Research project title:

The landscape of validation of global maternal and newborn health indicators through key informant interviews

Version: October 22, 2019

Table of Contents

Dataset description

We interviewed experts in measurement of maternal and newborn health indicators using purposive sampling until thematic saturation was achieved. First, a list of potential KIs was developed in discussion among the co-authors with input from the MoNITOR co-chairs, and further expanded using snowball methods to encompass experts on the five types of maternal and newborn indicators (health system and input, care access and availability, quality of care and safety, coverage and outcomes, as well as impact). We included experts in both qualitative and quantitative methods to assessing indicator validity. The final sample included 32 respondents, of which 22 were measurement experts based in academic institutions, two were from United Nations agencies, two from implementing agencies, four from organisations funding research and programmes in maternal and newborn health, and two from data collection organisations. Of the potential respondents approached, two suggested an alternative respondent within their organisation due to time constraints or better suitability of expertise. The suggested alternative respondents were interviewed in both cases. All the remaining potential respondents agreed to an interview.

Interviews were conducted during face-to-face meetings (six) or by telephone/Skype calls (25) between December 2017 and November 2018 and ranged between 45 and 90 minutes. Each respondent was interviewed once. Interviews were conducted in English, guided by a semi-structured interview guide (Appendix 1). The interview guide was pre-tested on the first five respondents and modified. The interviews were not recorded. Detailed notes were taken in shorthand during the interviews, transcribed and expanded immediately following the interview. Following each interview, further written materials and publications that were referred to by respondents were exchanged through email correspondence with several respondents.

Anonymisation

This dataset consists of anonymised notes from these 32 interviews. Sections where respondents discussed their own research were redacted to ensure anonymity. We do not replace names with respondent numbers or pseudonyms as there is the potential that respondents could be identified through the sequence of their responses to the five questions. The order of respondents is shuffled across the five questions. Each bullet point ("-") marks a response by one respondent. There might be several responses by one respondent to each question. Not all respondents replied to each of the five questions.

Abbreviations

AMSTL	Active management of third stage of labour
ANC	Antenatal care
AUC	Area under the curve
CD2030	Countdown to 2030
CHERG	Child health epidemiology reference group
ENAP	Every Newborn Action Plan
EPMM	Ending preventable maternal mortality
EWEC	Every Woman Every Child
DHIS2	District health information system
DHS	Demographic and Health Survey
FBD	Facility-based delivery
FP	Family planning
HIC	High-income country
HIV	Human immunodeficiency virus
HMIS	Health management information system
ICM	Improving Coverage Measurement
IF	Inflation factor
LMIC	Low- and middle-income country
mCPR	Modern contraceptive prevalence rate
MDG	Millennium Development Goal
MICS	Multiple indicator cluster survey
MMR	Maternal mortality ratio
MNH	Maternal and newborn health
NCD	Non-communicable disease
NMR	Neonatal mortality rate
PNC	Postnatal care
RBF	Results-based financing
RMNCAH	Reproductive, maternal, newborn, child and adolescent health
SBA	Skilled birth attendant
SDG	Sustainable Development Goal
SRH	Sexual and reproductive health
SRHR	Sexual and reproductive health and rights
UN	United Nations

Interview notes

Question 1. What does the concept of "validity" of indicators mean to various stakeholders?

- Validity can be seen on a continuum:
    - In French, validation ("endorsement") means to collect data and report findings to stakeholders and getting acceptance of findings – meaning – results are not true until accepted
    - It can also capture whether what you are trying to measure makes sense – is it actionable? Useful? – can stakeholders understand and use indicators?
- There are several issues in validation – the importance of numbers is recognised globally, but users are not critical about where they are from (only want "tweetable numbers") – not even confidence intervals are used/recognised (country example [country name redacted] of different maternal mortality ratio estimates which were available, the country chose to use the higher number for 1990 and lower for 2015 in order to be categorised as having met MDG5.
    - But where are data from, what does it measure, does it measure what impact is about?
    - Global donors are obsessed with indicators and cross-country comparison
    - But nationally – more contextualised, subnational and regularly updated indicators are needed/wanted
    - National policymakers don't care about cross country comparisons or comparability
    - There is a bias in global health funding – not supporting national decision makers to look at subnational variation and its sources
    - Primary goal – indicators used in MNH and of quality/coverage have not changed in 30 years, we need to advance measurement – including in DHS, routine data, global SDG reporting
- Validity speaks to whether the indicators are useful, tell us what we need to know, and whether they are scientifically valid
- "Validity" of indicator – how closely does it relate to a gold standard
    - What we want to know at population level?
    - Where do we need surveys and where can we use routine data instead?
- Do indicators measured relate to the facts in the real world? Do they provide a representation of the real world/true account of what we want to know about it?
    - Indicators of MNH financing – unlike mortality in terms of measurement – much more difficult to draw a line about what should be included – what do we know across countries?
    - SRH indicators - impossible to report properly on financing without deconstructing each project funded – on the donor side, concern about how little is disaggregated, categorisation of funding expenditures is not done with financing indicators in mind – questions validity of these indicators
- There is a widespread assumption that indicators based on the DHS are validated/valid, but that is not the case – most questions haven't been validated or even cognitively tested – but it's really nice to see work going on at the moment in this area
- The DHS contain questions primarily about what people are interested in measuring – experts have decided about how that was to be done, and while DHS have done work to pilot questionnaires, this is to get more general feedback not validation
- An indicator can be valid, but useless for the health system
- Indicators should show progress in real time, moving from population survey estimates to understand facility-level processes
- Importance of validity:
    - Policy makers are not interested in how things are measured – only want top line numbers (MMR/NMR) and how they compare to other countries
    - New indicators are getting accepted slowly
    - Bigger questions of reliability of data collection systems – why is data collected and what is it used for – the amount of data only going upwards in reporting systems is increasing – nothing comes down in form of feedback/supervision – all about numbers of services and not content of care (no incentives to report correctly)
- Historically, validity of indicators derived from data from household surveys was understood using different types of analyses of validity:
    - Internal consistency (indicators tracking together, stratified by wealth etc)

    - External consistency (across different types of data sources)
    - Reliability + lots of psychometric testing that informed DHS/MICS – such as respondents understanding the question
    - This was up until about 10 years ago, similar to FP/HIV world (sexual practices, FP method use), when the MNCH – CHERG Group raised concern about indicator validity (treatment for pneumonia with antibiotics, concern about denominators; maternal – SBA, FBD, csections – not measuring content of intrapartum care – and interested to see if we could do so in household surveys
        - Started conceptualising validation in a diagnostic way (against an independent gold standard)
        - A lot of interest in that approach, ICM continuing in this work currently
- Increasing recognition that there are some things women don't know and cannot report (acknowledging limitations of household survey data) and that facility data need to be used instead (intrapartum and postnatal indicators)
    - Ongoing work from several research groups (ENAP, EPMM, IDEAS) on how to get this facility estimates to population level (effective coverage)
- Methodological work on whether and how facility and routine HMIS data can be used to measure effective coverage; what is collected on facility surveys and how predictive that is of care content
- In French – validation means sitting down with experts to discuss and agree on indicators
- Social sciences - truth cannot be independently observed
- Indicators are important but not the end goal
    - "In general it might be good if we distinguished between indicators that support detailed work enabling us to plot a direction of travel based on need – which may be quite diverse and more complex - and indicators of health system performance that might be usefully measured at scale and over prolonged time periods." (quote from follow-up email)
- Key issues to highlight for MoNITOR work:
    - Rigour as part of indicators
    - What is "valid" – what are appropriate criteria to assess validity
    - Validity should be a consideration in assessing an indicator short-list
    - Establish a common language/understanding
    - Identify where indicators could be misleading
    - Plan on what to do with indicators – what is the action plan if poor validity found?
    - For some issues, do we need to be busy developing/measuring indicators (i.e. mistreatment in childbirth) or should we just prioritise doing something to remedy the situation?
- "An indicator is valid if the expenditure on action based on the indicator is justified – i.e. it is consistent with the values and goals of the health system at that point in time" (quote during interview)
- Validation seems to be becoming more important in the MNH field – this is to do with the development of science in the field – hard science behind it and agreement in the field on importance of measurement (diagnostic style with gold standard, concrete indicators), but we have 20+ years of data to analyse (unlike other fields such as child protection, literacy)
- MDGs were "health indicators on steroids" – pushed the field to improve measurement
- Clinical approach to validation seems reasonable and is fairly easy to do (observe women in facilities)
- Appetite for more data is there – increasing burden on surveys – when money on methodological work, including validation, is little.
    - There is no return on these additional indicators – how are they being used by countries to improve policy change?
    - New indicators should show impact within countries – not just published papers
    - Time to stop collecting ANC coverage (in the 90% in most countries) – need to improve quality – focus on that
    - Pockets of countries have issues – 5-8 countries that have very high rates – but why do all countries need to measure SBA?
    - Countries need to do a better job monitoring indicators, rather than just measuring
- The terminology – "a valid indicator" – needs to be understood separately. The indicator might be valid/sensible – i.e. meaningful, but the measurement of it might be poor – either the data,

the enumerators, or might be unmeasurable at all (e.g. % of children who needed resuscitation that got it)

- Within result-based financing, validity refers to verification of reported results – to enable payment based on performance
    - But validation is broader – do indicators we use actually measure what we are supposed to be measuring?
    - Indicators include structure, process and outcome to put together a picture of quality
        - But the vast majority of indicators captures structure
        - Structural things do matter – water/electricity/fridge
        - But once those are achieved, countries put on more process indicators – but especially MNH not quite getting at the indicators that matter – because results cannot be verified
- In a facility setting – got pretty far in measuring ANC (e.g. mosquito nets) – quantity (4+ visits, 1st trimester start, but intrapartum care – lot of the mortality cannot be verified – do you have oxytocin, was partograph filled in properly – doesn't mean it was done at the right time, forms can be completed retrospectively. We are not doing a good job with postnatal care measurement, very few indicators for newborn care and early newborn is particularly bad.
- Not getting indicators of an enabling environment – people motivated by environment/incentives/supplies – shared results/teamwork + more supportive supervision: but for this cannot find verifiable indicators
- What indicators to include – are they valid in what they seek to measure? (not necessarily just what is verifiable).
    - For example number of new contraceptive users – perfectly valid but useless – no quality component, discontinuation – can perform well but mCPR still low
    - Intrapartum care – very structural – no timing of AMSTL, newborn care - % breastfed within one hour – get from facility (issues), validate by women?

- What does validity mean?
    - Is a metric measuring what we expect it to be measuring?
    - Are the data collected and methods reliable to have confidence in what the indicator says? Includes understanding there are confidence intervals and imprecision
    - Challenges for newborn – small and sick care, all facility-based indicators
    - How do we operationalise these indicators and use them for programming?
    - We also need to consider that RMNCAH space is already crowded with indicators
    - We need to define at what level they are relevant and how they should be used
    - Also do not sacrifice programmatically useful data because they are not perfect.
- Lots of decision on validity of indicators have been based on few studies in limited variety of contexts
- Different concepts of validity
    - In epidemiology/demography: is criterion-based – compared against records/observations as a gold standard
    - Respectful care (no comparison/standard/truth) – depends on observer or respondent
    - Social science – face/content validity, does it respond to change/to other indicators logically
    - Misclassification is important – we want to be as close to truth as possible
- If an indicator is used to measure change, it needs to be fairly valid/precise – if one assumes the extent or error remains the same. However, this is not reasonable and nor has it been studied.
- Understanding validity and going for the best indicator is not overly critical – it is the right way to go for MNH.
    - We need to be careful about adding questions to survey instruments
- The danger is that the indicator is a proxy for a construct – but becomes conflated with the construct to mean that it is the single needed input to influence health outcomes (e.g. % f births with SBA link to maternal and newborn mortality)
- An indicator is a part of a broader picture – can be valid but other elements are needed – for example obsession with SBA (which is ok for normal deliveries) does not capture emergency obstetric care. We don't have a good indicator for this and csection rate is fraught because can be overused.

- Validity depends on what the indicator is for. This needs to be clear and understood; e.g. % of women who receive iron supplementation during pregnancy – is this a proxy for ANC quality or proxy for anaemia?
- Validity gets lost in translation – it can lead to unnecessary scepticism of indicators. The key question is whether the indicator is "good enough for its purpose". For example - oxytocin for PPH – question could be "Were you told you were getting oxytocin?" (then it might capture quality of care in a sense that patient consent was obtained). This is different from "Did you get oxytocin?" – women don't know this. Both approaches are very different from whether the oxytocin actually active (kept in fridge, etc) – that is an entirely different issue altogether (could be assessed through biomarker capture?)
- Another element of validity is gender dynamics – we need to listen to what women are saying – women are not stupid, they might not be told about interventions they are receiving. Applies to measurement of disrespectful care, self-reported conditions.

Question 2. What types of approaches are considered useful in assessing indicator validity?

- There are several approaches which can be taken:
    - Work on cognitive assessment of understanding questions/qualitative work in the process of questionnaire development is very useful.
    - Whether responses reflect accurately the respondent's experience
    - Understand whether coverage measure is correct using sampling approaches on a population level, for example – measure using two methods and compare (not matching individual data but effect on population level coverage – and assess whether confidence intervals overlap).
    - 1:1 match – whether each response matches the truth using a gold standard comparison – this approach is not always feasible and sample size considerations are important. This is often done at facility level; can be a rigorous method but with its challenges (expensive).
- Health system indicators – how to develop a meaningful indicator for a dynamic system? We also need to consider that measuring an indicator changes incentives.
    - Health system is non-linear and complex
    - How to measure responsiveness, autonomy, local control, motivation?
- Hierarchy of methodologies:
    - Comparison to gold standard – best would be a biomarker (such as in contraceptive use – hormone levels)
    - Second best would be an observer – but need to train observers well so additional error is not introduced
    - Triangulation – useful but does not strictly assess validity
        - For example, using medical records or HMIS; can examine consistency and systematic differences across various data sources. For example, % of maternal deaths that have audit conducted – there are issues with data sources for both numerator and denominators.
    - Policy indicators – would need to be creative but definitely worth assessing validity – asking beyond what is on paper (in the law).
- Health systems indicators: Qualitative aspects – intangible issues such as teamwork, leadership, trust, motivation, management, organisational culture - these are latent constructs – measurement of motivation/responsiveness. How do we measure and validate these?
- There is a disciplinary divide – epistemology of measurement and validation – these are loaded words/positivist perspective. Disciplines do not agree on whether there is one objective truth that can be captured. For example, for issues of power/hierarchy, there are differing perspectives.
    - Anthropology/organisational behaviour methods being used to see how people get work accomplished
    - Qualitative language for validity are rigor, robustness, trustworthiness of interpretation. There are tools to assess this based on prolonged engagement/audit trail, how are data interpreted by researcher etc.
- Useful approaches include:
    - How can individual indicators of effective coverage (such as blood pressure measurement in ANC) be correlated with evidence-based standards of care?
    - Understanding the efficiency of measurement (not all indicators, just some that perform well)

    - Assess “validity” of indicator – its reliability as a predictor of care
    - However, some indicators are very vague, need to define and flesh them out – PNC has very few indicators
- Indicators of policy and systems is where there is a tricky balance between advocacy and science
    - How to measure policies – existence of policy or its implementation?
    - Much harder to have standardised indicators in this field – for example, what is an essential list of policies on RMNCAH?
    - Validation is time consuming/sensitive and triangulation is required from various respondents.
- Methods of validation include several steps/stages:
    - Type of interviewers, is question understood?
    - Acceptability of a question
    - Will people report truthfully? Can the respondents know the answer?
    - Is an indicator validated in one place (context) good enough for use elsewhere?
    - Very few indicators currently in use have been tested – but we have to measure something
- There are issues with facility registers and clinical case notes - many interventions are done but not recorded, missingness – this affects on quality of data, liability to errors, performance management
- In observations of care, the Hawthorne effect (of observers) needs to be considered
- Many indicators are very hard to validate e.g., pre-discharge check after childbirth (hard to observe as the time period is long, not always recorded in case notes)
- Divergent approaches to validity in qualitative work:
    - People devise a question + pilot it on a few women – this is almost not done anymore
    - People devise a question and then do cognitive interviews – increasingly predominant in LMICs, but questioning the value – needs to be done well otherwise might worsen validity
    - Do qualitative work first to understand how people talk about issues – then design a question (this is not done frequently, but is the best approach)
- Verification is a method of checking validity, especially in pay-for-performance (or results-based financing) schemes
    - Verification of numbers – there might be lots of over-reporting but not necessarily due to cheating – unclear what was the definition of a “visit”, districts understood indicators differently, when verified the levels went down - so what do they really tell us about progress?
    - Should we verify indicators in all districts? This is very expensive and time consuming (“verification bonanza”)
    - What’s the right balance – make progress (verification is not the core business of pay-for performance programmes)
- Health systems indicators – routine monitoring of indicators requires routine data – financing data comes from National Health Accounts and National Budget reviews. Validity can be assessed through:
    - Data checks – seeing if data exist, whether they make sense, what are the completeness levels,
    - But mainly just use data that are there (no validation, triangulation) – if data extremely poor, then they don’t use at all
- Validation is complex and expensive, and the least thinking going into denominators (for example, among sub-samples who are in need of treatment). For numerators, people have been counting – it’s recording and collating that is the problem. Some of the barriers include:
    - Data literacy
    - Collecting too much data
    - Denominators not available or not clear
- In health systems/financing indicators - what do we need validation for:
    - Need facts and trend over time
    - Improve methodology for future tracking
    - We shouldn’t be too focused on details – might not be worth the effort to track financial resources for SRHR – is the actual number and % correct? If it is, is that enough resources? Is this the right way to show something is underfunded/funded at the right level? (what % of health financing should SRHR get?) if [country X] gives 60% of its budget to the field of MNH, is that enough/little/a lot? How do we interpret these levels?

    - Population-level surveys (household budget surveys) – massive variation in quality of data across countries, national statistics offices do this work
    - If collecting data oneself – i.e. From households on expenditures, can be checked – and enumerators trained etc.
- We also need to validate or assess indicator denominators (who needs particular intervention)
- Is it appropriate to measure certain indicators using facility records (HMIS/DHIS2)? Is it feasible to collect these?
- Measuring maternal recall validity to assess population level
- Some research can be understood as "pre-validation":
    - Understanding of how data are generated, recorded, collated; how measurement devices are maintained; the cultural meaning of certain procedures and of measurements
- Issue with using few facilities to do observations to gather gold standard– tells us about record keeping in those specific facilities (who are under observation), but not generally about facilities
- Indicator work is more than just validation, there are several phases:
    - For example, on a population survey, we first need to develop the question – cognitive interviews with women, do respondents understand the question and does their understanding reflect the intent of the question?
    - Testing the question in the field – understanding questions alone and their positioning in the questionnaire, are the levels seen and associations with other key variables reasonable? How do the levels seen compare to other sources?
    - Assessing how the question functions and comparing to a gold standard – to assess whether statistically valid (sensitivity, specificity)
    - But other methods – cognitive interviewing, field testing – are also important, but can be done well or poorly
    - There is lots of knowledge on these aspects (developing and testing a question), that are not necessarily written up and published
- Comparing two methods neither one of which is the truth (or gold standard) – Bland Altman method –correlation not a good way of comparing two methods, because it doesn't measure how well they agree. Propose a measure "limits of agreement' where differences between measure for each observation are summarised.
- Issues with measuring validity with subgroups when not everyone needs an intervention
    - All you can do is observe who received it and whether that was recorded
    - Doesn't speak to whether that particular child needed the intervention or whether other children who needed it did not receive it. In the second case, the observer needs to intervene, so impossible to record and analyse.
    - Analysing the use of different denominators for this (sometimes called benchmarking) might be a waste of time - like the optimal csection rate conversation
- Validity of indicators capturing rare events
    - Specificity needs to be 100% otherwise greatly inflating prevalence
    - Modelling should have reasonable estimates of sensitivity/specificity – from following up women with and without fistula/outcome to estimate) to use as assumptions/priors
- How much does verification cost – how it can be done (cheaper options such as risk-base, modelling)?
    - Indicators are only as useful as they are reliable
    - Can put in an exit interview – but women won't report abuse at exit from facility, then follow up by community survey (random sample) – how was service on day, their experience, payment for services (whether received for free) – this is counter-verification
    - Verification: extremely labour intensive, but reason why RBF might work – it's accepted that it's expensive but catches cheating, someone is watching/tracking/paying attention to data
- Conducting observations and exit interviews every 6 months – follow up 12-18 months postpartum in community: validity against observation and own exit interview
    - Interest in stability of validity measure (can vary)

Question 3. What is an acceptable level of indicator validity?

- Validation results depend on methodology used – we shouldn't be too excited about being able to differentiate between VALID/NOT VALID indicators – other factors have to be taken into account, for example representativeness of sample. Cut-off for validity is arbitrary (AUC/IF)
- If an indicator is not valid, it should not be used (e.g. pneumonia among children in household surveys)
- An indicator can be valid but not useful – assessment to see if indicator can be reported – some are useful but not validated (e.g. number of ANC visits) – can still tell us something about guideline recommendations
- Is the indicator measurable? – assess implementation research, feasibility, usefulness, usability – specifically at country levels
- Country-level studies give a sense of how an indicator performs in terms of validity, but the question is, combined how many country studies are needed to generalise whether an indicator is valid globally/across countries?
- Impact of findings showing poor validity of indicators:
    - People are not happy to hear such results – lots of reluctance to accept findings about indicators related to delivery (childbirth) care. But some results came of it – the question on the DHS about csections and skin-to-skin contact was changed (DHS open/willing to take comments)
    - Obstetric complications – has been shown not to work but people continue to try. Why? People don't bother to read the literature or new people come who don't know this has been done and rejected.
- EPMM/EWEC/CD2030 – long lists of hundreds of indicators – how can we distinguish between those which are worth measuring and feasible to measure? Methodological work can give some answers, but ultimately decisions are political.
    - We need a summary of which indicators have had scientific scrutiny. Validated or just expert opinion based? This should be recognised even for long established indicators.
- Uptake of validation evidence: Heartened to see 2016 DHS had the changed skin-to-skin question which was validated in 2013 paper in Mozambique
- Many indicators derived from DHS data are not validated – but people think it's a "validated tool" by the virtue of having been used for so long – but this doesn't mean it's measuring all issues well.
- DHS needs a higher level of validation (a higher standard/prevision in order to be considered valid), because it is meant for cross-country comparisons.
- HMIS/facility records which will only be used in that country might be ok with lower standard of validity – but how should the level of rigour differ?
- We know from evidence that we cannot get obstetric complications from women's recall
- Near miss measurement from facilities – woman has to be in a hospital or otherwise she would have died – easier as this is hospital-based but can it be used to measure prevalence in population? Women's interviews in audits – very powerful when brought to tell own story to audit committee. Measurement of near miss incidence – women cannot recall, problems/errors in medical records (trends over time issue as mortality declines, near miss might go up).
- Unmet obstetric need – absolute indications for csections – assuming there is a constant prevalence of these complications – can measure deficit between the csection level and theoretical need. But this doesn't work – assumptions don't hold, medical records can be an issue. It measures routine care but not the mortality element.
- Quality of care – needs observation, what happens in emergencies
- Care quality/content indicators need observation – especially of normal labour management (routine care improvement)
- Why are we interested in validity
    - Reducing the number of indicators and unpacking the health system – coverage with content of care
    - We are using indicators don't relate to health outcomes
    - Need to get indicators that are much more real and away from assumptions
    - All countries who need to report – data understandable and comparable and health move forward in global discussions of what is happening
- Poor validity is not an issue in maternal/newborn health specifically – the question is what is "good enough" for its purpose/use – it's not about that it's perfect, but about how wrong we can afford to be

- Indicator validity is not just black and white - there are degrees of truth – not binary. An indicator might be good in one country but not in another.
- Digesting of validation results: 2013 validation work in Mozambique was available when preparing DHS 7 – changed questions on skin-to-skin to be a two part question which performed better
- Many indicators seem to have face validity, but wouldn't be "valid" if obtained from routine data at the moment – we have to improve routine data, but how focussed should the efforts be – cannot devise a whole long list of indicators and expect all to be measurable
- Many indicators that are suggested for LMICs are not measured for HICs – we should understand why? If they are so important, why don't HICs measure them? – some reasons include: difficult to measure, gaming an issue, uncertain value. But when HICs do have certain indicators (mortality, for example), we should understand how they are used and useful.
- Country level - Governments don't have much of a view on indicator validity – they are running a chaotic system in an information vaccuum for decades, huge gaps in capacity to use indicators
- There is no set valid cut-off point – certain level of judgment is needed – also considering whether gold standard was done with good quality research approaches
- Decisions about indicators has to look at all the evidence – cognitive, field testing, acceptability, validity – all together as a body of evidence.
- Validation results is not a linear process – evidence comes in over time in bits and pieces, it builds up until tipping point is reached. Household survey programmes have to weigh in evidence in 50 countries experience versus 3 new published studies.
- AUC is useful for a diagnostic test that can produce continuous variable output – such as viral load, when changing the cut-off point for diagnosis then translates to different levels of sensitivity and specificity that can be plotted on a ROC and an AUC be calculated.
    - For binary measures (something was done or not), there is no cut-off, and the AUC will be an average of sensitivity and specificity – it doesn't say anything more than sensitivity and specificity do. Not additionally meaningful, the ROC is based on 3 points.
    - Cut off levels for good validity - Cut off of AUC >0.7 – this means average of sensitivity and specificity is 0.7 – so either one of them is terrible or both are poor
- The word "validate" is statistically misleading – statistics can only ever show things are untrue
- Acceptable validity – depends on how much imperfection you are willing to put up with – what purpose is the information for. AUC of 0.8 is pretty poor but that particular indicator might be worth collecting.
- The acceptable level of accuracy depends on what the indicator is for – a certain level of validity might be good for one purpose and poor for another
- We need to define, measure, and know the purpose of the indicator, and then ask – how wrong can we afford to be? What is it going to be used for? That will also answer whether an indicator needs to be valid across cultures/countries.
- Does validity mean technical validity or "usability" in a broader sense? What about indicators where there is no truth? For example, self-reported morbidity (again what matters is what it is used for)
- If an indicator is a proxy – there are two concerns – can it be measured reliably, and is it a good measure for the real thing/construct?
- We also need to understand sensitivities of measurement – for example, water running in a health facility one day but not next – how to capture volatility/stability in cross-sectional facility based assessments – valid at that point (applies to fuel, electricity, functioning ambulance)
- What's next after negative validation results:
    - Takes time and repeated discussions - continuity of measurement across time is hard to break even though there is scepticism about existing indicators
    - Most people who use data are not "measurement" people – they want some – any – estimates
    - Within countries – less interested in validation work (hasn't really been disseminated) – surveys such as the DHS is the "bible" and is not disputed
    - Translating to country level is tricky – if involves a whole new measurement approach, it is complicated to explain
- How should qualitative work impact cross-country surveys? Difficult question – if change improves validity only a bit, might not be worth changing and discontinuing from time series. But if substantial improvement, might be worth it.
- When does re-validation of an indicator need to happen? Changes in a country that would change the way women report:
    - Socio-economic circumstances (e.g. increase in education)

    - Prevalence of phenomenon being measured
    - Caregivers/care recipients are not told about the condition/treatment being received
    - But how to monitor these changes to trigger re-assessment of validity?
- What happens with results of validation:
    - Not much happens
    - With maternal morbidity, these questions were eliminated from the DHS – this was a success, because when data are collected, people will use them, sometimes uncritically, and also implicitly means DHS think these data are good enough to use.
    - There is generally a lack of communication about validation results

Question 4. What gaps exist in indicator validation work?

- Rigorous methods of assessing indicator validity (gold standard) should continue
- Assessment of which indicators matter – well measure the delivery and quality of interventions (not just coverage) – this is more difficult to capture
- Develop a feasible effective coverage measure – in the absence of gold standard validation
    - For example: whether indicators included in routine facility reporting and used for program improvement – this goes beyond strict validation but can be useful
- Valid indicators can be reportable through HMIS – platform for data collection matters – different data sources and platforms should be continually assessed, including routine facility data to asses population level indicators
    - Inspiration: adult NCD, environmental health, AIDS, malaria
- Work on patient-centered care experience/perspective/satisfaction – we don't know how to measure this
- Quality of care – what we are using now are indicators of comprehensiveness, not of actual quality. What's needed are audits/observations – that is expensive.
- Recent efforts to simplify obstetric care (guidelines, checklists, task-shifting) are not saving lives
- Household surveys:
    - More qualitative work on how women understand questions – insights from [country name] where women talk about a "fog" about delivery/postpartum period; questioning what "immediate" means. Women could give suggestions about how to better phrase questions
    - SBA indicator basic issue is that women don't report the way we understand the indicator – but we are stuck with it until 2030 (end of SDGs). FBD would be better and a good proxy, but we also need decent data on content of care – and need to work to determine which of these are valid/meaningful proxies for quality of care (blood pressure measured?)
    - ANC/PNC – better data on content of care needed – can be done by improving women's surveys, but also by health facility records and HMIS – linking across data sources and understanding facility capabilities
- How good are HMIS/routine data sources on routine/essential care? Also, logistical issues in medical notes where transfer occurs from delivery to postnatal ward. There is pressure with SDGs to measure through HMIS/routine data – so this would be along with SDG priorities.
- Context/quality of delivery/postnatal care indicators – measurement lacking (e.g. chlorhexidine needs to be incorporated into HMIS – but what is the quality of the data? From facilities? For home births?)
- Clinical care evidence is continually changing – new indicators are needed and what is the quality of those data? E.g. tranexamic acid for postpartum haemorrhage – how can it be tracked routinely? HMIS needs to be updated.
- Likewise, some indicators need to be dropped, but there is lots of reluctance to do this.
- Gaps in validation of indicators of:
    - Quality of care
    - Respectful care
    - Empowerment of women and health workers
- Understanding what countries use to make decisions and do we have the right indicators for them.
- Country representation could be better – country needs need to be translated to global discussions (each country has a different need)

- Things measured in Global Strategy and SDGs are useful to countries in resource allocation/priorities – global communication also has a role to play in inviting countries and asking for input
    - Choose a UN agency to help with this effort because can engage countries – WHO endorsement/guidelines is useful
    - Countries want to have an idea of what others are doing
- Gaps include:
    - Effective coverage – way of capturing health systems. ANC content of care – how valid based on women's responses? Are socio-economic inequalities seen in reports of care content real or an artefact of understanding/recall?
    - Stillbirth/abortion – ANC not asked at all for these outcomes
    - Need info from facility/health system – what is a gold standard for children's examination
    - Postnatal care – what should be the content and which is the best data source to use to measure it?
    - Are facility surveys a valid data source – transitioning to routine data, but in the meantime can we simplify and improve facility surveys to capture quality of care? Optimising data collection
- Survey data on household expenditures – strengthen national stats office capacity – ensure a minimal skill set especially with data cleaning, consistency, quality.
- Domestic funding tracking for MNH – variation in data systems across countries – increase in use of electronic systems might help (HMIS for financing) – at least get data in specific categories at the time it is being used – might not be correct but over time can improve system and it's a starting point for further work.
- So few questions on the DHS have been validated – but when they are looked at through validation research and found to perform poorly, DHS is happy to remove them. Countries want certain things to be measured (for time trends for instance) and have influence on items in questionnaire.
- Many questions don't perform well in validation studies but retained in DHS – remaining questions are:
    - We assumed women can recall these things (events, interventions), but they cannot
    - Can we get these indicators from women at all? How? Some items can be collected form facilities/registries, but others cannot. Increased burden on data collection in facilities
    - Do we have to do observations of care? This is expensive – introduces bias from high volume sites and issues with management of complications (rare events)
    - Data users need to prioritise indicators in terms of which to be validated first
    - Some indicators have issues (SBA) – questionnaire design can only improve validity so far – if women don't know the answer.
- Indicator validity for cross-country versus within-country use
    - It's a balancing act between usefulness (direct implication on program or policy) and how well it can be measured; how sensitive is it to change over time?
    - Often programmatic needs take priority over methodological robustness and relies on subjective assessment
    - Indicators within a survey compete with one another – for inclusion and attention
- Quality of care work – there is momentum and its important work but messages are not clear for countries: service readiness measures whether basic pre-requisites are in place. Facility quality improvement (hundreds of indicators) – on what basis? Confusing and distracting – MoNITOR could help reduce confusion. What is "service readiness"? There is no standard definition.
- Countries focus on mortality estimates, not methods or confidence intervals
- Gaps in indicator validation of maternal/newborn indicators:
    - Focus on intrapartum/postpartum
    - Gap in ANC, perhaps because not life saving? Although some work going on now
    - Care seeking for PNC visits – what is a "check"?
- Gaps and new developments: Technology might improve ability to measure aspects of care, but won't solve big problems quickly if the "data architecture" and system are broken – (door knob without a house)
- There are a number of proposed indicators that won't work – for example documenting newborn resuscitation (no idea which babies need it) – only valid if information on condition of the baby is available there – but hard to get/subjective/gaming
    - If it's a good indicator, why aren't HICs measuring it
    - HICs measuring far less than what LMICs are expected to

- Priority setting in indicators – setting them to context (limitations on measurement, for example in facilities). We need to sort out measurement of most important indicator, only then move onto measuring others.
- There are not a lot of resources for validation – development of questions, but building up. Any new question now needs to go through a battery of tests before even brought to countries to consider.
- Generalisability – validation studies done in a few places, but that is some of the best available evidence – it is not the only evidence, we can triangulate from other sources
- Validity studies using gold standard are all from clinical (facility-based) populations – that is a problem
- Cognitive interviews that were done early on in the development of these questions are not published
- Many questions/indicators being validated aren't good questions to start with – didn't need a validation study to tell us that it won't work
- A lot of inherited questions from 20 years ago – retroactively applying new validation methods to old questions
- Respectful care – no tools to capture this with a few questions
- Newborn indicators:
    - Quality of contacts (content of care, postnatal care)
    - Newborn morbidity and development
- Maternal indicators:
    - Respectful maternity care – validation for a more client-centered perspective, culturally diverse contexts
    - Work on measurement improvement is broad, different topics are at different stages in the pathway of measurement
- How do we make information on indicators (including validity status – yes, no etc) accessible in an easy way? (in indicator reference sheets?)
- Validity of routine data – should be available and used (we know a lot on population surveys), including value/assign/train staff who are responsible for data entry in facilities. Any intervention delivered by medical staff should be in medical records, not asked of the patient (self-report)
    - We need to use routine data quality tools, conduct period assessments, simplify processes
- It's WHO's job to advocate for a better understanding of validation - MoNITOR as advisory body should make recommendations of indicators taking into account validity
- We need really forceful data people – without easy guidance it's the programme people driving selection/development of indicators
- Gaps in validation work on indicators of:
    - Measurement of abortion – more work needed
    - Morbidity – measurement of maternal health/wellbeing in a broad sense
- The future is in facility records/HMIS – we need a standardised set of items to be captured routinely – as more women now in contact with health system (ANC, SBA/facility delivery).
    - These indicators should not rely on population surveys (e.g. intrapartum care) – and DHS can focus more on perceptions and subjective indicators from women that cannot be obtained in health facilities or by health workers.
- Epidemiology of maternal health is changing and we need to work on developing and validating indicators that capture chronic conditions/obesity and care received throughout pregnancy-delivery-postpartum over time, integration of maternal health into primary care in the postpartum period (new indicators, priorities)
- Gaps in implementation issues with questions: how is a question applied in the field by enumerators – especially for those including a measure of time, which enumerators often need to find a way to explain to women
- Validation studies tend to focus on countries where research is relatively easy – we have no studies conducted in large parts of Africa
- Gaps in work on facility records as sources of data for indicators. This has an added benefit of potentially improving data quality if the data is used.
- There is a fairly small group of people involved in validation – focusing on few priorities. One area left behind is maternal morbidity – what can we ask women? Which morbidities to prioritise?

Question 5. What are the recommendations for addressing these gaps?

- Keep track of groups working on validation and indicators – coordinate/keep everyone informed
    - Could convene a joint meeting to avoid duplication
    - A lot of this work is on global level – are there any country researchers and how do we bring them in?
- HMIS indicators – how good do they need to be? Are they part of a system where they are used? Is feedback given and improvement initiated?
- What is the purpose of the indicator? Comparing across countries or action within country? – are we engaging countries in this discussion – what do they want and use?
- Challenges: New guidelines and interventions–what is the best way to track these methods? DHIS2 – potential game changer in Africa/Asia, work on new modules and roll them out
- Paper on mapping of indicators was very useful - who is doing what. It also pointed out challenges with indicators – recommend better aspirational indicators going forward.
- Validation synthesis – shut down some of the noise – so many different indicators needed for different users? Aligning at global level help in the long-run to improve measurement, reduce frustration
- Reduce burden on health workers collecting data once, harmonising and efficiency for reporting the same thing, cutting down on one-off data collection, different data for different projects
    - This process can shift power to countries – push back at donors to say here is a list of indicators we can track using HMIS – empowers governments to own their own data – nobody can process so many indicators as exist now
- We need to be cognisant that it takes time for a questionnaire to be developed, collected and data to be made available – indicators need to be fairly constant otherwise data available is always outdated/unsuitable and keeps changing (not comparable over time)
- There isn't as much demand for maternal/newborn measurement from within-countries as there is from the global MNH community – requires more advocacy? Countries need to see a need for MNH indicators. The field is underappreciated/underfunded on the whole, MNC is never the "new sexy thing", with some exceptions.
- How can demand for good quality indicators be generated within countries?
    - Global community needs to be aware of survey dates and work directly with MOH and larger committee on inputs to questionnaires, also involve other ministries, UNICEF/UNFPA, local NGOs.
- Validity of MNH indicators compared to other global health fields (malaria, HIV) – for example malaria/HIV get lots more attention and these fields have some advantage with better measurement (HIV biomarkers, observing nets in households) which have better face validity and recall bias is minimized. In general, there is more introspection in MNH – more interest in validity than other fields, but this is not to the detriment in the long term – it is very powerful and can protect the questions that perform well from being cut from surveys/data collection instruments.
- We have had enough of certain indicators – they are not needed anymore – let's not collect them as often (tetanus toxoid, ANC) – let's focus on other indicators – get a shortlist and target certain countries
- Explore different modes of coming up with data
- Researchers focused on Africa and Asia – but this is not global, what are data needs on other continents and what other data sources can be used, like vital registration and facility records
- Countries should be pushing this agenda, Statistical Offices busy trying to provide indicators for SDFs and want to measure priority indicators – as many on one instrument as possible, but this means surveys are getting too large
- Prioritising fewer indicators will result in improvement in quality of the remaining data collected, and lessens burden on indicators and provides more time to improve health
- Build on what exists – effort on mapping, leverage on work done already, prioritise – being able to show results + put forward valid metrics at pace, especially newborn survival issue which is pressing
- When validation results from research become available, give guidance to countries on implications ASAP.
- MoNITOR advises WHO to reflect on urgency to include new metrics in programs
    - How best to support countries to take up best practices – what is it that facilitates uptake of indicators? Learn from other fields in health
    - Provide guidance to technical experts and advice to high-level policy-makers – different types of data/evidence

- Deliver clear guidance to countries
- Be a clearing house of resources – that are available to researchers
- Prioritise agenda for indicators under development (gaps)
- Don’t get bogged down managing partners and networks, there are existing technical working groups – don’t duplicate, have own defined agenda, define boundaries of involvement

Appendix 1. Semi-structured interview guide

Part 1: General – understanding key themes and issues in indicator validation

Introductions

Describe the purpose of interview

Provide scope of indicators that are of interest within **maternal/newborn health** (current/aspirational; survey/facility/ DHIS/policy; maternal/newborn; coverage/content).

Questions

A. Extent of engagement in validation work and landscape of validation work
- What kind of work relevant to validation of maternal/newborn indicators is your organisation currently doing, if any?
- Are you conducting, planning to conduct or have you conducted any validation studies? (if yes – also administer Part 2)
- Have you commissioned any validation work or thought about using data collected for other purposes to validate any indicators?
- Who are other stakeholders (organisations, researchers) conducting validation that you know?
- What do you think is the motivation behind work on validation?

B. What does validation mean?
- Is validation of indicators relevant to your work? If so, how?
- In your view, what types of research can be considered "validation"?
- What do you understand by the concept of "validity"?
- What is considered to be good versus poor validity?
- What is the perceived value of doing validation and where does funding for validation come from?
- What is the generalisability of validation findings (over time, across contexts and settings, data sources) How long is an indicator validated for (i.e., what is the need for re-validation)?

C. Recommendations of future validation work and uptake of findings
- What is the global and local appetite for validation and its results? Are validation results "accepted" and taken up?
- How are national governments and regional-level actors participating on validation work?
- Do you perceive a gap in which indicators are/aren't being validated and why?
- What other gaps in work on indicator validation do you see?

Wrap up

Is there anything else you'd like to share on this topic?

May I get in touch with you with any questions?

Thank you for your time. Agree on follow up if any materials were to be sent.

Part 2: For data collection on validation work (interviews or from written materials, such as reports, published literature and pre-publication drafts)

Information to be collected for each project, from the respondent or by email (study protocols, study findings, etc; published and unpublished)

Country(ies)

Setting - implementation, M&E, stand-alone work/opportunistic, facilities, households

Funding source

Timeline and expected results

Type of validation applied – methodology (sample size, gold standard, statistical indicators used)

What is considered to be good v poor validity and based on what criteria - statistical performance, variability?

Can you share the study protocol or any other materials to help understand the details of the work being conducted, including data collection and data analysis strategies?

Additionally, for projects that have results:

Can you share results with me (unpublished findings, study protocols, check that I identified all published materials)

What variability in validity has been found? What are the reasons for this variability?

How were results disseminated and received?

Anything you would have done differently?