Background document 2.b

Surveillance checklist of standards and benchmarks:

Summary of questions/issues that have arisen during the development process to date and how they have been addressed

A. IMPLEMENTATION

1. Process for assessment: self-assessment or peer-review?

During a meeting of the TF subgroup on TB surveillance in September 2011, both self-assessment and external peer review were considered important. It was also agreed that embedding the tool within the Global Fund grant processes would be very useful.

*Remaining question:* What are the incentives to conduct an assessment?

2. Do all standards need to be met for surveillance data to be considered a direct measure of disease burden?

It has been clarified in the introductory text of the checklist document that for surveillance data to be considered a direct measure of disease burden, all of the 11 data quality/coverage standards listed in B1 and B2 need to be met. For a system to measure TB deaths, the single VR standard (B3) must be met. Standards specific to HIV/MDR/TB in children can be assessed individually.

At the same time, benchmarks have been developed on a graded scale (e.g. Met, Partially met, Not met, Could not be assessed/not applicable) because of the strong feedback that some intermediate level (versus pass/fail) is needed to allow the demonstration of progress over time.

*Remaining questions:*
  1) Will ‘passing’ a standard be possible, if a valid explanation is provided, even if a benchmark is not met? (e.g. EP-TB in a few EU and other high-income countries);
  2) Should some form of grading scheme be adopted to allow broad categorization of the current status of progress (e.g. A*, A, B, C)?

3. Level of assessment: national, subnational?

It was agreed and has been clarified in the introductory text of the checklist document that the primary purpose of the tool is to assess the system at the national-level – because the objective is to assess if the national surveillance data provide a direct measure of TB cases and mortality.

The TF subgroup on surveillance also agreed that, as much as possible, the checklist should be assessed using data that are available at the national level. Otherwise, the process of undertaking the assessment at sub-national levels was perceived to be very burdensome, even in countries with low TB burden and with high performing systems, which would have a negative impact on uptake and utility of the tool.
4. **Time period included in assessment and periodicity of using checklist**

It has been agreed that the assessment is intended for the last calendar year for which data are complete, which may vary by setting. The periodicity of use is every 3-5 years. This is now explained in the introductory text at the beginning of the checklist document. It was also agreed that implementing the tool more often and covering more than one year would be too burdensome in many countries.

5. **Development of a user guide**

Field-testing showed that some users were unsure about the methods that should be applied to assess some of the standards and benchmarks, leading to lack of standardization in implementation. The agreed solution was to develop a user guide that, for each standard, would explain a) rationale b) methods to be used to assess if the accompanying benchmarks are met c) a worked example and d) suggestions for corrective actions if the standard is found to be not met. It was also agreed that a glossary of terms should be part of the user guide (see also “language” below).

6. **Use of published/historical data**

If there is no other evidence that the quality of data in a system has declined, then previous work/published studies can be used as evidence that S&Bs are met. This is now clarified in the introductory text at the beginning of the checklist.

**B. GENERAL CONTENT**

1. **Language**

Some words were found to be problematic – for example, not universally used or understood. Where this was the case, the language was improved/adapted. For example, the original wording related to “core variables” was changed to a “minimum set” of variables. Other wording perceived as a possible barrier to the use of the tool was also modified without changing the underlying principle. A glossary (list of terms provided at the end of this document) is being developed that will provide definitions for each of the terms that may not be familiar to all users.

2. **Case definitions**

Case definitions differ across countries, and the US and Europe in particular use definitions that are not identical to WHO definitions. However, since most definitions are basically consistent (a TB case as defined by WHO = a TB case defined elsewhere), the wording of Standard B1.1 (among others) was refined to ensure that the way in which case definitions are used in the checklist are comparable and that the checklist can be used (and perceived to be applicable) in all countries.

3. **Universal tool for electronic and paper-based systems**

Several discussions focused on how to implement a single tool across different systems, in particular paper-based versus electronic recording and reporting systems. S&Bs have been defined separately for paper and electronic systems within the same checklist, to be used when applicable.
However, S& Bs are much easier to assess when electronic systems are in place and a national electronic case-based database is available for analysis.

4. **Why is timeliness not specifically featured as a standard?**

Timeliness is a relevant aspect of data quality, but it doesn’t have to be assessed for the purposes of understanding if data provide a good proxy of cases and deaths. Text was added to the Introduction to the checklist to note that the timing of the assessment would be dependent on the timeliness of reporting/routine validation/data cleaning that is done. As the tool currently stands, it is not useful in assessing a system’s ability to detect aberrations for outbreaks.

5. **Elements dropped from checklist after first round of field-testing**

After the first round of field-testing in 10 countries (Brazil, China, Egypt, Estonia, Japan, Kenya, Netherlands, Thailand, UK, USA), there was common feedback that various standards should be dropped. These included:

- A standard focusing on human resources (e.g. person in charge of system, whether there is an M&E team). It was agreed that there was no obvious link to performance of surveillance and it provided information about ‘why’ a standard might not be met rather than whether the surveillance system was performing well or not.
- A standard about the budget for surveillance/M&E. This was hard to assess almost everywhere e.g. because it was not easily assessable from national level. It also does not show whether the system is performing in terms of outputs.
- A standard that a national report should be produced annually.
- Several elements originally included in the list of “core data items” (now called the “minimum set” of variables) were dropped, including treatment outcomes and risk factors, because they were beyond the aim of the tool and variables related to TB/HIV and drug-resistant TB were moved to specific standards about TB/HIV and DR-TB.
- A standard on contact tracing – this was difficult/impossible to assess in almost all countries that carried out field-testing, and there is no clear link with whether all cases are being captured by the surveillance system.
- A benchmark for external consistency (smear-positive among pulmonary 60-74% if HIV prevalence <20%). There is considerable variability among countries in the % of pulmonary cases that are smear positive (even within only high-income countries), which was considered too great to allow a useful benchmark to be defined.
- A benchmark on over-diagnosis. This was thought too vague and difficult to assess, if not impossible to measure.
- A benchmark for the range of extra pulmonary TB over pulmonary in childhood TB 1.5-4 was considered too wide to be meaningful.

6. **Some S& Bs are outside the purview of NTP**

To the extent possible, the S& Bs focus on the surveillance system itself, which the TB programme should be able to influence. However, to accurately assess whether TB surveillance is providing an accurate measure of TB cases and deaths, the assessment does need to consider wider health system and information system issues (e.g. access to care and VR system), and there are two
standards for this purpose (one on health system access, one on VR). This is now explained in the introductory text at the beginning of the checklist.

7. Standards developed specific to MDR TB, HIV/TB and children

These are priority areas that were not fully captured by the original set of general S&Bs, as the tool was designed to assess a system’s ability to measure total burden of TB cases. Additional standards have been added to allow the specific assessment of whether TB surveillance is good enough to allow accurate monitoring of the level of HIV+ TB among all TB cases, the level of drug-resistant TB among new TB cases and the burden of TB in children.

- **Drug resistant TB (B4.1):** The decision to focus the standard on new cases, instead of previously treated cases which are in many settings more often included in routine guidelines for testing, is because re-treatment cases are often double-counted in TB surveillance systems and the heterogeneity in re-treatment cases (e.g. treatment after default, failure, relapse, etc) make them hard to systematically assess. Currently the benchmarks don’t reflect the quality of labs within a country. Furthermore, this standard focuses on surveillance of rifampicin resistant TB vs MDRTB.

- **TB in children (B4.3):** Diagnosis of TB in children differs and is often more problematic than in adults. Under-reporting from pediatricians is a well-known phenomenon, especially in high TB endemic countries, while there are new country data that suggest under-diagnosis (especially among the very young) in the same countries is also a problem. There are questions that remain unanswered in childhood TB, necessitating: a) the improvement of recording and reporting of children with TB and b) the generation of new, and as much as possible, nationally representative data. The childhood TB subgroup was consulted in the development of the current benchmarks. Subsequent analysis of the data and discussions suggest, however, that they will require some additional work. Benchmarks will be revisited again by that group.

C. SPECIFIC STANDARDS

1. **TB surveillance system designed to capture a minimum set of variables for reported TB cases (standard B1.2) – requires variables not part of paper-based systems following 2006 WHO guidelines**

To assess trends and to look at internal and external consistency, it was decided that all cases would be included in the denominator, even though this means that paper-based systems following the 2006 WHO R&R guidelines will fail both age and sex benchmarks, which were only intended for smear-positive cases.

*Remaining question:* Is it appropriate for the benchmarks for standard B1.2 to require data (age and sex dis-aggregations for all case types) that will not be captured by paper-based R&R systems that follow the 2006 WHO guidelines?
2. **Extensive source data verification beyond verification of case counts is not being proposed (standard B1.4)**

Reviewing source documents, especially at the national level, is time-consuming and challenging to do in a representative sample. Although the methods to assess standard B1.4 require audit visits to facilities to review records, to the extent possible, this tool focuses on case counts (instead of data for individual cases) because it is key to the objective. Furthermore, misclassification is not likely to impact incidence in a particular direction, and major problems would be picked up by the external consistency standard. Also, other tools, such as the RDQA used by the Global Fund are available to do this.

3. **Duplicate cases within a system (standard B1.7) – currently only assessed for electronic systems**

Currently, there is a standard to assess the extent to which duplicate records are searched for and removed only for electronic systems. The reason is that it is recognized to be very difficult if not impossible to assess the extent to which there are duplicate records in paper-based systems. If such a standard is included for paper-based systems, they are likely to fail because of the inability to routinely and systematically check this, especially at the national level. However, this may be perceived to be important to the aim of the tool, even in paper-based systems.

*Remaining question: Is there a need for a standard related to duplicate records for paper-based systems?*

4. **Internal consistency (standard B1.8) – hard to define good benchmarks**

It has proved challenging to identify suitable benchmarks to assess internal consistency. A benchmark examining internal consistency across different geographic areas within a country (in addition to consistency over time) has been discussed but not included because, by chance, a benchmark for over-dispersion may not be met in some parts of a country due to stochastic fluctuations. Among the last two benchmarks for this standard, there are also challenges. There is a time lag between trends in the national prevalence of HIV and trajectories of reported TB cases. Also, in countries where HIV prevalence and GDP per capita are either both increasing or both decreasing, it will be hard to assess the consistency of each indicator with trends in TB notifications because HIV and GDP per capita have opposite effects on TB. For example, there are at least 43 countries with increasing GDP/capita and increasing HIV in TB since 2007 (with missing data for 32 other countries) and for those countries it isn’t possible to predict the net effect of GDP and TBHIV on the direction of incidence.

*Remaining question: Can better/additional benchmarks be suggested to examine internal consistency?*

5. **External consistency benchmarks, notably for % of cases with pulmonary vs. extrapulmonary TB (standard B1.9)**

The benchmark that the percentage of TB cases with pulmonary TB should be between 60-90% is not met in a few EU and other high-income countries with high levels of EP TB (e.g. UK, Netherlands, Norway, New Zealand); it is also not met in many countries in the Eastern Mediterranean Region where reported levels of EP-TB are high. This issue is unresolved in the current version (except that countries are required to meet only 2 of the three proposed
benchmarks so they could still pass if they met the external consistency benchmarks related to M/F ratio and the age distribution of cases).

Remaining questions: What would be the best way to accommodate the high levels of EP-TB reported in certain countries, especially those that have high-quality surveillance systems according to the results for other standards?

6. Evidence on underreporting (standard B2.1) is often not yet available

It is recognized that many countries will not have the evidence needed to demonstrate current levels of under-reporting. Using the checklist will help to draw attention to the lack of evidence about under-reporting and should help to highlight the value of inventory studies (for which a guide is in the advanced stages of development) as well as regulatory mechanisms to better ensure reporting of all diagnosed cases.

Remaining questions:

1) Given the challenges to measuring in a standardized and quantifiable way the second benchmark (“TB reporting is a legal requirement that is strongly and systematically enforced, with penalties (financial and other) for non-reporting or incentives for reporting”), are there other recommendations to assess under-reporting?

2) Is the cut-off of the past 2-3 years appropriate or too recent for when a previously conducted inventory study was to have taken place?

7. Coverage (B2.2.) - How best to assess access to care?

There are challenges to finding good benchmarks for this standard. Measures of human resources (e.g. # health care professionals per 10,000 population) and total health spending per capita are potentially useful but not directly linked to TB surveillance systems. People in WHO working on health financing and HRH have been contacted regarding this standard, but benchmarks aren’t finalized yet.

Remaining question: Are the current benchmarks sufficient? If not, can better/additional benchmarks be suggested to examine access to care (or another way to get at the level of undiagnosed cases)?

8. Vital registration (standard 3.1) – Health Metrics Network tool vs. simpler benchmarks

There has been discussion of whether to use two relatively simply benchmarks for coverage and quality, or a more detailed tool developed by the Health Metrics Network. (See below. Additional information available at: http://www.who.int/healthmetrics/tools/Version_4.00_Assessment_Tool3.pdf)

The original version of the standard for VR had two relatively simple benchmarks, one related to national coverage and other to quality. After the meeting of the TF surveillance subgroup in September, this was modified to use a more detailed set of checks developed by the Health Metrics Network. Subsequent retesting and feedback suggested that this was more difficult to use compared with the original two benchmarks.

Remaining question: Are the two benchmarks currently listed sufficient for assessment of whether national surveillance data provide a direct measure of TB mortality or should the HMN tool be used?
### HMN tool: Civil registration

<table>
<thead>
<tr>
<th>Core dimensions</th>
<th>Items</th>
<th>Highly adequate</th>
<th>Adequate</th>
<th>Present but not adequate</th>
<th>Not adequate at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>III.B.1 Contents</td>
<td>B.1.1: There is a reliable source of nationwide vital statistics: civil registration; Sample Registration System (SRS); or Demographic Surveillance System (DSS)</td>
<td>Nationwide civil registration</td>
<td>Sample Registration System</td>
<td>Demographic Surveillance Systems</td>
<td>There is no reliable source</td>
</tr>
<tr>
<td></td>
<td>B.1.2: Coverage of deaths registered through civil registration</td>
<td>90% or more</td>
<td>70% - 89%</td>
<td>50% - 69%</td>
<td>&lt;50%</td>
</tr>
<tr>
<td></td>
<td>B.1.3: Cause-of-death information is recorded on the death registration form if civil registration is in place</td>
<td>90% or more</td>
<td>70% - 89%</td>
<td>50% - 69%</td>
<td>&lt;50%</td>
</tr>
<tr>
<td></td>
<td><strong>Note:</strong> Skip this item if civil registration is not in place</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III.B.2 Capacity &amp; practices</td>
<td>B.2.1: The country has adequate capacity to: (1) implement data collection; (2) process the data; and (3) analyse the data from civil registration or SRS or DSS</td>
<td>Adequate capacity for all 3</td>
<td>Adequate capacity for 2 of the 3</td>
<td>Adequate capacity for only 1 of the 3</td>
<td>Adequate capacity for none of the 3</td>
</tr>
<tr>
<td></td>
<td>B.2.2: Frequency of the assessment of completeness of civil registration</td>
<td>Each time census is conducted (every 5 to 10 years)</td>
<td>Each time census is conducted</td>
<td>Less periodically than census</td>
<td>Never conducted or do not know</td>
</tr>
<tr>
<td></td>
<td>B.2.3: <em>The International Statistical Classification of Diseases and Related Health Problems</em> (ICD) is currently in use for cause-of-death registration</td>
<td>ICD-10 detailed</td>
<td>Tabulation List ICD-10</td>
<td>ICD-9</td>
<td>No ICD used or ICD-8 or earlier or there is no cause-of-death registration</td>
</tr>
<tr>
<td></td>
<td><strong>Note:</strong> Score 0 if there is no cause-of-death registration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B.2.4: Proportion of all deaths coded to ill-defined causes (%) (garbage codes)</td>
<td>Less than 5%</td>
<td>5% - 10%</td>
<td>11% - 19%</td>
<td>20% or more</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>B.2.5: Published statistics from civil registration or SRS are disaggregated by: (1) sex; (2) age; and (3) geographical or administrative region (or urban/rural) <strong>Note:</strong> Score 0 if there is no civil registration or SRS</td>
<td>All 3</td>
<td>2 of 3</td>
<td>1 of 3</td>
<td>None of 3, or there is no civil registration and no SRS</td>
<td></td>
</tr>
<tr>
<td>B.2.6: Sample Registration System (SRS) developed and generating timely and accurate data <strong>Note:</strong> Skip this item if there is no SRS</td>
<td>Nationally representative</td>
<td>Partially representative</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.2.7: Demographic Surveillance System (DSS) sites developed and generating timely and accurate data <strong>Note:</strong> Skip this item if there is no DSS</td>
<td>Partially representative (at least 1 urban and 2 rural sites)</td>
<td>Non-representative</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.2.8: Verbal autopsy (VA) tool <strong>Note:</strong> Skip this item if there is no DSS or SRS</td>
<td>VA tool validated; questionnaire publicly available and consistent with international standards</td>
<td>VA tool validated</td>
<td>VA not validated</td>
<td>No verbal autopsy used by SRS and/or DSS</td>
<td></td>
</tr>
<tr>
<td>III.B.3 Dissemination</td>
<td>B.3.1: Lag between the time that data were collected and the time that statistics from civil registration/SRS/DSS were published <strong>Note:</strong> Score 0 if there is no civil registration or SRS or DSS</td>
<td>Less than 3 years</td>
<td>3 years</td>
<td>4 or 5 years</td>
<td>More than 5 years or statistics not published or no vital statistics system (civil registration, SRS, DSS) exists</td>
</tr>
<tr>
<td>III.B.4 Integration &amp; use</td>
<td>B.4.1: Information from civil registration/SRS/DSS on: (1) mortality rates; and (2) causes of death is used for national and subnational analysis <strong>Note:</strong> Score 0 if there is no civil registration or SRS or DSS</td>
<td>Both mortality rates and cause-of-death information are used</td>
<td>1 of 2 used</td>
<td>Not used or statistics not published or no vital statistics system (civil registration, SRS, DSS) exists</td>
<td></td>
</tr>
</tbody>
</table>
Surveillance checklist of standards and benchmarks and user guide:

Glossary of terms

Age of patient
All forms of TB
Automatic batch data
Capture-recapture analysis
Case of tuberculosis
Case types
Case-based surveillance system
Clinically diagnosed TB case
Confidence interval
Consistency and validation checks
Contingency and recovery plan
Coverage of vital registration of deaths
Culture or equivalent
Data completeness
Data dictionary
Data entry validation processes
Definite TB case
Drug-resistant TB
duplicate notification of TB
Electronic system
Expected rate of change (k)
Extra-pulmonary case of TB
First sub-national level
Garbage codes
GDP per capita
Geographic location
Health Expenditure
Health Metrics Network (HMN) tool
Human immunodeficiency virus (HIV) status
Ill-defined causes
Income level (low, middle, and high-income countries)
International Classification of Diseases (ICD)
Inventory study
Lab confirmed TB case
Mortality due to TB per 100,000 population
Mortality rate
Multidrug-resistant tuberculosis (MDR-TB)
National TB Program
Nationally representative sample
New case of TB
Out of pocket expenditure on health
Paper-based system
Patient identifier
Patient-based surveillance system
Previously treated case of TB
Probabilistic matching
Pulmonary case of TB
Real-time automatic data validation
Recording and reporting
Recurrent case of TB
Report completeness
Reported TB case
Reporting units
Re-treatment TB cases
Rifampicin susceptibility
Sample vital registration system
Scheduled periodic reports
Source documents
Standard operating procedures (SOPs)
Surveillance system
Total out-of-pocket health expenditure
Under 5 mortality rate
Unique identifier
Vital registration
Year of registration