unicef

World Health Organization

Meeting report

Report

WHO/UNICEF JMP Task Force on Methods

WHO / UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP)

Contents

Executive Summary

The WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP) has been producing estimates of global coverage and trends since 1990 and is officially mandated to monitor progress towards the Millennium Development Goal MDG) target 7c for water and sanitation. With the MDG period coming to an end in 2015, the JMP has facilitated international consultations on post-2015 targets and indicators and is also reviewing the current JMP method for deriving estimates of coverage.

From 2-4 December 2014, a Technical Task Force meeting of the WHO/UNICEF Joint Monitoring Programme (JMP) for Water Supply and Sanitation was held in New York. A group of 18 technical experts, 6 members of the JMP, were in attendance.

As originally set forth in the meeting Concept Note, the objectives of this meeting were to:

1. review evidence for non-linear patterns in the JMP data and to assess the sensitivity of JMP estimates to data updates.
2. evaluate the suitability of alternative statistical methods for estimating the population using drinking water and sanitation at national level in future
3. establish guiding principles for constructing baseline estimates from partial data sets for hand washing with soap, water quality and WASH in schools and health facilities

Through a combination of presentations and plenary discussions reflecting on background papers and a series of questions prepared in advance of the meeting, the taskforce found:

Evidence of non-linear patterns. There is evidence of non-linearity in the JMP data and at least a few different patterns but many countries still have too few data to detect non-linear patterns. Of those countries with over 10 data points, 15-25% have non-linear patterns depending on the indicator.

Criteria for model selection. Criteria were proposed for evaluating alternative models and it was agreed that some degree of subjective assessment is needed to establish whether the predicted trends are “credible” based on sector knowledge. It is acceptable to use different models based on availability of data as long as they are selected in an objective and transparent manner and the number of models is limited to enable comparison.

Alternatives to linear regression. Linear regression was compared to piecewise, logit and quadratic regression as well as the use of generalized additive models (GAM). More flexible ‘semi-parametric’ approaches can better account for curvature. A more detailed side by side comparison including piecewise linear regression and well as smoothing methods (B-spline and Gaussian process regression. Splines, lowess/loess) was recommended.

Generating global and regional estimates with scarce data. The JMP only reports regional estimates if at least 50% of the population are represented. This rule should be retained and the JMP should avoid reporting national estimates for countries with no data. Covariates should also be avoided where data allow. Exploratory “modelled” information should be clearly separated in reporting. A decision tree can be used by the JMP to select an appropriate approach depending on data availability.

Data sources and their quality. Data quality is just as important as the choice of model**.** Primary concerns relate to whether data are representative of the target population and from a reliable source. The JMP should clearly record the reason for excluding any survey or census in the country files and should formalize and document the rules governing data adjustments. In future the JMP should report some measure of uncertainty and consider assessing data quality.

Approaches to integrating and combining data. The JMP will need to combine data from multiple sources in order to report on new indicators including "safely managed" drinking water and sanitation. The JMP should initially report the new components separately and explore approaches for a composite indicator or index in future.

Monitoring inequalities. JMP should first consolidate analysis of disparities between urban and rural and wealth quintiles. Given the difficulties in defining informal urban settlements the JMP should focus on reporting water and sanitation coverage amongst the urban poor. Using a poverty line, income or expenditure data would allow regional and global aggregation of wealth-related inequalities. Census data could be used to complement household survey data for monitoring small population groups and geographic areas.

Projections and rates of change. Projection rules would need to be revisited if a new model is introduced. Continuity with the MDG period should be considered. Baselines may become less important for the Sustainable Development Goals (SDGs) and the JMP will need to select an appropriate metric for rate of progress.

Introduction

During December 2 – 4, 2014, a Technical Task Force meeting of the WHO/UNICEF Joint Monitoring Programme (JMP) for Water Supply and Sanitation was held in New York, USA. A group of 18 technical experts, 6 members of the JMP were in attendance. A list of meeting participants is included in Annex 1. The Task Force was convened to consider the methods currently used by the JMP and alternative approaches which could be used by the JMP in the post-2015 Sustainable Development Goal era.

Background

The WHO/UNICEF JMP produces yearly national, regional and global estimates of population using improved drinking water sources and improved sanitation facilities. Since 2000, the JMP has been using data from National Statistics Offices, collected by national censuses and nationally representative household surveys. In countries where such data are no longer collected through censuses or household surveys, the JMP uses data from regulatory authorities or administrative records compiled by national authorities.

The JMP does not merely report on the latest survey findings but instead publishes model estimates using simple linear regression. Linear regression averages small differences in coverage between surveys, provides estimates for years in which no survey data are available, and is relatively easy to explain to policy makers and practitioners responsible for water and sanitation service delivery. However, with coverage trends that are depicted linearly, rapid changes are not well captured and it is less suitable at the low and high coverage levels where progress may no longer be linear.

Over the years the JMP has invested a great deal in the promotion of a standard measure of access across the main international household surveys, national censuses and other nationally representative survey instruments. This has led to a much greater comparability of data across surveys and between countries. The increasing number of new household surveys and improved accessibility of older survey data has also enabled the JMP to increase the median number of surveys per country on file from 2.3 in 2006 to 10 in 2014. With the MDG period coming to an end in 2015, and with more data now available, WHO and UNICEF would like to review the JMP's current method of deriving coverage estimates.

Looking ahead to the Post-2015 Development Agenda

The review must also take account of emerging priorities for national and global monitoring post-2015. Over the past few years the JMP has supported a series of international consultations resulting in technical recommendations for enhanced monitoring of drinking water, sanitation and hygiene within the post-2015 development agenda. The proposals call for an expanded set of indicators with additional qualifiers for access, such as water quality, but such data are currently only available for a small number of countries. The JMP is seeking methodological advice on how to incorporate additional qualifiers into its current dataset. Some modeling has been done to adjust the JMP dataset for compliance with drinking water guidelines and feedback is sought on the suitability of this and similar approaches for a potential new baseline for post-2015 monitoring.

The post-2015 technical consultation also recommended that the JMP report on the progressive elimination of inequalities among disadvantaged groups. We would like feedback from the Task

Force on the suitability of proposed measures and advice on alternative ways to assessing progressive elimination of inequalities.

Recent calls for a post-2015 'Data Revolution' have highlighted the need to further strengthen national level monitoring systems and increasingly draw on a variety of national level data for global monitoring. The survey and census data currently used by the JMP are generally produced by National Statistics Offices (NSO) and could potentially be complemented with other sources of data available at national level. The proposed expanded set of indicators for post-2015 monitoring will likely require going beyond household surveys and the JMP would like advice from the Task Force on the use of data from administrative, regulatory and other sources and the possibility of combining these with existing survey data for the purposes of global monitoring in the coming decades.

Objectives

As originally set forth in the meeting Concept Note, the objectives of this meeting were to:

4. review evidence for non-linear patterns in the JMP data (e.g. saturation, acceleration and stagnation) and to assess the sensitivity of JMP estimates to data updates.

5. evaluate the suitability of alternative statistical methods for estimating the population using drinking water and sanitation at national level in future, against the following criteria
    a. accuracy of estimates and projections
    b. consistency in estimating service levels (e.g. open defecation, piped on premises)
    c. consistency in estimating disparities in access (e.g. wealth quintiles, urban and rural)
    d. applicability for all countries and over time
    e. replicability by non-specialists
    f. ease of communication to non-specialists and decision makers

6. establish guiding principles for constructing baseline estimates from partial data sets for hand washing with soap, water quality and WASH in schools and health facilities.

Proceedings

Opening plenary

Ms Holly Newby (UNICEF) opened the meeting by emphasising the importance of the JMP's mandate and noting that global monitoring of water and sanitation is often cited as a good example by other sectors. The culmination of the Millennium Development Goals (MDGs) and beginning of the Sustainable Development Goals (SDGs) provides a strategic moment and an opportunity to review and make better sense of the available data. The opening sesson continued with a brief introduction by each task force participant and presentations by the chairs.

Mr Rifat Hossain (WHO) reflected on the expectations of the taskforce meeting. The JMP seeks guidance not on prescriptions from the taskforce and wishes to establish whether new methods are warranted and if so what steps should be taken. It is unlikely that one model will fit all of the requirements for the different data the JMP will be reporting on post-2015. Rifat emphasised that this is not simply a curve fitting exercise but that the aim of the JMP is to capture the reality and nuances on the ground. As such if a set of countries has very different patterns the JMP's estimates should capture these. Describing one of the background paper (Wolf et al. 2013) he noted that a conscious decision was made not to use covariates but this is something the taskforce could consider. More broadly, the taskforce should bear in mind the longer term developments including reduced distinctions between "developed" and "developing" countries, and try to predict what should be done over the next 20 to 30 years.

Mr Tom Slaymaker (UNICEF) outlined the agenda (Annex 2) and objectives (see above) of the meeting in addition to explaining the criteria for selecting participants. The JMP sought experts from outside the WASH sector, experts in WASH sector monitoring and finally frequent users of the JMP data, those with strong interest in any potential changes in the methods. The aim was to ensure a balance between technical aspects of the discussion and concern about how the data are used to inform decisions (regional and national level as well as global). Mr Slaymaker described the draft evaluation criteria and the background papers. The meeting format was based around six main sessions, each focusing on one topic followed a review of recommendations and next steps by on the third day. The findings from each session are described below.

Reflecting on the meeting objectives, participants decided to keep in mind the following through the proceedings:

- The shifting baseline during the MDGS has been difficult to communicate but this may not be as important for the SDGs since the targets are expected to be universal and/or absolute
- The need for a clear articulation of multiple purposes for JMP monitoring and the criteria for selecting a model. The JMP is not simply after the best curve that goes through the data.
- Post-2015 the JMP will initially be dealing with a rich dataset of existing indicators (basic water and basic sanitation) as well as relatively scarce data for new indicators (e.g. water quality)
- How to manage a change to a new method while ensuring historical continuity; this is partly a question of communication and can be addressed once we know what change will be made.

Methodological developments in other sectors

Ms Holly Newby described developments in selected sectors, with a view to providing insights of particular relevance to the JMP. The following were highlighted:

Child mortality. Child mortality estimates are produced by the inter-agency Group for Child Mortality Estimation (IGME). In comparison to the JMP, there is a lot of data even for relatively data poor countries. Under-five mortality rate is a very political indicator and there have at times been controversies over the UN IGME estimates. Modeling has become a lot more complicated in the last five years and we are now at the point where during missions to countries even explaining the models to the technical people is challenging. The goal of the missions is to have a mutual understanding of different approaches. What is often of most importance is what data are being used (are all of the country's data being used?) and how close estimates are to latest available estimate used by the government. An area that is being explored is the assessment of data quality parameters within a single data set and it is proving very challenging to quantify quality or account for it in the model. The data visualization provided on *CMEinfo* (http://www.childmortality.org) is great for explaining to non-specialists why we do what we do and what the problems are. For example the Nigerian parliamentarians were able to understand the problem of coming up with a single estimate from a huge spread of nationally produced data even if they did not understand how the model was generated. [Note running the models takes a very long time and steps are being taken to look at the use of high performance computing.]

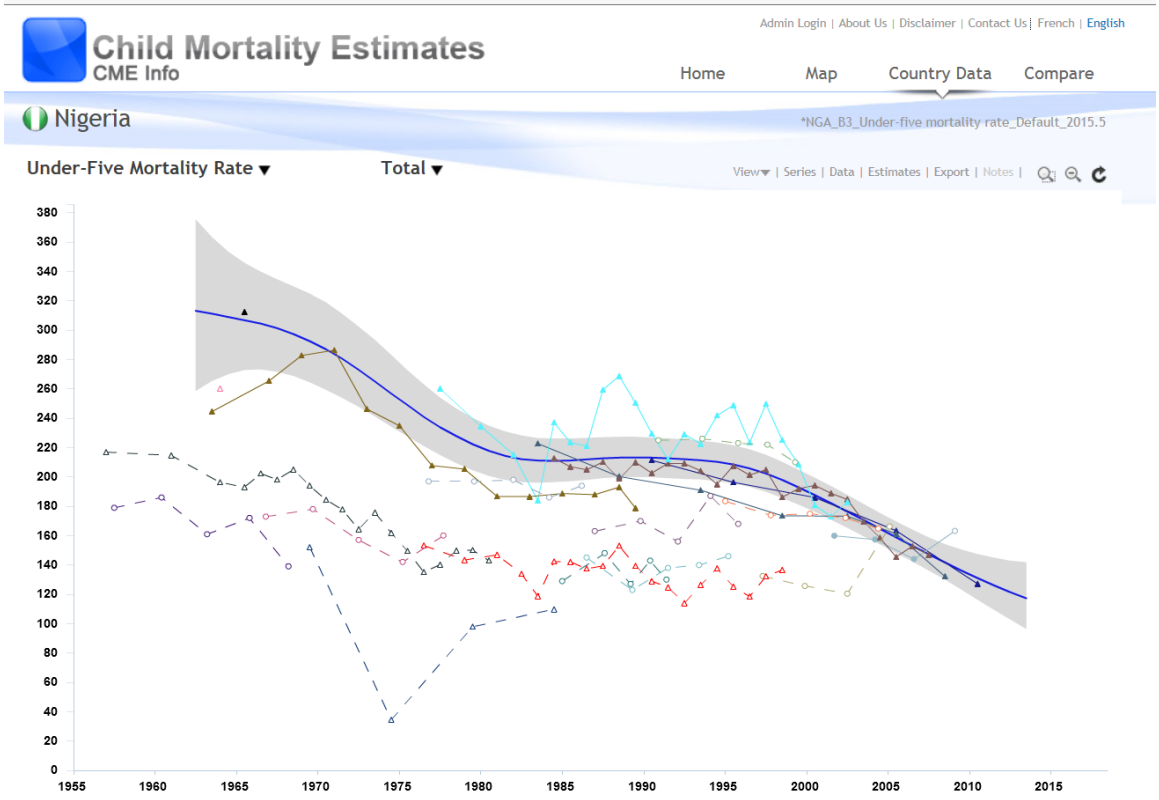


Figure 1: Snapshot for Nigeria under five mortality from childmortality.org (CMEinfo)

Maternal mortality. The data for maternal mortality are sparse and inconsistent. Civil registration data for 97 countries are incomplete and other sources such as household surveys need to be used. The sisterhood approach is very problematic (7 to 9 year recall), with rare events meaning that the confidence bands are wide. There are 20 countries for which no nationally representative data have ever been available and others where the last available data are 15 or more years old. It is not possible to do anything sub-nationally except where household surveys are very large (India, for example). The approach used is a hierarchical regression model with covariates (GDP, general fertility rate, skilled attendant at birth). As the estimates have been particularly controversial due in part to the very poor underlying evidence base as well as the imprecision of the indicator itself, in SOWC UNICEF also reports the country's own "reported" data point alongside the UN interagency "adjusted" figure. Adjustment is made based on known underestimation. For online graphs a decision was taken not to shown original data points – (should unadjusted or adjusted be shown?). There are issues with outliers --for example Indonesia -- and it is difficult to justify a different approach for just one country. The entire approach is under review.

Immunization. Uses data from different sources: household surveys and admin reported. Developments include a qualitative assessment of "grade of confidence" and the use of computational logic to determine which data point to use for the estimates. Although the inter-agency group has cautioned that the estimates should not be used for funding decisions this advice is not always heeded. The uncertainty around the estimates of immunization coverage estimates has not always been fully appreciated, for example the attention that was given to a "decline" of the DPT3 coverage estimate from 84 to 83%. [A further presentation was given on the assessment of grades of confidence and use of computational logic – see *Data sources and their quality.*]

Malnutrition. The joint monitoring initiative between WHO/UNICEF/World Bank is relatively recent. The nutrition community currently generates regional and global estimates on children's anthropometry indicators but does not report country-level estimates. There are plans to develop approaches for national and sub-national (geographic areas) as well as other background variables such as wealth quintiles and urban-rural residence. A key challenge is how to assess annual rates of reduction when the baseline is in 2012 for the WHA targets – should a 5, 10 or even 15 year period be used in the calculations?

Generally speaking, there has been a move towards modeled estimates of various indicators and increasingly the use of complex modeling approaches. An appropriate balance needs to be had between model sophistication on the one hand and confusing audiences and making more than we should on the basis of scarce and/or poor quality data on the other.

In the following discussion it was noted that a key difference between WASH and several of the examples given above lies in the lack of an absolute truth – for WASH defining and setting a benchmark is really important. Holly Newby suggested that may be the case but that there are other examples where definitions are equally important (e.g. skilled attendant, educational attainment etc.).

Data revolution

Dr Robert Chen, member of UNSG's Independent Expert Advisory Group on data revolution and professor at the Earth Institute at Columbia working at the Center for International Earth Science

Information Network (CIESIN) was invited to give a guest talk. Dr Chen highlighted water-related data work at CIESIN and discussed his role on the advisory group on data revolution. Selected examples of water-related work included: the climate vulnerability index; transboundary water assessment programme (TWAP); human water stress (based on freshwater availability and population density); a wastewater treatment map; and the environmental performance index in collaboration with Yale. Forthcoming working includes collaboration with the Blacksmith institute on hazardous waste, assessing air quality using satellite data and utilizing data from an upcoming NASA mission that will assess soil moisture levels.

The Independent expert advisory group on data revolution stemmed from a recommendation by SG High Level Panel. Its report "*A world that counts: mobilising the data revolution for sustainable development"* was released in Nov 2014 shortly before the taskforce meeting. The report underscores the need to focus on strengthening national capacity. Key questions include how to integrate traditional with new sources of information and get to the point when governments can use this information for decision making. Dr Chen noted that it is not clear how data are integrated in the US - The chief statistician and NASA may not be in good communication. Whilst new social science data could be substantially cheaper, for example Orange micro can look at what mobile phones can tell you about development in a region, it may tell you little about the SDG targets.

WASH Post-2015

Tom Slaymaker provided participants with an overview of the JMP international consultation on post-2015 WASH targets and indicators and an update on the latest in terms of the political process to negotiate and agree the SDG framework, including the Open Working Group targets.

The shared vision is universal access to water, sanitation and hygiene (WASH). Addressing the unfinished business and shortcomings of MDGs, proposals include hygiene, equality, going beyond the household (schools, health facilities) and achieving sustainable services. Sector proposals are shown in Box 1 below.

Box 1: Sector proposals for post-2015 WASH targets

By 2030 to:

- eliminate open defecation;
- achieve universal access to basic drinking water, sanitation and hygiene for households, schools and health facilities;
- halve the proportion of the population without access at home to safely managed drinking water and sanitation services; and
- progressively eliminate inequalities in access.

Open defecation, access to basic WASH at home and monitoring of inequalities in access are possible to monitoring globally based on available information and current JMP monitoring practice. Monitoring WASH in schools and health facilities will require new data collection mechanisms and safely-managed services will also need to draw on information that cannot be collected solely from household surveys (e.g. verified risk management plans and fecal sludge management).

Reflecting on the proposals, members of the taskforce noted that (i) countries may in fact decide on their individual targets which would complicate regional assessments (ii) it is likely that not all of these indicators will have an SDG target (iii) the JMP has always monitored more indicators than required by the MDGs, for example it continued to monitor sanitation before this was included in the MDGs and it has reported water and sanitation ladders.

Plenary discussions on specific questions

On each of the first two days of the meeting, findings and recommendations were determined through consensus after a plenary discussion in which questions (shown in blue) from the meeting Concept Note were shown to members of the taskforce. The summary of the discussion and related findings and recommendations (shown in green) are presented in the sections below. On the final day, these topics were revisited and the taskforce discussed additional topics identified as particularly important. Findings and recommendations are intended to convey the consensus view of the Task Force.

Session 1: Evidence of non-linear patterns

Questions posed to taskforce

What evidence is there for non-linear trends in coverage estimates? What patterns are observed?

Using the current JMP method, how much variation in baseline and current coverage estimates can be expected when new datapoints are added during regular updates?

Non-linear patterns in the JMP data

Mr James Fuller presented the key findings on non-linearity of JMP data to task force members, summarizing the proportion of countries with evidence of non-linear trends based on the background paper prepared for the taskforce (Fuller et al. 2014). He noted that the assessment was subjective and based on the inspection of graphs from all countries with at least 10 data points (and at least two before 2000). A variety of linear patterns were observed: linear growth, linear decline, 100% coverage and no change. Non-linear patterns observed in the JMP data included:

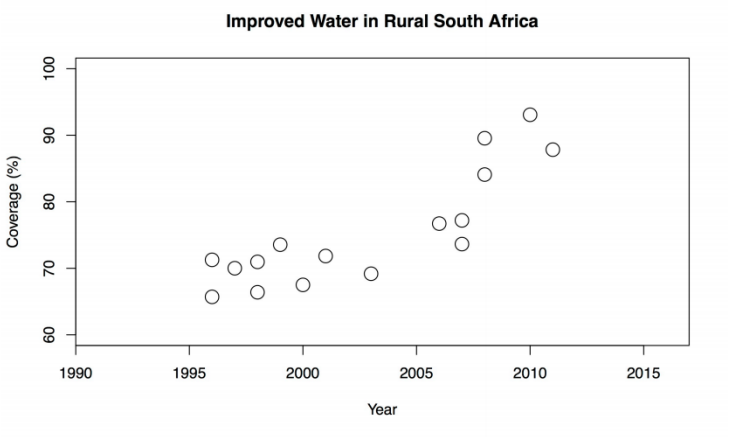
Saturation: Progress slows as coverage approaches 100%.

Acceleration: Substantial progress after a period of no, little or less progress.

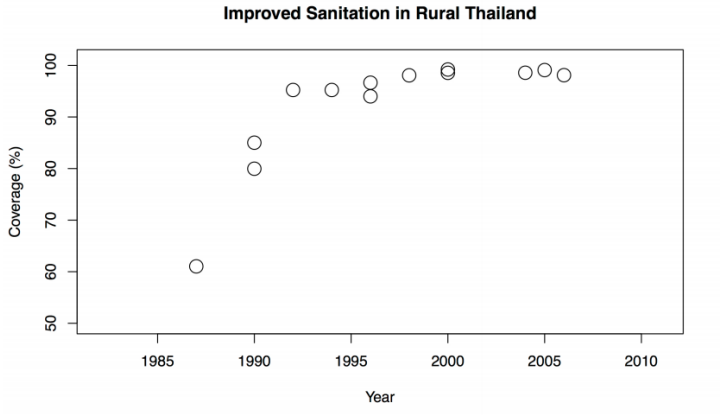
Deceleration: Progress is stalling or slowing down but coverage is below 100%.

All three of these patterns can either be gradual or abrupt and they have implications for whether linear regression will tend to under or over-estimate coverage and rates of progress. Examples of each of these patterns are shown in Figure 2. Of those countries with sufficient data (>10 data points with at least two before 2000), 15-25% were found to show evidence of non-linearity depending on the indicator (Table 1).

Acceleration Example



Saturation Example



Deceleration Example

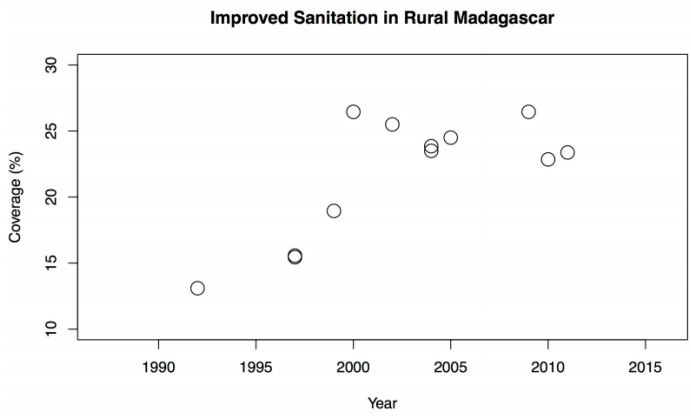


Figure 2. Non-linear patterns observed in the JMP data and illustration of sensitivity to data updates (Fuller et al. 2014, in draft)

Table 1: Proportion of countries showing evidence of non-linear trends (Source: Fuller et al. 2014, in draft)

Indicator	**Countries*** (n)	**Proportion showing non-linearity** (%)
Improved sanitation (Rural)	48	25
Improved sanitation (Urban)	48	20.8
Open defecation (Rural)	35	14.3
Open defecation (Urban)	34	14.7
Improved water (Rural)	50	18.0
Improved water (Urban)	48	20.8
Piped Water (Rural)	50	28.0
Piped Water (Urban)	47	25.5

*Countries for which over 10 data points were available for the JMP 2014 update provided at least two were prior to 2000.

Sensitivity to data updates

James Fuller also presented results from an analysis of the implications of removing recent data points from the JMP and the impact on the MDG baseline and target. First the 1990 baseline was estimated for each country using data points before 2001. Then it was estimated using all data points. The change is the difference between the 1990 estimate between all data points currently available in 2014 and data points from before 2001. The impact on the baseline, current estimates and MDG target is illustrated in Figure 3. The summary table in the background paper shows that small changes are typical but there is a wide range (see Table 2).

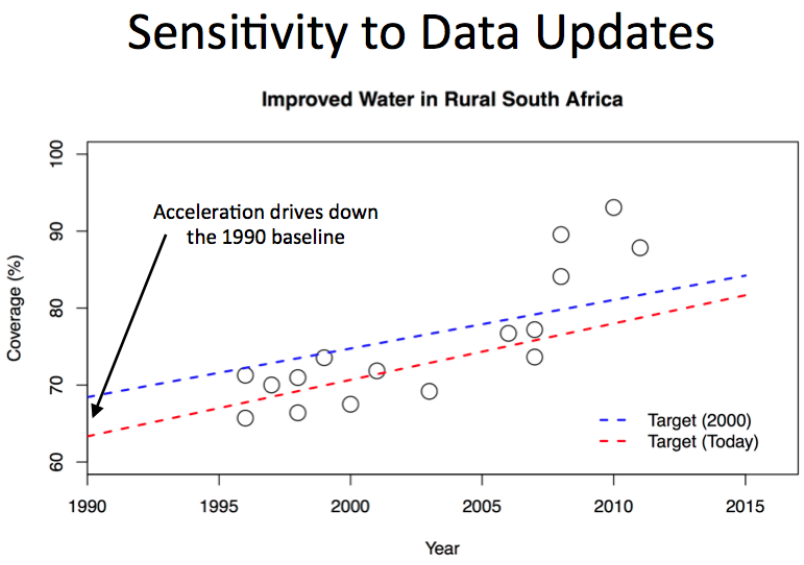


Figure 3: Example of sensitivity to data update, including impact on baseline and MDG target (Source: Fuller et al. 2014, in draft)

The taskforce discussed the implications of a shifting baseline (Table 2), noting that this can be the result of both the addition of a recent survey as well as old surveys which had previously not been accessible. The differences were small for most countries (<5% for 71 to 96% of countries) but there are some countries in which these are more substantial as shown by the ranges in Table 2. Any statistical model will face challenges if old data points are added. Shifting baselines were a particular problem during the MDGs but may be less of an issue during the SDGs since targets will most likely be either "universal" or absolute rather than based on proportional reduction (e.g. halving). Whilst the baseline will become less important, metrics for rates of progress (e.g. average annual reduction) will need to be defined and an interval selected for their calculation.

Recommendation 1. There is evidence of non-linearity in the JMP data. There are at least a few different patterns. Of those countries with over 10 data points, 15-25% have non-linear trends depending on the indicator. Without a reasonable number of data points (>10) spread out over several years curvature cannot be assessed.

The taskforce considered whether it would be possible to conduct an entirely objective analysis of the patterns. Several approaches can be used include leave one (or many) out cross-validation. Whilst it would be possible to generate tests of significance comparing different models and to summarize the fit for any given model (goodness of fit statistics) it was agreed that there will need to be some degree of subjective assessment to capture sector knowledge. In particular this relates to assessment of the plausibility of rapid changes that may result from overfitting.

Table 2. Comparison of the change in the 1990 baseline estimates between 2000 and 2014 using linear regression. (Fuller et al. 2014)

		Linear Model			
			Percent of Countries with change less than[3]:		
Indicator	**Number of countries[1]**	**Range[2]**	**1%**	**2%**	**5%**
Sanitation Indicators					
Improved (Rural)	47	-17 to 8	47%	57%	85%
Improved (Urban)	47	-22 to 6	55%	70%	89%
Open Defecation (Rural)	35	-35 to 15	29%	46%	71%
Open Defecation (Urban)	34	-21 to 7	59%	76%	88%
Drinking Water Indicators					
Improved (Rural)	48	-13 to 6	50%	65%	85%
Improved (Urban)	46	-6 to 7	57%	76%	96%
Piped (Rural)	49	-15 to 2	61%	82%	92%
Piped (Urban)	46	-16 to 9	39%	70%	85%

[1]With >10 data points and >2 data points before on or before 2000.
[2]The most extreme changes, largest decrease to the largest increase in the estimated 1990 baseline.
[3]The change is less than XX% in either direction. For 5%, the change was between -5 and +5 percentage points.

Recommendation 2. A combination of quantitative and subjective assessment will be needed to adequately compare alterative models. Subjective assessments are needed to establish whether the predicted trends are "credible" based on knowledge of the water and sanitation sector and the plausible rate of change in different countries.

Recommendation 3. It is acceptable to use different models based on availability of data as long as there is a transparent approach to select these and the number of models is limited to enable comparison.

Session 2: Alternatives to linear regression

Questions posed to taskforce

Are there any major methods that should be considered but were not included in the background paper?

How suitable are alternative statistical methods in terms of the criteria identified in the previous discussion?

Should the JMP consider using different modeling approaches for countries with different patterns and/or availability of data? If so, how should the JMP decide which approach to use for a given country?

Are there any rules the JMP could use to decide what modeling approaches to use for different indicators? What are the criteria that need to be considered?

Linear, piecewise linear, logit and quadratic regression and GAM

Mr James Fuller provided participants with a summary of the findings for the background paper which evaluated four alternative approaches to linear regression. These included quadratic, logit, piecewise linear and generalized additive models (GAM) – examples of which are given overleaf in Figures 4a and 4b.

Quadratic. Regressions with a quadratic term force a direction change but this was not a pattern that was observed in the JMP data. This approach can yield extreme results with few data points (n<5) and is not considered a viable option for the JMP.

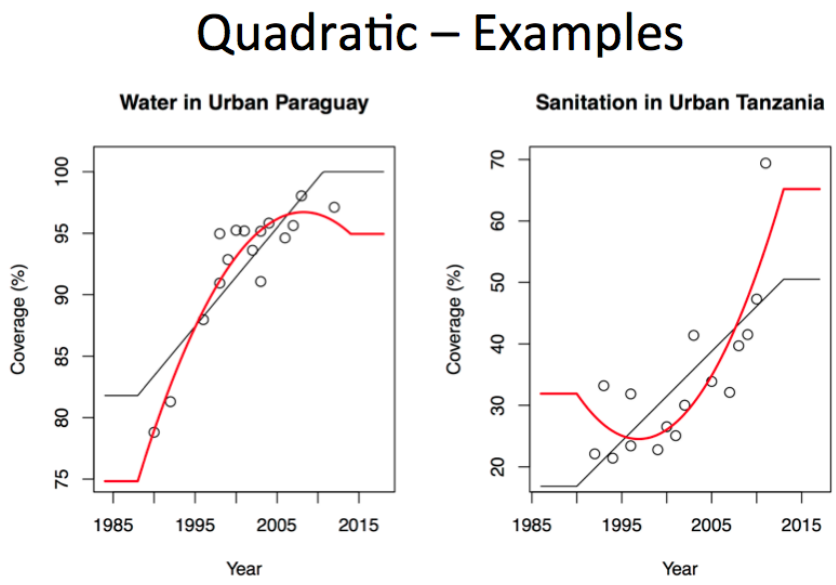
Logit. Transformation of the dependent variable using a logit (log odds) results in an "S-shaped" curve. In Bostoen (2006) there was a suggestion that this might be the general trend for water and sanitation services however this is not observed in the JMP data. The models works well for saturation but if acceleration takes place at intermediate levels of coverage it is essentially linear. Similar issues occur with deceleration occurring far from 0% or 100%.

Piecewise linear. Linear regression can be extended by joining two or more lines. Where the two lines intersect is known as the knot. In the simple case of a single knot, this is a basic analysis and possible using Excel. The key decisions are the number of knots and their locations. R has a package to optimise the location and software can also determine the appropriate number of "knots" within a given dataset.

GAM. Generalised additive models are an approach that can be used to derive a "smooth" curve. GAM does not have a specific functional form but rather relies on penalizing curves for too much curvature given the amount of data. When there are few data points the algorithms will choose a simpler model and it defaults to linear regression when there are very few points. The models can be fit using statistical software (R, Stata etc.) but not in Excel.

Mr James Fuller also noted a few other methods that were not considered in the background paper but could be considered for use by the JMP: higher order polynomials (cubic etc.), logarithmic or exponential growth, cubic splines or loess/lowess.

Table 3 provides a summary of evaluation of the aforementioned models on criteria including accuracy of the curve in reflecting the data, ease of explanation and ease of execution. Further details are available in the background paper (Fuller et al. 2014).



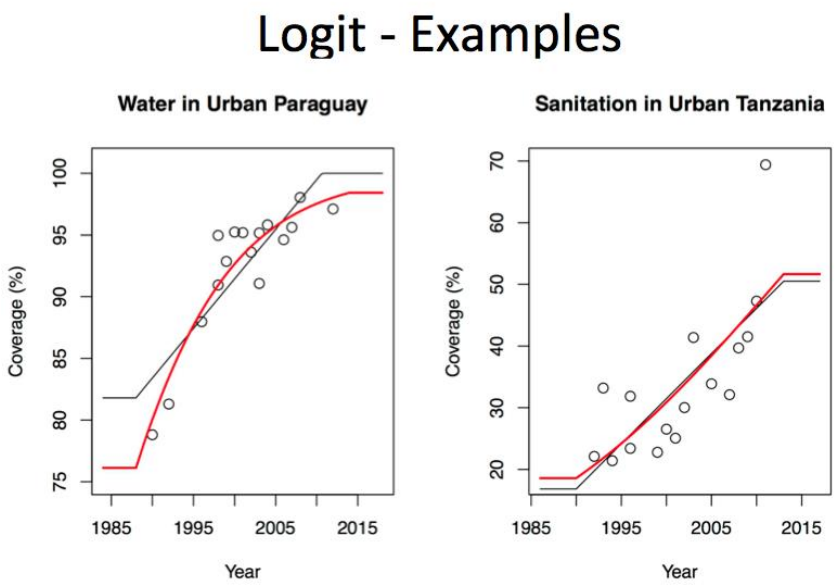
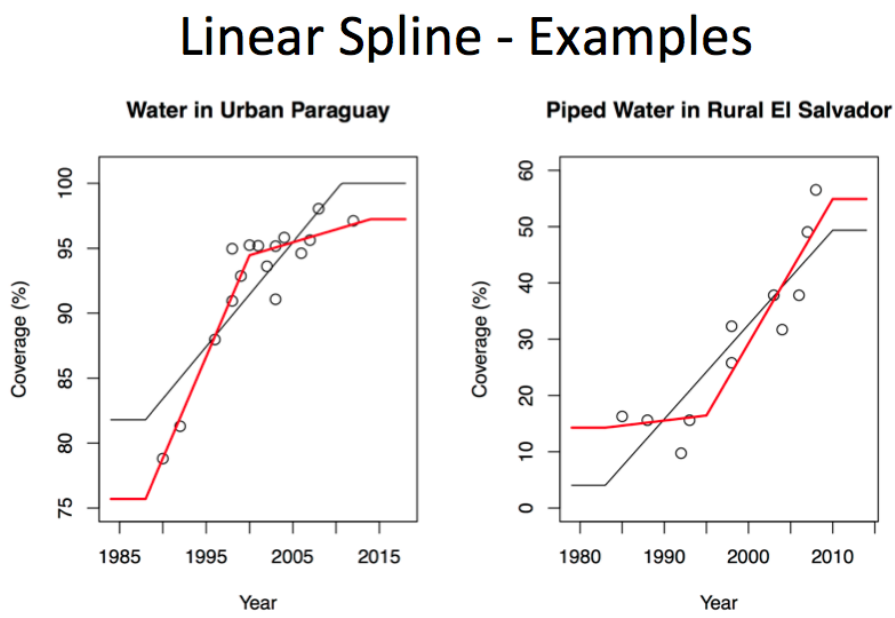


Figure 4a: Quadratic and logit regression applied to improved water in urban Paraguay and Improved sanitation in urban Tanzania (Source: Fuller et al. 2014)



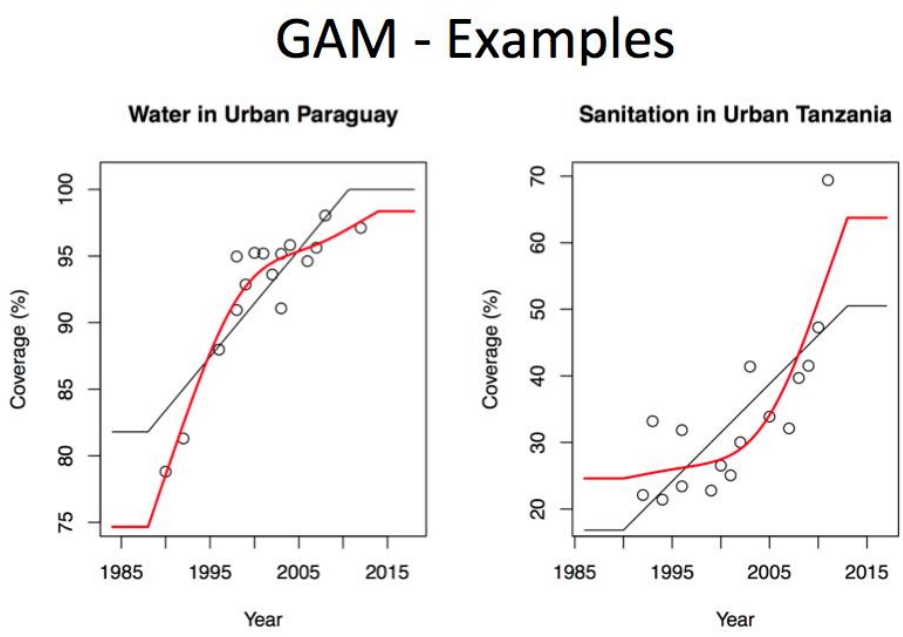


Figure 4a: Linear spline and GAM applied to improved water in urban Paraguay and Improved sanitation in urban Tanzania (Source: Fuller et al. 2014)

Evaluation of four alternatives

Table 3: Comparison of four alternative methods (source Fuller et al. 2014)

Comparison

Method	Pros	Cons	Ranking		
			Accuracy	Explanation	Execution
Quadratic	1. Fits all patterns 2. Easy to implement 3. Easy to explain	1. Parabola 2. Sparse data → unlikely trajectory	3	2	1
Logit	1. Fits saturation 2. Easy to implement 3. Easy to explain	1. Other trajectories	4	3	2
Linear Spline	1. Fits all patterns 2. Easy to explain 3. Fairly easy to implement	1. Location of breakpoints 2. Gradual changes	2	1	3
GAM	1. Fits all patterns 2. Often defaults to linear	1. Difficult to explain 2. Requires specialized statistical software	1	4	4

In the plenary discussion following the presentation the taskforce discussed the appropriateness of these methods for the JMP and highlighted a number of additional models that could be considered. There was consensus around the need to better articulate the different purposes of the JMP monitoring and to revise the evaluation criteria (see session 1). It was also agreed that it would be important to evaluate models for all countries and when doing so that candidate models are all displayed on the same graph. It was suggested that at a minimum this should include piecewise linear regression and smoothing methods (for example B-spline or lowess/loess). There was general agreement that it is defensible to use different models for countries with differing amounts of data as long as the rules are transparent and objective.

In relation to evaluating models the following remarks were made:

- It is usual practice to look at the data before selecting models (this is not possible for some of the new SDG indicators).
- Models should fit the data well but not too well - the model needs constraints so that the data are not overfit
- Projections need to be credible. Accuracy cannot be assessed prospectively but we can conduct objective tests to look at accuracy of past projections.
- More flexible, semi-parametric, models may perform poorly if the first or last point is an outlier. However it may be possible to identify and exclude outliers.
- The longer the time period for which the JMP has (and uses) data the less credible a linear regression appears. One option may be to exclude data before a particular year (for example 2000).

To model service levels multinomial logit regression could be used. A further suggestion was to investigate the possibility of using an isometric log transform or other similar transformation designed for compositional data (e.g. proportions that sum to 1). These transforms may offer a means to use standard statistical modelling approaches whilst retaining the compositional nature of the data.

Recommendation 4. A more detailed side by side comparison of alternative models is recommended and should include piecewise linear regression and well as considering smoothing methods (B-spline and Gaussian process regression. Splines, lowess/loess).

Recommendation 5. To produce consistent estimates of different service levels the JMP could consider multinomial logit and/or isometric log transforms.

Criteria for evaluating models

Throughout the meeting the taskforce members were asked to reflect on criteria for evaluating models. Recommended criteria are shown in Table 4 and take into consideration the primary purpose of the JMP to: provide country estimates and trends, though regional and global estimates are often reported; use comparable methods across countries and over time; and disaggregate service levels and show disparities.

Table 4: Criteria for evaluating the suitability of alternative modeling approaches

Criterion	Illustrative question
Credibility of the model	
Accuracy of estimates and credibility of predictions	How closely do trend lines fit the data and are the predictions credible?
Consistency across JMP SDG indicators	Can the modeling approaches be used in a consistent manner across all SDG indicators?
Objectivity of analysis	Are subjective judgments needed to run the model?
Understanding and implementing the model	
Ease of explaining the model	How easily can the model be explained to the JMP's audience including WASH sector stakeholders and policymakers?
Ease of explaining and communicating trends	How easily can the trends be explained and communicated to the JMP's target audience including WASH sector stakeholders and policymakers?
Ease of generating and displaying results	What software is needed to run the model and would it be possible to do so with simple graphical displays (including on websites)?
Within capability of the JMP team	Is the JMP team capable of conducting and explaining the analysis?

Several criteria were discussed but were not included in the final list:

- The need for developing countries to be able to replicate and verify estimates was considered as a criterion but there was disagreement over how often this takes place with the current model. Instead it was felt that the most important aspect was to have a transparent and objective approach. In examples of country data-related missions a common theme was that the most important aspects were determining whether all relevant data were included. As such the taskforce felt it was important to make a clear distinction between ease of explaining the model and the estimates – many users of the JMP data do not need to understand exactly how the curves are fit but may be more interested to see that they fit the data well.
- Consistent estimation of service levels was not considered to be a priority since normalization procedures can be used and since this would be restrictive (e.g. multinomial regression – see *alternative models*).
- It was felt that consistent estimation of disparities should not be a criterion to judge models though it was recognized that differences can arise when subsets of the JMP dataset are used for equity analyses.
- As outlined above, the stability of baselines was of particular importance during the MDGs but this is expected to be less critical during the SDGs.

Session 3: Generating global and regional estimates with scarce data

Questions posed to taskforce

What examples are available where other information (e.g. covariates) have been used for the purposes of generating regional and global data in global monitoring?

How should the JMP determine when is it reasonable to report global and regional aggregates? Should the current JMP rules be revised?

What are the implications of data scarcity for choice of methods?

Mr Rob Bain described a systematic review and modelling study to investigate exposure to water sources containing *E. coli* (Bain et al. 2014). Similar to earlier studies, almost two billion people were estimated to use a drinking water source contaminated with faecal indicators. The magnitude of this gives the JMP evidence of the need to track water quality. In the context of data scarcity, global estimates (and regional estimates) can help raise the profile of an issue. Other recent WASH-related estimates have been generated for wastewater treatment and handwashing.

Dr Annette Pruss-Ustun provided some reflections based on her work using multilevel modelling. A single model can be used for all countries. Expert opinion should be used to evaluate whether the fit and trends are credible. Visual inspection of each estimate in every country should be done. Multilevel modeling helps to prevent overfitting which would otherwise happen with the cubic spline. Care needs to be taken in the evaluation of different models since error metrics (e.g. RSME) will tend to favour overfitting

Mr James Fuller presented finding on an evaluation of multilevel modeling (without curvature), noting some potential disadvantages relative to linear (least squares) regression such as "shrinkage". A comparison was made between the current JMP method and multilevel linear regression based on the root squared mean error (RSME) from removing different numbers of data points from each country in turn. In doing so the researcher investigated how much error arises due to "borrowing" information from the region: are regional trends a good estimate for countries with no data? Multilevel model performed very poorly when there no data points for a given country, slightly worse than linear regression for 2-4 data points and similarly for more data points. James also illustrated this by showing data from Bolivia.

Use of covariates

Examples in other sectors which use covariates include cause of death trends, unsafe abortion and work on international futures (projections). Cause of death estimates developed by WHO and Johns Hopkins are very complex predictive models. Unsafe abortion estimates are in preparation by WHO and will use a Bayesian covariate model but will not report at the country level. International futures work seeks to predict patterns of potential human progress including infrastructure (improved water and sanitation) and uses a multinomial model with covariates. The taskforce was in agreement that the preference should be to avoid the use of covariates where data allow. It was also noted that the selection of covariates would depend on the indicator and would likely vary considerably between indicators (for example infrastructure vs behavior).

The use of regions in multilevel models is in effect a form of covariate and its use makes strong assumptions about the relationship between countries within a region. There was a discussion on the sensitivity of models to the use of different regional groupings and the possibility of using five WASH country clusters derived by hierarchical cluster analysis. Concern was expressed about assuming countries with no data follow regional trends. In the absence of evidence to support a strong link it was felt that reporting modelled estimates at the country level should not be done where no data are available. If using modelled estimates, it was agreed these should be clearly indicated as such in reporting.

Reporting regional and global aggregates

The taskforce members clearly distinguished between two differing purposes: (i) highlighting a problem and (ii) monitoring. The value of reporting "no estimate" to highlight gaps in monitoring was discussed and the potential disincentive if modelling is used to derive estimates for countries with insufficient or no data. There was concern that modeling could amount to reporting on something we would like rather than what information we have and that it should be driven by data availability.

The JMP only reports estimates at the regional level if estimates are available for at least 50% of the region's population in a given year. This is a rule used in many other global monitoring programmes. For many new SDG indicators data will initially only be available for selected countries. With small numbers of studies the taskforce recommended simply reporting these country estimates and refraining from anything else (e.g. modelled estimates) though it recognised the utility of collaborating with others (for example academics) and referring to peer reviewed articles as a means of highlighting issues for which data are few. This form of collaboration was seen as mutually beneficial and steps to systematise such relationships could be taken. It was noted that "modelling" is common practice for other areas (e.g. WHO burden of disease and risk factors analyses) but did not recommend for water, sanitation and hygiene.

Ms Holly Newby noted that this is a decision which needs to be taken across sectors and that a key action point will be to discuss with other colleagues at UNICEF and WHO. In light of the SDGs this rule is being revisited by UNICEF's D&A section. When data are not available for a populous country (China, India) it is common practice to report estimates for the region and to add a footnote. When there are data for only a few countries (e.g. zinc), these can be grouped by region so as to make the most of the available data. A further consideration is whether the extent of population coverage should be reported alongside WASH coverage.

Recommendation 6. The JMP only reports regional estimates if at least 50% of the population are represented. A change to the minimum 50% of population rule was not recommended.

Recommendation 7. There is value in collaborating with academics to establish first estimates which have been referred to in JMP reports but do not fall within the remit of global monitoring. This exploratory information should be clearly separated in reporting.

Towards a Model decision tree

Reflecting on the implications of data scarcity and the "rules" the JMP should use to determine the most appropriate approaches, the taskforce members and JMP agreed on the need for a decision tree to map out the potential methods and how these relate to different scenarios that the JMP may encounter. An initial attempt at this is given in the box below.

Recommendation 8. A decision tree can be used by the JMP to select an appropriate approach depending on the availability, an illustrative example of which is shown below:

Scenario	Potential Methods
>10 data points (enough) per country/time period with reasonable spread	Semi parametric model of coverage/trends to last data point. Project scenarios based on recent rate of progress
<10 data points per country/time period or unreasonable spread	Simple linear model of coverage/trends Project scenarios based on overall rate of progress
<50% population per region	Existing regional rules apply (incl w/out China/India)
<50 countries (e.g. 10 per region)	Estimates for countries with data.
5 studies	Illustrative only no country estimates

Note: Table assumes that the data cleaning has already been done. Other methods for generating regional and global estimates based on scarce data can be used but are not recommended for use by the JMP in global monitoring.

Session 4: Data sources and their quality

Dr Richard Johnston provided an overview of the errors and uncertainties relating to JMP estimates and outlined sources of data that could be used to complement household surveys and that would be needed to provide information on proposed indicators for the SDGs. Errors can arise at the data gathering level (interviewers misclassifying facilities), during data harmonization (survey data not matching global definitions, adjustments applied) and when modeling (excluding surveys, linear regression). Current JMP practice relies on representative surveys accepted by national statistical offices and no attempt is made to assess quality and take this into account in reporting. Potential improvements could include: primary data generation (calculation of sample error, internal data checks – and their quantification), data harmonization (protocols for adjustments, weighting) and modeling (weighting, confidence bounds). Whilst it is relatively straightforward to calculate sampling errors, other sources of error are much more difficult.

Future sources of data that could be included in the JMP's estimates are: administrative data from regulators or service providers, novel data from non-state actors and earth observations. These pose great challenges for quality assurance and error management.

Mr Rouslan Karimov described the developments in immunization monitoring in the past decade. Each year a long and complex questionnaire is sent to 196 countries. Survey data and administrative data are both considered and a single "best point" is selected based on a set of rules (heuristics) applied consistently to all countries. Initially an expert working group met once a year to adjudicate but this was perceived as lacking transparency and not being replicable. A computer program (Prolog) was used to apply the same rules using Boolean logic and generates automatic reports for all countries. The program usually makes similar choices as the expert panel and only a small number of countries need to be manually adjusted. The software is open source and published. The grade of confidence is an assessment of the confidence in a given estimate. Data are rated as "low", "medium" or "high" depending on a few measures including internal consistency and plausibility. Denominators were often found to be a problem and these are now compared against those from UN Population Division. All decisions are recorded and the grades of confidence can be downloaded. For further information about the grade of confidence see Brown et al (2013).

In response to questions from the taskforce:

- The expert working group is primarily made up of staff at UNICEF and WHO headquarters but others have been invited (Supply Division, regional health advisers, GAVI representative) on an ad hoc basis. One year five populous countries took part to observe how estimates are derived.
- Around two thirds or more of the immunization data are derived from administrative records. There is plenty of data for immunization.
- To the frustration of many no sub-national estimates are available for immunization
- It took about a year to develop the system and this was run in parallel for one year. Based on the results they agreed to switch to the new method.

Questions posed to taskforce

What are the main concerns with different data sources? Are there any established criteria for judging data quality?

In modeling coverage, should the JMP consider weighting data by quality and/or precision or should quality criteria only be used on an include/exclude basis?

Should JMP report confidence intervals or other measures of uncertainty on its estimates? If so, which types of error should be included in the measures of uncertainty?

Main concerns with data sources

Following the presentations, the taskforce discussed their key concerns with the new forms of data which the JMP may use post-2015. Key concerns with data sources include representativeness for a given population and the reliability of the source of information. These concerns are also applicable to household surveys and the JMP must be careful about the criteria it sets for new sources of data. It was suggested that non-representative administrative data

might be shown alongside representative data (for example from household surveys) however others noted concern that these may not converge since they may measure different things (coverage by design standards versus use). Dr Richard Johnston suggested that there may be value in attempting to link different types of information – for example looking across the spectrum from inputs (GLAAS) and service provision (admin) to use (JMP) and health burden (GBD). The possibility of reporting information side by side rather than numerically combining to provide insights into trends in different measures was also considered (see *integrating and combining data*).

As an example of non-representativeness, the recent MICS in Mali was not conducted throughout the whole country due to conflict and is not representative in the strict sense. In these cases it may be appropriate to apply a correction factor and the JMP does for some countries. However, such adjustments need to be clearly recorded and may be more problematic for equity analyses.

Weighting by quality of precision

The JMP currently either includes or excludes a survey - equivalent to assigning weights of 1 or 0 respectively. Mr Rifat Hossain noted the difficulties that have occasionally been encountered when countries have challenged the JMP on why individual data points have not been used in the estimates. Although in recent years these have been recorded together with any adjustments applied to, this information is not publicly available. The JMP country files contains a "comment" box which as a minimum should clearly indicate why a survey has or has not been used. Furthermore, the taskforce indicated that the JMP should formalize its rules regarding inclusion of data and potentially consider a software implementation similar to that used by the immunization community.

Recommendation 9. The JMP should clearly record the reason for excluding any survey or census in the country files. To ensure consistency there is a need to formalize the rules for excluding/including data points and potentially consider their implementation using software.

Assessing the quality of surveys can be difficult for several reasons. It is challenging to assess quality based on reports since these are typically written in a positive light and because in many cases the relevant information may not be included. Whilst it is possible for some indicators to judge quality using the microdata (for example weight for height) this may not be possible for WASH indicators. The JMP often relies on identification of data points that are out of trend but this is only possible when there are sufficient data for a given country. As noted by Ms Holly Newby in the presentation on methodological developments in other sectors, IGME has been trying to quantify quality but has not found a systematic way to do so. In practice, UNICEF and WHO rely on internal processes whereby country offices submit and confirm estimates in consultation with national counterparts. Documentation is also required for surveys. Across the different sectors there are many examples of criteria that are being used to either fully accept or exclude surveys. For example, for some indicators it is essential that surveys conform to standard questions. Improvements could be made to make this process more transparent and consistent. As an alternative approach, Mr Tom Slaymaker suggested that perhaps it would be more useful to grade the confidence in an entire country file rather than trying to assess quality for individual surveys.

Recommendation 10. The assessment could be either at the country level or at the survey level. The latter may be more feasible given the many challenges in assessing individual household survey reports.

There was a lengthy discussion of the relative merits of weighting by study quality versus by confidence intervals/standard errors but no agreement on which was preferable. Other sectors are using confidence intervals to weight data points (including IGME) by these cannot account for non-sampling errors. A concern with this is that it implies that sampling error is the dominant form of uncertainty (or at least proportionate to other forms). There is also a conceptual difficulty that arises with census data (which in theory are a complete count). Use of study quality suggests that it is acceptable to include data from surveys that do not meet all criteria. Approaches to calculating confidence bounds are described below.

Reporting confidence bounds

The taskforce agreed that it would be valuable for the JMP to report confidence bounds where possible. Some measure of uncertainty is generally expected by the scientific community. A variety of approaches could be considered and the JMP would need to investigate which uncertainties can be incorporated. A first approach to deriving confidence bounds can rely on removal of data points from the model ("bootstrapping") as was implemented in Wolf et al. (2013). This only works if there are many data points and would not initially be appropriate for the new indicators the JMP is proposing to monitor. Furthermore, this only captures model uncertainty and may substantially overstate confidence. For HIV, a pragmatic approach that has been used is that sampling errors are increased by 0, 10 or 20% based on a rating of the quality of a particular survey. Reflecting on the discussion and the inability to account for "non-sampling" or systematic errors, it was noted that it may be more important to be transparency about the level of uncertainty than being able to calculate quantitative measure that could be misleading.

Recommendation 11. The JMP should consider report some measure of "uncertainty bounds" and a few alternatives should be examined.

Session 5: Approaches to integrating and combining data

Dr Kristof Bostoen noted that prior to 2000, the JMP relied entirely on administrative data. Many of the conclusions from earlier reports showed similar trends to those found based on data from household surveys and censuses. An example of combining data for a single criterion is assessing the number of people with access to a drinking water supply based on multiplying the number of water points by the number of people that can be served by these. Whilst this helps to make figures more reasonable they are clearly not equate to survey data. For example using this logic administrative data were found to over report coverage in rural areas and substantially underreport in urban areas. Reflecting on the shift from provider to user reported data, Kristof noted that this has led to a distancing of the JMP from the WASH sector in each country. Whilst administrative data have their problems they have value, providing information that is otherwise not possible to collect. It also possible that 'administrative data' are in fact surveys.

Mr Ricard Gine Garriga highlighted that post-2015 proposals require a move from unidimensional "coverage" to multidimensional definitions of access. Key considerations in developing indices

include what weights to use and what type of aggregation (e.g. additive). There are many ways to combine data and these should ideally be adapted for a given audience, for example to reflect the human right to water and sanitation. In aggregating we will of course lose information and hide changes within the individual components of the index. When composite index has two inputs from different data sources, the joint distribution of the two indicators is unknown. For this reason it may be valuable to also report each component separately. A separate analysis of proposed indicators may also provide a complementary perspective on the progressive realisation of the right to water and sanitation.

Dr Jim Wright described examples of geospatial analysis combining data from multiple sources. A secondary analysis of geo-referenced water quality data from the RADWQ study in Nigeria (2004-2005) sought to understand the relative risk of contamination in rural and urban areas and found that the results are heavily dependent on the definition of urban extent. As a further example there have been attempts to assess the risk of exposure to geogenic drinking water contaminants (including arsenic and fluoride) that combine a risk map based on geological information with population data (e.g. Amini et al 2008). It may be possible to use spatial approaches to combine non-nationally representative data but this would need to be explored.

Questions posed to taskforce

What examples are there of approaches that have successfully combined data in WASH and other sectors relating to a *single* access criterion (e.g. coverage of improved sanitation)?

What examples are there of approaches that have successfully combined data in WASH and other sectors relating to *multiple* access criteria (e.g. coverage of improved water and water quality information)?

How suitable are these techniques in terms of key criteria including ease of interpretation and feasibility of the JMP reporting at national, regional and global levels?

Approaches for a single access criterion

Examples of approaches focusing on a single criterion include:

- Imputation based on household characteristics or GIS. For example census and household survey data have been integrated using imputation and hierarchical models to generate estimates for poverty in small areas.
- Vital registration systems and survey data have been integrated to derive trends
- Linking facilities with household survey data to obtain information from children's health records

The immunization sector does not use combinations of data from different sources but rather makes a choice between competing estimates. Triangulation using multiple sources of data may be an important means to assess new sources of data but the taskforce were not aware of good examples that have achieved this.

Approaches for multiple access criteria

In terms of multidimensional or multiple criteria, there are many examples:

- Multidimensional indices such as HDI, MPI and MODA. A good source of guidance on the construction of these is available from COIN, JRC-OECD.
- Assessment of vulnerability of water supplies and sanitation facilities to climate change
- Geogenic mapping of arsenic and fluoride
- Water points and urban extent using geo-referenced drinking water quality data from the RADWQ study in Nigeria

Utility and feasibility of integrating and combining data

The taskforce cautioned that some multidimensional indices have not proven to be of great utility. There is often a need to "decompose" the indices to understand trends in the individual components and their contribution to the index value. Taskforce members stressed the need to ensure that any index is sufficiently "meaningful". As highlighted by Ricard's presentation, there are several frameworks which could be used to create a meaningful index for different audiences. In order to promote quality of composite indicators and ensure the meaningfulness of the results, the use of uncertainty and sensitivity analysis should be an essential step in index construction (Saisana et al 2005).

Recommendation 12: The JMP should initially report new indicators separately and explore approaches to developing a composite indicator or index once more data becomes available.

With regards to geogenic mapping, Jim Wright asked whether prior knowledge could be used to optimize water quality sampling (and thus reduce costs). Holly confirmed that this kind of approach had been used by DHS for selecting where to do biomarker testing but that it is not done systematically. An example was the testing of some biomarkers only in the capital of Uzbekistan and in the Verghana valley were certain known problems are. It was felt that this could potentially be used for Arsenic and fluoride, reducing the number of test kits and teams requiring training. Rick Johnston noted that similar considerations might apply in the use of chlorine residual tests as a proxy for microbial contamination in urban areas but that this would not typically apply in rural areas.

In general, the JMP uses national data with permission from statistical agencies. A concern was raised about the potential difficulties around the use of data from some new sources (especially geospatial) but it was agreed that the JMP is unlikely to move away from publically available data. Were data from small studies or those conducted by academic institutions or NGOs to be incorporated, subject to criteria to be determined by the JMP, a memorandum of understanding would be need to ensure transparent process and the ability to report said information.

Taskforce members reflected on the difficulties of integrating and combining new data sources, especially when these cannot be linked to individual households and the facilities they use. This would certainly be the case for wastewater treatment and some aspects of safely-managed drinking water (verified risk management plans).

Recommendation 13. There may be value in using prior information (including risk maps) to determine where to conduct water quality testing for geogenic contaminants.

Session 6: Monitoring Inequalities

Mr Martin Evans provided reflections on addressing equity in global monitoring. Three main reflections were: How sensible is it to align WASH with other SDG goals that promote equity? Which indicators can help understand relationships between goals? What inequality measure could be used? Poverty reduction will be an SDG and this will likely include targets for the elimination of extreme poverty ($1.25 per capita PPP) by 2030 as well as a target relating to shared prosperity. National metrics will likely include monetary (e.g. poverty lines) as well as multidimensional poverty (e.g. OPHI MPI) that includes aspects of education, health and living standards. It may be strategic to link water and sanitation with poverty goals and other targets such as those relating to nutrition. Expenditure surveys have "dwelling related" questions and this would mean not needing to rely on DHS and MICS etc. These can also capture the costs of water and sanitation. WASH is an important part of poverty reduction strategies and the relationship operates both ways – inadequate WASH is both a cause and consequence of poverty. Mr Martin Evans explained the potential to use a concentration index. Benefits include the ability to decompose the index using standard tools and assess the relative contribution of factors most closely associated with wealth-based inequalities. A key advantage is the widespread use in the health sector and the understanding of its strengths and limitations.

Reflecting on the recommendations of the END working group, Ms Meg Satterthwaite highlighted that this was very much a within sector exercise, with a focus on the empirical and normative (Human Rights) and is the outcome of a series of compromises. The working group wanted to present a series of issues that should be considered rather than outlining specific metrics (for example the equality checklist) and to stress the need for both universal basic access and progressive realisation. With regards to specific stratifiers, urban/rural is seen as important and should be retained. Wealth quintiles are perceived as useful and have strong advocacy purchase. More challenging inequalities include gender (beyond who collects water), age (the elderly can have difficulty accessing facilities) and marginalized groups (which differ between countries. A key point for consideration is the need to continue to emphasize individual-level disparities despite the lack of data; these should be a priority of the future. Meg also noted the working group recommendation to consider the results of surveys conducted by small groups, academics or NGOs when these are representative and meet certain criteria.

Questions posed to taskforce

For which disparities would it be meaningful to report aggregated data and trends? What examples do we have from other sectors?

What do you consider to be the high priority inequality metrics and stratifiers for the JMP?

What challenges do you foresee with monitoring inequalities in WASH?

Regional and global aggregation of inequalities

Not all inequalities can be meaningfully aggregated to the regional or global level. There was general agreement that it is meaningful to report aggregated trends for:

- ***Rural/urban***. Long reported by the JMP, there is a pronounced difference in coverage between rural and urban in many countries. It was recognised that this is still a useful although definitions can vary between countries.
- ***Service levels.*** Monitoring service levels is a form of inequality monitoring once comparisons are made between or within countries. The JMP current report water and sanitation service levels using ladders.

Whilst in theory possible to aggregate to regional or global levels the following pose challenges

- ***Wealth***. Asset indices are relative though use of poverty line may be possible and approaches are in development to "standardise" assets.
- ***Intra-urban.*** There is no standard definitions of informal urban areas which can include "slums" and "peri-urban" and few surveys make a distinction. These areas are also closely associated with lack of municipal infrastructure and services, including water and sanitation. Using geospatial analysis to identify such areas is an active area of research. Until standard definitions become available it was agreed that the JMP should instead focus on reporting access for the urban poor.
- ***Intra-rural.*** Similar to intra-urban there are no standard definitions. An example which has been used by the JMP is rural with and without road access (Lao PDR MICS11) as featured in the JMP 2014 update.
- ***Individual-level.*** With the exception of the person responsible for collecting drinking water (which is no longer a standard question in DHS), inequalities within the household cannot currently be monitored though in principle these can be aggregated.

With the exception of some specific groups (e.g. indigenous in the Americas) it is usually inappropriate to aggregate across group stratifiers (such as ethnicity, race, nationality, language, religion or caste etc.) and it is not possible to standardise across countries (advantaged vs disadvantaged). Furthermore for some groups there are often problems with the sample size of household surveys unless oversampling is specifically conducted (for example Roma population in Serbia MICS2013).

Other sectors prioritise different equity stratifiers. For example for under five mortality gender is the major inequality. The World Bank is currently focusing reporting on the bottom 40% as part of its strategic realignment to focus on poverty and "shared prosperity". In the recent *Global Monitoring Report* a combination of asset indices and expenditure-based metrics were used depending on the indicator even though it is well known these are poorly correlated. It was agreed that inequality monitoring for WASH will be heavily dependent on what others do such as the equity stratifiers included in household surveys. As such the JMP should consider strategically engaging in this dialogue with a view to ensuring that, to the extent possible, important inequalities from a WASH perspective are captured alongside WASH information in future data collection efforts. There will also be pressure to report against some metrics which may not be prudent and the JMP should be prepared to push back against these. There was general agreement on the need to prioritise though limited discussion on the highest priority inequalities.

Measures of wealth inequality

Measures of wealth-based inequality are particularly problematic. For example, asset indices are known to be related to household size and there is a potential issue with water and sanitation being included in the wealth index. It was noted that poverty lines and consumption expenditure

measures do not include water and sanitation unlike the wealth index. Poverty lines could potentially be aggregated whereas this procedure is complex for asset indices (which are relative measures) though several attempts have been made by DHS and academics. There is a possibility that poverty lines could be imputed on the basis of asset ownership.

Mr Rifat Hossain explained that the use of wealth quintiles has been gaining ground. In the latest JMP report (2014) an analysis was presented based on data from 75 countries and over 400 datasets, approximately half of which were DHS or MICS. This analysis has been conducted for both rural and urban quintiles. Aggregating these remains a challenge and the taskforce discussed a few alternatives which could be considered including (i) reporting an average of the progress in the poorest quintile within each country and (ii) reporting the number of countries where the poorest quintile is "on-track". However it was generally agreed that the priority for monitoring inequalities lies in disaggregation not aggregation. Disaggregation is most useful for global monitoring as this is what is useful at the country level.

Recommendation 14. Given the difficulties in defining "slums" or "peri-urban areas" the taskforce recommended that the JMP focus on reporting water and sanitation coverage amongst the urban poor (e.g. using an asset index), and by extension rural poor as a measure of intra-rural inequalities.

Recommendation 15. Assessment of wealth-related inequalities using a poverty line, expenditure or income would provide a measure which can be aggregated.

Use of census data and geospatial analysis

Geospatial analysis using census data or geocoded household survey data may offer valuable means of assessing inequalities. A recent analysis of data from MICS and DHS (Pullan et al. 2014) demonstrated the possibility of using small area estimation techniques but there is concern that the precision may not be that high and further work would be need to validate these approaches. The question of whether the JMP could investigate these was posed. It was felt that there would be value in knowing where water and sanitation coverage was lowest and that maps would be valued by governments even if they do not explain the underlying reasons for the patterns. Census data were thought to be particularly useful for monitoring access for small disadvantaged groups for which sample sizes would be too small in some household surveys.

Recommendation 16. Census data could be used to complement household survey data and as a means to validate spatial inequalities as assessed by household surveys.

Use of representative surveys from non-state actors

The taskforce noted the importance of being open to the use of representative data from sources other than household surveys and administrative data. More entities are thinking about collecting data in more rigorous ways. This could include surveys conducted by NGOs and research institutes. These could be particularly valuable for monitoring inequalities. The JMP would need to specify the criteria and create a mechanism for collecting this information. It was noted that there is value in sub-nationally representative data but where data are not representative it is unclear what value the JMP adds from collecting the information.

Recommendation 17. The JMP was encouraged to develop criteria for inclusion of representative data from non-state actors (NGOs, universities) which could in some cases be used to track inequalities.

Additional topics of discussion on day 3

Managing a change in method

In changing to a new method, approaches would be needed to ensure continuity and to explain the reason for the change as well as its impact on current estimates. In the past the JMP has been criticized for not making this comparison (even though estimates were similar) on adopting the linear regression approach. The credibility of the JMP would be at stake if changes were substantial or unexplained. The impact on historical trends (e.g. MDG targets) would be of lesser importance to countries once the final assessments have been made in 2015. Restricting analysis to years where individual estimates are less uncertain (e.g. since 2000) may lessen the differences. The use of a web-based interface so stakeholders can run complex models was seen as important and may help to facilitate a transition. The reporting of confidence bounds can also help in explaining changes in methods as has been the case for HIV – and the controversy surrounding a substantial decline in HIV prevalence estimates in India (5 vs 3 million).

Projections

An important aspect of models is their ability to provide plausible estimates outside the range of the data. Projections can serve two related purposes: (i) to provide comparable estimates for countries for the same year and (ii) in order to provide estimates that are closer to the present day. It should be noted that the JMP's projections are in fact "projecting the future of the past". Taskforce members had differing views on whether the JMP should continue to report estimates beyond the latest data point and how this projection should be made. The current JMP approach is to use a projection based on the linear regression for two years after the last datapoint and then extend the last point for up to four years. As such the JMP reports estimates for up to six years after the last datapoint. For countries with high (>95%) or low (<5%) coverage the last estimate is extrapolated indefinitely. A few alternatives were considered ranging from reporting the latest estimate and a rate of change metric to providing the "most likely" estimate or a series of projections (low, business as usual, high). A concern was raised that unlike some other sectors where a model is theory driven (e.g. disease transmission) for water and sanitation coverage we don't have any reason to believe the model is mechanistically right. The first of these would be problematic from the perspective of being seen as a backward step by many users and since countries often request that the JMP provide current rather than historic estimates (the JMP 2015 report will be an exception as 2015 data will be reported). Whilst consensus on projections was not reached, it was agreed that any change in methods would require revisiting the JMP projection methods.

Recommendation 18. If the JMP changes its method for it would also need to revisit the projection rules.

Rates of progress

Since the SDG framework is likely to be based around universal access or an absolute target the baseline is less of an issue. However, assessing the rates of progress to the target end date (provisionally 2030) and progress in narrowing inequalities will require the development of new metrics. Other sectors such as nutrition have been working on the use of an annual rate of reduction (ARR). A key point of discussion is what time period to use in the calculation. For some models it might be appropriate to use a derivative, for example the slope at the year of the final data point. Such assessments all implicitly assume a linear trend. The JMP has in the past reported whether or not countries are "on track" to meet the MDG target; this assessment was based on a comparison of countries' coverage relative to the level required if progress between the baseline and target had been linear (e.g. within 5% of the coverage level required in 2012 or <5% progress insufficient).

Recommendation 19. Baselines may become less important during the SDGs since the targets are expected to be absolute (e.g. universal). The JMP will need to select appropriate measures of rate of progress.

A further idea is to focus on *benchmarking rates*. Jamie Bartram outlined ideas which look at the rate of progress for a given level of coverage, an approach that has previously been used in economic and social rights and appears to be justified based on historical patterns in census and survey data from high-income countries. A frontier is defined and used to judge countries relative to those achieving the greatest historical rate of progress at their level of coverage; for further details see Luh et al. (2013). An interesting idea for further exploration and could provide a nuanced interpretation of the JMP data in a similar way to the % of the current population having gained access metric which attempts to account for population growth. A concern was expressed that historical rates do not represent the best possible future performance and these benchmarks should not be interpreted too conservatively – for example wastewater treatment in the US rose steadily whereas in Chile concerted efforts increased coverage from less than 10% to nearly 100% in urban areas in less than 20 years.

Annex 1. List of Task Force Participants

Task Force members

1. Professor James K. Bartram
   Director, Global Water Institute
   Gillings School of Global Public Health
   University of North Carolina at Chapel Hill
   Chapel Hill, NC, USA
   T: +1 919 966 3934
   jbartram@email.unc.edu

2. Dr Kristof Bostoen
   Senior Programme Officer
   International Research Centre on Water and Sanitation
   The Hague, The Netherlands
   bostoen@ircwash.org

3. Dr Joseph Eisenberg
   Professor
   Department of Epidemiology, University of Michigan
   Ann Arbor, MI, USA
   jnse@umich.edu

4. Mr Martin Evans
   Policy Advisor
   Division of Data, Research and Policy
   United Nations Children's Fund (UNICEF)
   New York, United States of America
   T: +1 212 303 7275
   mcevans@unicef.org

5. Mr James Fuller
   PhD candidate
   Department of Epidemiology, University of Michigan
   Ann Arbor, MI, USA
   jnse@umich.edu

6. Mr Ricard Gine Garriga
   Researcher
   University Research Institute for Sustainability Science And Technology
   Universitat Politècnica de Catalunya
   Barcelona, Spain
   Ricard.gine@upc.edu

7. Mr Gérard Payen
   UN Secretary General's Advisory Board on Water and Sanitation
   gerard.payen@wanadoo.fr

8. Dr Annette Prüss-Üstün
   Department of Public Health, Environmental
   and Social Determinants of Health
   World Health Organization
   Geneva, Switzerland
   pruessa@who.int

9. Dr Meg Satterthwaite
   Faculty Director
   New York University School of Law
   New York, USA
   margaret.satterthwaite@nyu.edu

10. Dr Neff Walker
    Senior Scientist
    Deptartment of Global Disease Epidemiology and Control, Johns Hopkins University
    nwalker@jhsph.edu

11. Ms Laura Wescott
    Statistician
    Department for International Development
    London, UK
    l-westcott@dfid.gov.uk

12. Dr Jim Wright
Associate Professor of GIS
University of Southampton
Southampton, UK
J.A.Wright@soton.ac.uk

Members of the JMP

13. Mr Robert Bain
Statistics Specialist (WASH)
Division of Data, Research and Policy
United Nations Children's Fund (UNICEF)
New York, United States of America
T: +1 212 303 7275
rbain@unicef.org

14. Mr Rifat Hossain
Technical Officer, Water, Sanitation, Hygiene and Health
PHE/SDH
World Health Organization (WHO)
Geneva, Switzerland
hossainr@who.int

15. Dr Richard Johnston
Technical Officer, Water, Sanitation, Hygiene and Health
PHE/SDH
World Health Organization (WHO)
Geneva, Switzerland
johnstonr@who.int

16. Ms Holly Newby
Chief of Data Analysis Unit
Division of Data, Research and Policy
United Nations Children's Fund (UNICEF)
New York, United States of America
T: +1 212 303 7932
hnewby@unicef.org

17. Mr Tom Slaymaker
Sr. Statistics Specialist (WASH)
Division of Data, Research and Policy
United Nations Children's Fund (UNICEF)
New York, United States of America
T: +1 212 303 7269
tslaymaker@unicef.org

18. Dr Andrew Trevett
Senior Advisor, Water, Sanitation and Hygiene (WASH)
United Nations Children's Fund (UNICEF)
New York, USA
T: +1 212 303 7913
atrevett@unicef.org

Rapporteur

19. Mr Robert Bain
Statistics Specialist (WASH)
Division of Data, Research and Policy
United Nations Children's Fund (UNICEF)
New York, United States of America
T: +1 212 303 7275
rbain@unicef.org

Annex 2. Agenda of the Methods Task Force

<u>Tuesday, 2 December</u>: Review of non-linearity in JMP data and suitability of alternative modelling approaches

Chair: Tom Slaymaker, UNICEF

08:30	Morning coffee
09:00	Welcome and tour de table *Holly Newby, UNICEF*
09:15	Objectives, meeting format and draft criteria for assessing methods *Rifat Hossain, WHO* *Tom Slaymaker, UNICEF*
09:45	Methodological developments in other global monitoring programmes *Holly Newby, UNICEF*
10:30	Refreshment
11:00	Session 1: Evidence of non-linear patterns *James Fuller, University of Michigan*
11:30	Session 1: Plenary discussion on non-linear patterns *Moderated by Rick Johnston, WHO*
12:30	Lunch break
14:00	Session 2: Empirical review of alternative modelling approaches *James Fuller, University of Michigan*
14:30	Session 2: Plenary discussion on statistical methods for modelling non-linearity *Moderated by Rifat Hossain, WHO*
15:30	Coffee break
16:00	Plenary discussion to evaluate suitability of alternative methods *Moderated by Rifat Hossain, WHO*
17:00	Data revolution for sustainable development *Robert Chen*, *CIESIN Columbia University*
17:30	Closing remarks by the chair

18:00 Drinks and group dinner

<u>Wednesday, 3 December</u>: Guidance on integrating new indicators and new sources of data post-2015

Chair: Rifat Hossain, WHO

08:30 Morning coffee

09:00 Recapitulation of Day 1
Rapporteur, Rob Bain, UNICEF

09:15 Proposed new elements for Post-2015 monitoring
Tom Slaymaker, UNICEF

09:45 Session 3: Generating global and regional estimates with scarce data
Joseph Eisenberg, University of Michigan
Rob Bain, UNICEF
Annette Prüss-Üstün, WHO

10:00 Session 3: Plenary discussion on modelling approaches to address data scarcity
Moderated by Jamie Bartram, UNC

10:45 Refreshments

11:15 Session 4: Data sources and their quality
Rick Johnston, WHO
Rouslan Karimov, UNICEF

11:30 Session 4: Plenary discussion on criteria for evaluating data quality
Moderated by Tom Slaymaker, UNICEF

12:30 Lunch break

14:00 Session 5: Approaches to combining data
Kristof Bostoen, IRC
Jim Wright, University of Southampton
Ricard Giné Garriga, University of Catalunya

14:30 Session 5: Plenary discussion on approaches to combining data
Moderated by Rick Johnston, WHO

15:15 Coffee

15:45 Session 6: Monitoring inequalities

Martin Evans, UNICEF
Meg Satterthwaite, New York University

16:00 Session 6: Plenary discussion on monitoring inequalities

Moderated by Rick Johnston, WHO

17:00 Closing remarks by the chair

<u>Thursday, 4 November</u>: Recommendations and ways forward

Chair: Tom Slaymaker, WHO

09:00 Recapitulation of days 1&2

Rob Bain, Rapporteur, UNICEF

09:30 Discussion of implications for JMP

Moderated by Rick Johnston, WHO

10:30 Refreshments

11:00 Identification of outstanding issues and actions to address them

Moderated by the Chairs

12:30 Confirmation of key recommendations and actions

Moderated by the Chairs

13:00 Closure of the meeting followed by lunch

Selected References

Amini et al (2008) Statistical Modeling of Global Geogenic Arsenic Contamination in Groundwater. Environ Sci Technol. 42: 3669–3675

Bain et al (2014) Global assessment of exposure to faecal contamination through drinking water based on a systematic review. Trop Med Int Health. 19(8):917-27.

Bartram et al (2014) Global Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges. International Journal of Environmental Research and Public Health. 2014; 11; 11(8):8137-65.

Brown et al (2013) An Introduction to the Grade of Confidence Used to Characterize Uncertainty Around the WHO and UNICEF Estimates of National Immunization Coverage. The Open Public Health Journal. 6: 73-76

WHO/UNICEF (2013) Final report of the JMP post-2015 working group on Equity and Non Discrimination (END). Available from: http://www.wssinfo.org/fileadmin/user_upload/resources/JMP-END-WG-Final-Report-20120821.pdf.

Fuller et al (2014) Background paper on alternative modelling approaches and non-linearity of JMP data (*in draft*)

Luh et al (2013) Equity in water and sanitation: developing an index to measure progressive realization of the human right. Int J Hyg Environ Health. 216(6):662-71.

Pullan et al (2014) Geographical Inequalities in Use of Improved Drinking Water Supply and Sanitation across Sub-Saharan Africa: Mapping and Spatial Analysis of Cross-sectional Survey Data. PLOS Med.

Saisana et al (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. Journal of the Royal Statistical Society Series a-Statistics in Society, 168, 307–323.

WHO/UNICEF (2013) WASH Post-2015: Proposed targets and indicators for drinking-water, sanitation and hygiene Available from: http://www.wssinfo.org/fileadmin/user_upload/resources/post-2015-WASH-targets-factsheet-12pp.pdf

Wolf et al (2013). An exploration of multilevel modeling for estimating access to drinking-water and sanitation. Journal of Water and Health. 2013; 11(1):64-77.