 August 2020).

3. Maternal mortality. In: UNICEF Data [website]. New York: United Nations Children's Fund; 2019 (https://data.unicef.org/topic/maternal-health/maternal-mortality/, accessed 14 August 2020).

4. Neonatal mortality. In: UNICEF Data [website]. New York: United Nations Children's Fund; 2019 (https://data.unicef.org/topic/child-survival/neonatal-mortality/, accessed 14 August 2020).

5. Alkema L, Chou D, Hogan D, Zhang S, Moller AB, Gemmill A, et al. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. Lancet. 2016;387(10017):462–74. doi:10.1016/S0140-6736(15)00838-7.

6. Levels and trends in child mortality report 2019. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation. New York: United Nations Children's Fund; 2019 (https://www.who.int/maternal_child_adolescent/documents/levels_trends_child_mortality_2019/en/, accessed 14 August 2020).

7. Newborns: reducing mortality. In: World Health Organization [website]. Geneva: WHO; 19 September 2019 (https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality, accessed 14 August 2020).

8. Every Woman Every Child. The Global Strategy for Women's, Children's and Adolescents' Health (2016–2030). New York: United Nations; 2015 (http://globalstrategyeverywomaneverychildorg/, accessed 14 August 2020).

9. Strategies toward ending preventable maternal mortality. Geneva: World Health Organization; 2015 (http://who.int/reproductivehealth/topics/maternal_perinatal/epmm/en/, accessed 14 August 2020).

10. Every newborn: an action plan to end preventable deaths. Geneva: World Health Organization; 2014 (https://www.who.int/maternal_child_adolescent/documents/every-newborn-action-plan/en/, accessed 14 August 2020).

11. Countdown to 2030: tracking progress towards universal coverage for women's, children's and adolescents' health. New York: United Nations Children's Fund; 2017 (https://www.countdown2030.org/reports-and-publications/countdown-2017-report, accessed 14 August 2020).

12. Jolivet RR, Moran AC, O'Connor M, Chou D, Bhardwaj N, Newby H, et al. Ending preventable maternal mortality: phase II of a multi-step process to develop a monitoring framework, 2016–2030. BMC Pregnancy Childbirth. 2018;18(1):258. doi:10.1186/s12884-018-1763-8.

13. Mother and Newborn Information for Tracking Outcomes and Results (MONITOR) technical advisory group. In: World Health Organization [website] (https://www.who.int/maternal_child_adolescent/epidemiology/monitor/en/, accessed 14 August 2020).

14. Moran AC, Moller AB, Chou D, Morgan A, El Arifeen S, Hanson C, et al. 'What gets measured gets managed': revisiting the indicators for maternal and newborn health programmes. Reprod Health. 2018;15(1):19. doi:10.1186/s12978-018-0465-z.

15. Moran AC, Jolivet RR, Chou D, Dalglish SL, Hill K, Ramsey K, et al. A common monitoring framework for ending preventable maternal mortality, 2015-2030: phase I of a multi-step process. BMC Pregnancy Childbirth. 2016;16:250. doi:10.1186/s12884-016-1035-4.

16. Moller AB, Newby H, Hanson C, Morgan A, El Arifeen S, Chou D, et al. Measures matter: a scoping review of maternal and newborn indicators. PLoS One. 2018;13(10):e0204763. doi:10.1371/journal.pone.0204763.

17. Munos MK, Blanc AK, Carter ED, Eisele TP, Gesuale S, Katz J, et al. Validation studies for population-based intervention coverage indicators: design, analysis, and interpretation. J Glob Health. 2018;8(2):020804. doi:10.7189/jogh.08.020804.

18. Arnold F, Khan SM. Perspectives and implications of the Improving Coverage Measurement Core Group's validation studies for household surveys. J Glob Health. 2018;8(1):010606. doi:10.7189/jogh.08.010606.

19. McGlynn EA, Asch SM. Developing a clinical performance measure. Am J Prev Med. 1998;14(3 Suppl):14–21. doi:10.1016/s0749-3797(97)00032-9.

20. Saturno-Hernandez PJ, Martinez-Nicolas I, Moreno-Zegbe E, Fernandez-Elorriaga M, Poblano-Verastegui O. Indicators for monitoring maternal and neonatal quality care: a systematic review. BMC Pregnancy Childbirth. 2019;19(1):25. doi:10.1186/s12884-019-2173-2.

21. Global reference list of 100 core health indicators (plus health-related SDGs). Geneva: World Health Organization; 2018 (http://www.who.int/healthinfo/indicators/2018/en/, accessed 14 August 2020).

22. Benova L, Moller AB, Moran AC. "What gets measured better gets done better": the landscape of validation of global maternal and newborn health indicators through key informant interviews. PLoS One. 2019;14(11):e0224746. doi:10.1371/journal.pone.0224746.

23. Benova L, Moller A, Hill K, Vaz LME, Morgan A, Hanson C, et al. What is meant by validity in maternal and newborn health measurement? A conceptual framework for understanding indicator validation. PLoS One. 2020;15(5). doi:10.1371/journal.pone.0233969.

24. Family planning and reproductive health indicators database. Selection of indicators. In: Measure Evaluation [website] (https://www.measureevaluation.org/prh/rh_indicators/overview/rationale2, accessed 14 August 2020).

25. Health indicators: conceptual and operational considerations. In: Pan American Health Organization [website] (https://www.paho.org/hq/index.php?option=com_content&view=article&id=14401, accessed 14 August 2020).