WHO/HTM/TB/2011.6b

Annex 2. GRADE glossary and summary of evidence tables

GRADE glossary

Absolute effect

The absolute measure of intervention effects is the difference between the baseline risk of an outcome (for example, in patients receiving control interventions or estimated in the observational studies) and the risk of outcome after the intervention is applied; that is, the risk of an outcome in people who were exposed to or received an intervention. Absolute effect is based on the relative magnitude of an effect and baseline risk.

Bias

A systematic error or deviation in results or inferences from the truth. In studies of the effects of health care, the main types of bias arise from systematic differences in the groups that are compared (selection bias), the care that is provided, exposure to other factors apart from the intervention of interest (performance bias), withdrawals or exclusions of people entered into a study (attrition bias) or how outcomes are assessed (detection bias). Systematic reviews of studies may also be particularly affected by reporting bias, where a biased subset of all the relevant data is available.

Critical outcome

An outcome that has been assessed as 7–9 on a scale of 1–9 for the importance of the outcome when making decisions about the optimal management strategy.

Dose response gradient

The relationship between the quantity of treatment given and its effect on outcome. This factor may increase confidence in the results.

Evidence profile

A table summarizing the quality of the available evidence, the judgements that bear on the quality rating and the effects of alternative management strategies on the outcomes of interest. It includes an explicit judgement of each factor determining the quality of evidence for each outcome. It should be used by guideline panels to ensure that they agree about the judgements underlying the quality assessments and to establish the judgements.

High quality evidence

We are very confident that the true effect lies close to that of the estimate of the effect.

Important outcome

An outcome that has been assessed as 4–6 on a scale of 1–9 for the importance of the outcome when making decisions about the optimal management strategy. It is important but not critical.

Imprecision

Refers to whether the results are precise enough. When assessing imprecision, guideline panels need to consider the context of a recommendation and other outcomes, whereas authors of systematic reviews need only to consider the imprecision for a specific outcome. Authors should consider width of confidence intervals, number of patients (optimal information size) and number of events.

Incomplete accounting of patients and outcome events

Loss to follow-up and failure to adhere to the intention-to-treat principle in superiority trials; or in non-inferiority trials, loss to follow-up, and failure to conduct both analyses considering only those who adhered to treatment, and all patients for whom outcome data are available.

Inconsistency

Refers to widely differing estimates of the treatment effect (that is, heterogeneity or variability in results) across studies that suggest true differences in underlying treatment effect. When the magnitude of intervention effects differs, explanations may lie in the patients (e.g. disease severity), the interventions (e.g. doses, co-interventions, comparison interventions), the outcomes (e.g. duration of follow-up) or the study methods (e.g. randomized trials with higher and lower quality risk of bias).

Indirectness

Refers to whether the evidence directly answers the health-care question. Indirectness may occur when we have no direct or head-to-head comparisons between two or more interventions of interest; it may occur also when the question being addressed by the guideline panel or by the authors of a systematic review is different from the available evidence regarding the population, intervention, comparator or an outcome.

Influence of all plausible residual confounding

A factor used in GRADE to upgrade the quality of evidence. When the influence of all plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results show no effect than the confidence in the results may be increased.

Lack of allocation concealment

Those enrolling patients are aware of the group (or period in a crossover trial) to which the next enrolled patient will be allocated (major problem in "pseudo" or "quasi" randomized trials with allocation by day of week, birth date, chart number, etc.).

Lack of blinding

Patients, caregivers, those recording outcomes, those adjudicating outcomes or data analysts are aware of the arm to which patients are allocated (or the medication currently being received in a crossover trial).

Large magnitude of effect

A factor used in GRADE to upgrade the quality of evidence. When the estimates of the magnitude of a treatment or exposure effect are large or very large and consistent, then we may be confident about the results. A large effect may mean that the effect is real but it may not be large.

Low quality evidence

Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect

Moderate quality evidence

We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.

Not important outcome

An outcome that has been assessed as 1–3 on a scale of 1–9 for the importance of the outcome when making decisions about the optimal management strategy.

Observational study

A study in which the investigators do not seek to intervene but simply observe the course of events. Changes or differences in one characteristic (e.g. whether or not people received the intervention of interest) are studied in relation to changes or differences in other characteristic(s) (e.g. whether or not they died), without action by the investigator. There is a greater risk of selection bias than in experimental studies.

Optimal information size

Used to assess the precision of results for dichotomous and continuous outcomes. The threshold for precision is met when the total sample size is as great as or greater than the number of patients generated by a conventional sample size calculation for a single trial.

Quality of evidence

Refers to a body of evidence not to individual studies (that is, means more than risk of bias of studies). It includes consideration of risk of bias, imprecision, inconsistency, indirectness and publication bias, as well as the magnitude of treatment effect and the presence of a dose–response gradient. In the context of a systematic review, the ratings of the quality of evidence reflect the extent of our confidence that the estimates of the effect are correct. In the context of making recommendations, the quality ratings reflect the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation.

Randomized controlled trial

An experimental study in which two or more interventions are compared by being randomly allocated to participants. In most trials, one intervention is assigned to each individual but sometimes assignment is to defined groups of individuals (for example, in a household) or interventions are assigned within individuals (for example, in different orders or to different parts of the body).

Relative effect

The relative effect for a dichotomous outcome from a single study or a meta-analysis will typically be a risk ratio (relative risk), odds ratio or, occasionally, a hazard ratio.

Selective outcome reporting bias

Incomplete or absent reporting of some outcomes and not others on the basis of the results.

Stopping early for benefit

Trials stopped early (before protocol, in particular in the absence of adequate stopping rules) overestimate treatment effects.

Strength of a recommendation

The degree of confidence that the desirable effects of adherence to a recommendation outweigh the undesirable effects. Either strong or weak/conditional.

Strong recommendation

Most patients would want the recommended course of action, and only a small proportion would not; therefore, clinicians should provide the intervention. The recommendation can be adapted as policy in most situations.

Study limitations (risk of bias)

The risk of misleading results as a result of flawed design or conduct of randomized or observational studies. It is one of the five categories of reasons for downgrading the quality of evidence. It includes lack of allocation concealment; lack of blinding; incomplete accounting of patients and outcomes events; selective outcome reporting bias; and other limitations, such as stopping early for benefit, use of non-validated outcome measures, carryover effects in crossover trials, and recruitment bias in cluster-randomized trials.

Summary of findings table

A table summarizing the quality of the available evidence, the judgments that bear on the quality rating and the effects of alternative management strategies on the outcomes of interest. It does not include detailed judgements about each factor determining the quality of the evidence, but instead an overall summary of the quality.

Surrogate outcome

Outcome measure that is not of direct practical importance but is believed to reflect an outcome that is important; for example, blood pressure is not directly important to patients but it is often used as an outcome in clinical trials because it is a risk factor for stroke and heart attacks. Surrogate outcomes are often physiological or biochemical markers that can be relatively quickly and easily measured, and that are taken as being predictive of important clinical outcomes. They are often used when observation of clinical outcomes requires long follow-up. Also called: intermediary outcomes or surrogate end-points.

Very low quality evidence

We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.

Weak/conditional recommendation

The majority of patients would want the suggested course of action, but many would not. Clinicians should recognize that different choices will be appropriate for individual patients, and that they must help each patient arrive at a management decision consistent with his or her values and preferences. Policy-making will require substantial debate and involvement of various stakeholders.

Question 1. At what prevalence of MDR-TB in any group of TB patients is rapid drug-susceptibility testing warranted to detect resistance to rifampicin and isoniazid or rifampicin alone on all patients in the group at the time of TB diagnosis, in order to prescribe appropriate treatment at the outset?

Quality assessment							No of outcomes		Effect		Quality	Importance
Setting	**Design**	**Limitations**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other**	**No DST**	**Alternative strategy**	**Absolute effect vs no DST**	**ICER/outcome**		
Total deaths (non MDR-TB and MDR-TB) per 1000 new cases and ICER – no DST vs rapid H&R at diagnosis:								**Rapid H&R at diagnosis**	**Deaths averted**			
Moderate drug resistance (no HIV)	Simulation model based on observational studies	Serious[1]	Not assessed	Serious[2]	Not assessed	None	43.1	39.1	4.0	$34 218	⊕○○○	Critical
Low MDR-TB and high mono-H							39.3	37.6	1.6	$34 755		
High drug resistance							78.0	51.3	26.8	$27 771		
Moderate drug resistance (HIV)							123.6	117.7	5.8	$18 825		
Total MDR-TB cases per 1000 new cases and ICER – no DST vs rapid H&R at diagnosis								**Rapid H&R at diagnosis**	**MDR-TB cases averted**			
Moderate drug resistance (no HIV)	Simulation model based on observational studies	Serious[1]	Not assessed	Serious[2]	Not assessed	None	24.1	22.3	1.8	$75 972	⊕○○○	Critical
Low MDR-TB and high mono-H							9.4	6.4	3.0	$19 005		
High drug resistance							159.5	153.3	6.2	$120 553		
Moderate drug resistance (HIV)							23.8	22.1	1.6	$68 598		
Total DALYs per 1000 new cases and ICER – no DST vs rapid H&R at diagnosis								**Rapid H&R at diagnosis**	**DALYs averted**			
Moderate drug resistance (no HIV)	Simulation model based on observational studies	Serious[1]	Not assessed	Serious[2]	Not assessed	None	52 357	52 539	182	$744	⊕○○○	Critical
Low MDR-TB and high mono-H							54 592	54 695	103	$556		
High drug resistance							54 263	55 752	1489	$499		
Moderate drug resistance (HIV)							51 044	51 204	160	$687		

$ refers to US dollars

DALY = disability-adjusted life-year; DST = drug susceptibility testing; H = isoniazid; H&R = DST for isoniazid and rifampicin; HIV = human immunodeficiency virus; ICER = incremental cost-effectiveness ratio; MDR-TB = multidrug-resistant tuberculosis; XDR-TB = extensively drug-resistant tuberculos

Quality assessment							No of outcomes		Effect		Quality	Importance
Setting	Design	Limitations	Inconsistency	Indirectness	Imprecision	Other	No DST	Alternative strategy	Absolute effect vs no DST	ICER/outcome		
Total deaths (non MDR-TB and MDR-TB) per 1000 new cases and ICER – no DST vs rapid H&R at 2 months post diagnosis:								**Rapid H&R at 2 months**	**Deaths averted**			
Moderate drug resistance (no HIV)	Simulation model based on observational studies	Serious[1]	Not assessed	Serious[2]	Not assessed	None	43.1	40.3	2.8	$46 678	⊕○○○	Critical
Low MDR-TB and high mono-H							39.3	38.1	1.1	$39 264		
High drug resistance							78.0	59.0	19.0	$45 320		
Moderate drug resistance (HIV)							123.6	119.1	4.6	$30 007		
Total MDR-TB cases per 1000 new cases and ICER – no DST vs rapid H&R at 2 months post diagnosis								**Rapid H&R at 2 months**	**MDR-TB cases averted**			
Moderate drug resistance (no HIV)	Simulation model based on observational studies	Serious[1]	Not assessed	Serious[2]	Not assessed	None	24.1	23.1	1.0	$127 076	⊕○○○	Critical
Low MDR-TB and high mono-H							9.4	7.5	1.9	$23 696		
High drug resistance							159.5	155.4	4.1	$211 110		
Moderate drug resistance (HIV)							23.8	22.7	1.1	$127 570		
Total DALYs per 1000 new cases and ICER – no DST vs rapid H&R at 2 months post diagnosis:								**Rapid H&R at 2 months**	**DALYs averted**			
Moderate drug resistance (no HIV)	Simulation model based on observational studies	Serious[1]	Not assessed	Serious[2]	Not assessed	None	52 357	52 518	161	$800	⊕○○○	Critical
Low MDR-TB and high mono-H							54 592	54 679	87	$504		
High drug resistance							54 263	55 605	1342	$640		
Moderate drug resistance (HIV)							51 044	51 180	136	$987		

[1] All models are always simplifications of complex processes. They are based on assumptions and are sensitive to input parameters. Models are useful for generating hypotheses and making predictions, but they are not a substitute for true clinical data on outcomes generated through well designed evaluative and epidemiologic trials. Model results must always be carefully interpreted in light of these limitations.

[2] Sensitivity analysis performed around epidemiological parameters and conditions. Results were found to be quite robust.

Question 2. Among patients MDR-TB receiving appropriate treatment in settings with reliable direct microscopy, is monitoring using sputum smear microscopy alone, rather than sputum smear and culture, more or less likely to lead to the outcomes listed below?

A. Question: Should monitoring be performed using smear alone (monthly) in patients with MDR-TB?

Quality assessment							Absolute estimates		Effect	Quality	Importance
No of studies	Design	Limitations	Inconsistency	Indirectness	Imprecision[1]	Other	Monthly culture	Monthly smear	Relative risk monthly smear or monthly culture (95% CI)		
Treatment failure							Frequency/N (%)		Hazard of failure, compared with monthly culture		
11	Observational studies (IPD)	Serious limitations[2]	Serious inconsistency[3]	No serious indirectness	No serious imprecision	None	770/5958 (12.9%)	466/5958 (7.8%)	HR 0.36 (0.32, 0.40)	⊕⊕○○[4]	Critical
Prompt initiation of appropriate treatment – not measured											
Avoiding the acquisition or amplification of drug resistance – not measured											
Survival (death from TB) of failures detected by monthly culture in patients who die, what % are detected by smear?							Frequency/N (%)		% of failures (among deaths) detected by culture, also detected by monthly smear		
10	Observational studies (IPD)	Serious limitations[5]	No serious inconsistency	No serious indirectness	Serious imprecision[6]	None	579/909 (63.7%)	502/909 (55.2%)	86.70 (80.32, 93.58)	⊕○○○	Critical
Accelerated detection of drug resistance – not measured											

CI = confidence interval; HR = hazard ratio; IPD = individual patient data; MDR-TB = multidrug-resistant tuberculosis

[1] Imprecision due to publication bias or bias in available data sets is possible for all outcomes. There were a large number of reports for which no response was received after repeated requests for data. Insufficient bacteriology data was the reason for non-inclusion of data from six additional reports, some of which used only sputum smear for monitoring. A larger proportion of missing studies used standardized treatment than did included studies. In individualized treatment, regimens may be adjusted according to early results. This could affect the sensitivity and specificity of absence of conversion to predict death or failure. This could also affect the degree of discordance between smear and culture.

[2] These results are crude, unadjusted as event strata were too small to perform multivariable regression.

[3] Stratified analyses revealed a difference (significant =0.05) in the relative hazard of failure by smear positivity at baseline and individualized vs standardized treatment; differences were not significant by BMI (≤18.5 vs >18.5), HIV status, X-ray findings (no cavitation, unilateral disease with cavitation, bilateral disease with cavitation) or resistance to first-line drugs (MDR-TB only as referent, compared with MDR-TB +1 and MDR-TB +2). Because stratified analysis was performed and reveals sources of inconsistency, evidence is not downgraded.

[4] Downgraded for limitations, but upgraded for estimate of large effect.

[5] Nearly half of patients do not meet failure definition – according to any monitoring strategy – before death. Quality is not downgraded for this limitation as it is downgraded for imprecision.

[6] Quality is downgraded for imprecision due to heterogeneity among sites.

Quality assessment							Absolute estimates		Effect	Quality	Importance
No of studies	Design	Limitations	Inconsistency	Indirectness	Imprecision[1]	Other	Monthly culture	Monthly smear	Relative risk monthly smear or monthly culture (95% CI)		
Cost to the TB control programme in past 12 months (cost reduction relative to monthly culture and smear in US$ 2003)							**Mean of 12 months monitoring costs**		**Mean difference smear only vs culture and smear**		
12[2]	Observational studies	No serious limitations	Serious inconsistency[3]	Serious indirectness[4]	Serious imprecision[5]	None	$164.82[6] (±$150.14)	$88.34[7] (±$95.54)	$105.01 (range: 16.56–244.68)	⊕○○○	Important
Months to first of 2 consecutive negative results among patients with baseline culture positivity (relative to monthly culture)							**Median months (IQR) Patients converted/N**		**Hazard of conversion by monthly smear compared with conversion by monthly culture**		
11	Observational studies (IPD)	Serious limitations[8]	No serious inconsistency	No serious indirectness	No serious imprecision	None	3 (3, 3) 2897/4171	3 (2, 3) 3021/4171	1.16 (1.10, 1.22)	⊕⊕○○	Critical
Ability to detect delayed conversion associated with probability of poor outcome (death & failure)							**Sensitivity/specificity of non-conversion at 6 months**		**Relative sensitivity or specificity of non-conversion by monthly smear vs monthly culture at 6 months**		
11	Observational studies (IPD)	No serious limitations	Serious inconsistency[9]	No serious indirectness	No serious imprecision	None	65.1/61.6	55.3/61.7	0.85/1.00	⊕○○○	Not rated

[1] Imprecision due to publication bias or bias in available data sets is possible for all outcomes. There were a large number of reports for which no response was received after repeated requests for data. Insufficient bacteriology data was the reason for non-inclusion for data from six additional reports, some of which used only sputum smear for monitoring. A larger proportion of missing studies used standardized treatment than did included studies. In individualized treatment, regimens may be adjusted according to early results. This could affect the sensitivity and specificity of absence of conversion to predict death or failure. This could also affect the degree of discordance between smear and culture.

[2] From 8 countries over the time period from 1999 to 2011 (projected).

[3] There was substantial variability in range of 12-monthly cost estimates for smear alone (from $6 to $237.24) and for smear and culture (from $37.20 to $447.36).

[4] Cost data from sites (and times) other than study sites and duration were among the estimates used. Evidence is not downgraded for this as it is already downgraded for imprecision due, in part, to the same causes.

[5] Combined costs of smear and culture were, in some cases, the sum of the costs of the two components; this is believed to represent an upper limit of the cost of smear and culture. Other estimates were less than the sum of the component parts, presumably due to the use of a single sputum sample for the two tests. Cost data for combined smear and culture were only available for 1; cost data on one or the other component parts were available for 6 sites (2 in WHO's European Region, 1 in the Western Pacific Region, 1 in the South-East Asia Region, 1 in Americas and 1 in the African Region). Additional data on smear and culture cost were available from 4 other sites worldwide. Only represents health system cost.

[6] This represents the mean cost of 12 months of monthly smear and culture.

[7] 12 months of monthly smear only.

[8] This analysis was performed on patients with baseline culture positivity; this is the group most likely to start MDR-TB treatment. Interval to 2 consecutive negative smears was counted irrespective of baseline smear status. The alternative considered (and rejected) was to limit the risk group to patients with baseline smear and culture positivity. This would have further reduced the observations. It is likely that results from baseline smear-and culture-positive patients would not be generalizable to the remainder of the study population or to other populations with MDR-TB. Evidence was not downgraded for this limitation.

[9] Although not formally evaluated, variability among site estimates is very high. Evidence is downgraded.

B. Question: Should monitoring be performed using smear alone (bimonthly) in patients with MDR-TB?

Quality assessment							Absolute estimates		Effect	Quality	Importance
No of studies	Design	Limitations	Inconsistency	Indirectness	Imprecision[1]	Other	Monthly culture	Bimonthly smear	Relative Risk Bimonthly smear/ Monthly Culture (95% CI)		
Treatment failure							**Frequency/N (%)**		**Hazard of failure, compared to monthly culture**		
11	Observational studies (IPD)	No serious limitations[2]	Serious inconsistency[3]	No serious indirectness	No serious imprecision	None	770/5958 (12.9%)	440/5958 (7.4%)	HR 0.32 (0.28, 0.36)	⊕⊕○○[4]	Critical
Prompt initiation of appropriate treatment – not measured											
Avoiding the acquisition or amplification of drug resistance – not measured											
Survival (death from TB) of failures detected by monthly culture in patients who die, what % are detected by smear?							**Frequency/N (%)**		**% of failures (among deaths) detected by culture, also detected by monthly smear**		
10	Observational studies (IPD)	Serious limitations[5]	No serious inconsistency[6]	No serious indirectness	No serious imprecision[7]	None	579/909 (63.7%)	465/909 (51.2%)	80.31 (74.12, 87.02)	⊕○○○	Critical
Accelerated detection of drug resistance – not measured											

[1] Imprecision due to publication bias or bias in available data sets is possible for all outcomes. There were a large number of reports for which no response was received after repeated requests for data. Insufficient bacteriology data was the reason for non-inclusion for data from six additional reports, some of which used only sputum smear for monitoring. A larger proportion of missing studies used standardized treatment than did included studies. In individualized treatment, regimens may be adjusted according to early results.

[2] These results are crude, unadjusted as event strata were too small to perform multivariable regression.

[3] Stratified analyses revealed a difference (significant at 0.05) in the relative hazard of failure by smear positivity at baseline and individualized vs standardized treatment; differences were not significant by BMI (≤18.5 vs >18.5), HIV status, X-ray findings (no cavitation, unilateral disease with cavitation, bilateral disease with cavitation), or resistance to first-line drugs (MDR-TB only as referent, compared with MDR-TB +1 and MDR-TB +2). Because stratified analysis was performed and reveals sources of inconsistency, evidence is not downgraded.

[4] Downgraded for limitations, but upgraded for large effect estimate.

[5] Approximately half of patients do not meet failure definition – according to any monitoring strategy – before death.

[6] Nearly half of patients do not meet failure definition – according to any monitoring strategy – before death. Quality is not downgraded for this limitation as it is downgraded for imprecision.

[7] Very substantial heterogeneity among sites. Quality is downgraded for imprecision.

Quality assessment							Absolute estimates		Effect		
No of studies	Design	Limitations	Inconsistency	Indirectness	Imprecision[1]	Other	Monthly culture	Bimonthly smear	Relative Risk Bimonthly smear/ Monthly Culture (95% CI)	Quality	Importance
Cost to the TB Programme in last 12 months (cost reduction relative to monthly culture+smear in USD 2003)							**Mean (SD) of 12 months of monitoring costs**		**Mean difference smear only vs culture & smear**		
12[2]	Observational studies	No serious limitations	Serious inconsistency[3]	Serious indirectness[4]	Serious imprecision[5]	None	$164.82[6] (±$150.14)	$29.91[7] (±$39.33)	$134.92 (range: 19.80–346.02)	⊕○○○	Important
Months to first of 2 consecutive negative results among patients with baseline culture positivity (relative to monthly culture)							**Median months (IQR) Patients converted/N**		**Hazard of conversion by bimonthly smear compared to conversion by monthly culture**		
11	Observational studies (IPD)	limitations[8]	No serious inconsistency	No serious indirectness	No serious imprecision	None	3 (3, 3) 2897/4171	3 (3,3) 3021/4171	1.07 (1.02, 1.13)	⊕⊕○○	Critical
Ability to detect delayed conversion associated with probability of poor outcome (death and failure)							**Sensitivity/specificity of non-conversion at 6 months**		**Relative sensitivity/Specificity of non-conversion by monthly smear vs monthly culture at 6 months**		
11	Observational studies (IPD)	Serious limitations	No serious inconsistency[9]	Serious indirectness	No serious imprecision	None	65.1/61.6	57.0/60.8	0.88/0.99	⊕○○○	Not rated

IPD = individual patient data; OR = odds ratio; SD = standard deviation

[1] Imprecision due to publication bias or bias in available data sets is possible for all outcomes. There were a large number of reports for which no response was received after repeated requests for data. Insufficient bacteriology data was the reason for non-inclusion for data from six additional reports, some of which used only sputum smear for monitoring. A larger proportion of missing studies used standardized treatment than did included studies. In individualized treatment, regimens may be adjusted according to early results.

[2] From 8 countries over the time period from 1999 to 2011 (projected).

[3] Cost data for combined smear and culture were available only for 1; cost data on one or the other component parts were available for 6 sites (2 in WHO's European Region, 1 in the Western Pacific Region, 1 in the South-East Asia Region, 1 in Americas and 1 in the African Region). Additional cost data on smear and culture were available from 4 other settings worldwide.

[4] Cost data from sites (and times) other than study sites and duration were among the estimates used. Evidence is not downgraded for this as it is already downgraded for imprecision due, in part, to the same causes.

[5] There was substantial variability in range of cost estimates. Combined smear and culture costs were, in some cases, the sum of the costs of the two components; this is believed to represent an upper limit of the cost of smear and culture. Other estimates were less than the sum of the component parts, presumably due to the use of a single sputum sample for the two tests. Additional limitations are described in the attached cost analysis.

[6] This represents the mean cost of 12 months of monthly smear and culture.

[7] 12 months of bimonthly smear (6 samples) only.

[8] This analysis was performed on patients with baseline culture positivity; this is the group most likely to start MDR-TB treatment. Interval to 2 consecutive negative smears was counted irrespective of baseline smear status. The alternative considered (and rejected) was to limit the risk group to patients with baseline smear and culture positivity. This would have further reduced the observations. It is likely that results from baseline smear-and culture-positive patients would not be generalizable to the remainder of the study population or to other populations with MDR-TB. Evidence was not downgraded for this limitation.

[9] Although not formally evaluated, variability among site estimates is very high. Evidence is downgraded.

Question 3. When designing regimens for patients with MDR-TB, is the inclusion of specific drugs (with or without documented drug susceptibility) more or less likely to lead to the outcomes of interest?

Quality assessment							No of patients[1]		Effect estimate: OR (95% CI) treatment success vs failure or relapse	Quality	Importance
No of studies	Design	Limitations[2]	Inconsistency[3]	Indirectness[4]	Imprecision[5]	Other	Drug used	Not used	Adjusted Odds		
Resource use – not measured											
Use of pyrazinamide											
32	IPD meta analysis	Very serious	No concerns	No concerns	No concerns		6571	2582	1.2 (0.9, 1.7)	⊕○○○	Critical
Use of ofloxacin											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		6489	2664	1.7 (1.1, 2.7)	⊕⊕○○	Critical
Use of a "later-generation" quinolone											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		1258	7895	2.5 (1.1, 6.0)	⊕⊕○○	Critical
Use of ciprofloxacin											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		986	9153	2.3 (1.0, 5.2)	⊕⊕○○	Critical
Use of kanamycin											
32	IPD meta analysis	Serious	Moderate concerns	No concerns	No concerns		5002	3165	1.0 (0.4, 2.2)	⊕⊕○○	Critical
Use of capreomycin											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		1757	7396	1.0 (0.4, 2.4)	⊕⊕○○	Critical
Use of ethionamide/prothionamide											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		1757	1824	1.7 (1.3, 2.3)	⊕⊕○○	Critical

Continues..

Continued

Quality assessment							No of patients[1]		Effect estimate: OR (95% CI) treatment success vs failure or relapse	Quality	Importance
No of studies	Design	Limitations[2]	Inconsistency[3]	Indirectness[4]	Imprecision[5]	Other	Drug used	Not used	Adjusted Odds		
Use of cycloserine											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		5344	3809	1.1 (0.8, 1.7)	⊕⊕○○	Critical
Use of p-aminosalicylic acid (PAS)											
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		3196	5957	0.9 (0.6, 1.4)	⊕⊕○○	Critical
Use of Group 5 drugs (see guidelines text for list of drugs)[5]											
32	IPD meta analysis	Very serious	No concerns	No concerns	No concerns		2709	3196	0.5 (0.4, 0.7)	⊕⊕○○	Critical

IPD = individual patient data; OR = odds ratio; SD = standard deviation

[1] Overall results shown. These do not account for results of drug sensitivity testing.

[2] Limitations – Analysis based on individual patient data meta-analysis. All of the original studies were observational studies. As well, in the majority of studies therapy was individualized; this may have led to bias in that certain drugs may have been given to more seriously ill patients with worse initial drug resistance, or to patients who were not responding well.

[3] Inconsistency – Based on estimated I squared

[4] Imprecision – Based on 95% CI, which were narrow, reflecting the large patient population

[5] DST was rarely available for Group 5 drugs and therefore estimates shown for all patients.

Other notes:

- Later-generation quinolones and ofloxacin were significantly superior to no quinolone for all three outcomes. Later-generation quinolones included levofloxacin (750mg/day or more), moxifloxacin, gatifloxacin, sparfloxacin. These were significantly superior to ofloxacin: aOR 1.7 (1.1, 2.7). Ofloxacin was equivalent to ciprofloxacin (aOR 1.1(0.5, 2.5)).
- Kanamycin was also compared with capreomycin. In all patients, kanamycin was somewhat superior to capreomycin aOR 1.5 (0.8, 3.0). Kanamycin was significantly superior to capreomycin for success vs fail /relapse/died and success vs fail/relapse/died/ default. However, in kanamycin-resistant strains that were not resistant to capreomycin, kanamycin was insignificantly inferior for all outcomes (kanamycin vs capreomycin for success vs fail/relapse: aOR 0.3 (0.1, 0.8)).
- Use of ethionamide or prothionamide was significantly superior to non-use of these drugs for all outcomes.
- Use of cycloserine was significantly higher than non-use for success vs fail/relapse/died and success vs fail/relapse/died/default.

Question 4. When designing regimens for patients with MDR-TB, is the inclusion of fewer drugs in the regimen (depending on the drug used, the patient's history of its use and isolate susceptibility) more or less likely to lead to the outcomes of interest?

Quality assessment							No of patients[1]					Effect estimate: OR (95% CI) adjusted only success vs failure or relapse					Quality	Importance
No of studies	**Design**	**Limitations[2]**	**Inconsistency[3]**	**Indirectness**	**Imprecision[4]**	**Other**											⊕⊕○○	Critical
Resource use – not measured																		
How many susceptible drugs in total?							**0–2**	**3**	**4**	**5**	**6+**	**0–2**	**3**	**4**	**5**	**6+**		
32	IPD meta analysis	Serious	No concerns	No concerns	No concerns		328	439	1148	2296	1453	1.0 (ref)	1.4 (0.9,2.0)	2.8 (1.9,4.1)	3.9 (2.7,5.6)	3.7 (2.5,5.6)	⊕⊕○○	Critical
How many susceptible drugs in the initial phase?							**0–2**	**3**	**4**	**5**	**6+**	**0–2**	**3**	**4**	**5**	**6+**		
32	IPD meta analysis	Very serious[5]	No concerns	No concerns	No concerns		119	163	469	814	812	1.0 (ref)	1.3 (0.7,2.4)	2.3 (1.3,3.9)	2.2 (1.1,4.3)	2.4 (1.4,4.1)	⊕○○○	Critical
How many susceptible drugs in the continuation phase?							**0–2**	**3**	**4**	**5+**		**0–2**	**3**	**4**	**5+**			
32	IPD meta analysis	Very serious[5]	No concerns	No concerns	No concerns		255	551	600	564		1.0 (ref)	2.7 (1.7,4.1)	2.8 (1.6,5.1)	2.1 (1.4,3.3)		⊕○○○	Critical

[1] Analysis of results for success vs fail/relapse shown. Results with other outcomes were similar.

[2] Limitations – Analysis based on individual patient data meta-analysis. All of the original studies were observational studies. In the majority of studies therapy was individualized; this may have led to bias in that certain drugs may have been given to more seriously ill patients with worse initial drug resistance, or to patients who were not responding well.

[3] Inconsistency – Based on estimated I squared.

[4] Imprecision – Based on 95% CI, which were narrow, reflecting the large patient population

[5] In many studies the number of drugs in the initial phase and continuation phase was not provided. Hence the number of subjects on which this analysis is based is more limited

Question 5. In patients with MDR-TB, is shorter treatment, compared with the duration currently recommended by WHO, more or less likely to lead to the outcomes of interest?

Quality assessment							No of patients[1]								Quality	Importance
No of studies	**Design**	**Limitations[2]**	**Inconsistency[3]**	**Indirectness**	**Imprecision[4]**	**Other**									⊕⊕○○	Critical
Resource use – not measured																
Number of months of total treatment (no prior MDR-TB treatment)							6.0–12.5	12.6–15.5	15.6–18.5	18.6–21.5	21.6–24.5	24.6–27.5	27.6–30.5	30.6–36		
32	IPD meta analysis	Serious	Moderate concerns	No concerns	No concerns		743	384	1646	612	435	207	106	48	⊕⊕○○	Critical
Number of months of Initial treatment (all patients)							1–2.5	2.6–4.0	4.1–5.5	5.6–7.0	7.1–8.5	8.6–20				
32	IPD meta analysis	Serious[5]	Moderate concerns	No concerns	No concerns		308	1406	481	377	172	792			⊕⊕○○	Critical

Quality assessment							Effect estimate: OR (95% CI)adjusted only								Quality	Importance
No of studies	**Design**	**Limitations[2]**	**Inconsistency[3]**	**Indirectness**	**Imprecision[4]**	**Other**									⊕⊕○○	Critical
Resource use – not measured																
Number of months of total treatment (no prior MDR-TB treatment)							6.0–12.5	12.6–15.5	15.6–18.5	18.6–21.5	21.6–24.5	24.6–27.5	27.6–30.5	30.6–36		
32	IPD meta analysis	Serious	Moderate concerns	No concerns	No concerns		1.0 (ref)	2.4 (1.5,3.6)	4.6 (2.0,10.4)	9.3 (5.8,15.0)	6.8 (4.2,11.1)	8.2 (4.2,15.9)	2.4 (1.2,5.0)	1.3 (0.6,2.7)	⊕⊕○○	Critical
Number of months of initial treatment (all patients)							1–2.5	2.6–4.0	4.1–5.5	5.6–7.0	7.1–8.5	8.6–20				
32	IPD meta analysis	Serious[5]	Moderate concerns	No concerns	No concerns		1.0(ref)	1.2 (0.5, 2.9)	2.4 (1.3,4.3)	3.7 (1.9,7.1)	5.1 (2.1,12.7)	2.2 (1.2,3.9)			⊕⊕○○	Critical

[1] Analysis of results for success vs fail/relapse shown. Results with other outcomes were similar.

[2] Limitations – Analysis based on individual patient data meta-analysis. All of the original studies were observational studies. As well, in the majority of studies therapy was individualized; this may have led to bias in that certain drugs may have been given to more seriously ill patients with worse initial drug resistance, or to patients who were not responding well.

[3] Inconsistency – Based on estimated I squared.

[4] Imprecision – Based on 95% CI, which were narrow, reflecting the large patient population.

[5] Fewer studies provided information on number of drugs in the initial or continuation phase. All provided information on number of drugs used in total.

Question 6. In patients with HIV infection and drug-resistant TB receiving antiretroviral therapy, is the use of drugs with overlapping and potentially additive toxicities, compared with their avoidance, more or less likely to lead to the outcomes of interest?

Quality assessment							No of patients		Effect	Quality	Importance
No of studies	**Design**	**Limitations[1]**	**Inconsistency[2]**	**Indirectness**	**Imprecision**	**Publication bias[3]**	**ART use**	**No ART Use**	**Relative (95% CI)**		
Cure (failure)											
9	Observational studies	Serious[4]	Serious	No serious indirectness	No serious imprecision	Possible	33/72 (46%)	7/53 (13%)	HR 3.17[5, 6] (1.46,6.90)	⊕⊕○○	Critical
Prompt initiation of appropriate treatment											
see Table 2											
Avoiding the acquisition or amplification of drug resistance											
Studies not identified to evaluate this outcome											
Death from TB											
10	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	No serious imprecision	Possible	34/124 (27%)	48/83 (58%)	HR 0.41[6, 8] (0.26, 0.63)	⊕⊕○○	Critical
Staying disease-free after treatment; sustaining a cure (relapse)											
Studies not identified to evaluate this outcome											
Case holding so the TB patient remains adherent to treatment (default or treatment interruption due to non-adherence)											
9	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	Serious[9]	Possible	6/72 (8%)	9/53 (17%)	HR 0.48 (0.18, 1.31)	⊕○○○	Critical
Population coverage or access to appropriate treatment of drug resistant TB – not measured											
Studies not identified to evaluate this outcome											
Smear conversion during treatment											
5	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[9]	Possible	11/13 (85%)	15/23 (65%)	HR 2.21[6] (0.97, 5.04)	⊕⊕○○	Critical
Culture conversion during treatment											
6	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[9]	Possible	28/71 (39%)	24/56 (43%)	HR1.04 (0.61, 1.80)	⊕○○○	Critical
Accelerated detection of drug resistance											
not evaluated in the context of our question											
Avoid unnecessary MDR-TB treatment											
Studies not identified to evaluate this outcome											
Population coverage or access to diagnosis of drug resistant TB											
not evaluated in the context of our question											
Prevention or interruption of transmission of DR-TB to other people, including other patients, health care workers											
Studies not identified to evaluate this outcome											
Shortest possible duration of treatment											
Studies not identified to evaluate this outcome											

Continues..

Continued

Quality assessment							No of patients		Effect		
No of studies	Design	Limitations[1]	Inconsistency[2]	Indirectness	Imprecision	Publication bias[3]	ART use	No ART Use	Relative (95% CI)	Quality	Importance
Avoiding toxicity and adverse reactions from TB drugs											
7	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	Serious[9]	Possible	19/59 (32%)	8/51 (16%)	HR 1.79 (0.79, 4.06)	⊕○○○	Important
Cost to patient, including direct medical costs as well as others such as transportation, lost wages due to disability											
Studies not identified to evaluate this outcome											
Resolution of TB signs and symptoms; ability to resume usual life activities											
9	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	No serious imprecision	Possible	34/86 (40%)	9/72 (13%)	HR 2.59[6] (1.34, 5)	⊕⊕○○	Important
Interaction of TB drugs with non-TB medications											
5	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	very serious[5]	Possible			–	⊕○○○	Important
Cost to the TB control programme											
Studies not identified to evaluate this outcome											

ART = antiretroviral therapy; CI = confidence interval; DR-TB = drug-resistant tuberculosis; HR = hazard ratio; MDR-TB = multidrug-resistant tuberculosis; XDR-TB = extensively drug-resistant tuberculosis

[1] All studies were longitudinal cohort studies in which use of second-line drugs was based on drug susceptibility testing. Heterogeneity in the measurement of exposures and outcomes (including CD4-cell count, definition of cure and record of adverse events) may have limited this analysis.

[2] Our analysis included patient-level data from 10 separate studies. We adjusted for baseline CD4-cell count and stratified by drug resistance pattern, and reported these results only when they differed from the unadjusted or unstratified results. To assess inter-study heterogeneity we conducted a likelihood ratio test to test the hypothesis that the magnitude or direction of a given association depended on which study the patients came from. We downgraded for inconsistency when the result of this test was statistically significant at the 0.05 level. In cases when the number of patients per study was too small for the likelihood test for interaction to be valid, we downgraded for inconsistency only when patients in one comparison group came from an entirely different study than that of the referent group. We did not downgrade for heterogeneity in outcomes of cure, adverse events and default, because we already downgraded for variation in definition as noted in 4.

[3] We did not restrict our search to English language articles, and contacted authors of manuscripts published in French, German, Italian, Russian and Spanish. In all cases, we contacted authors on at least three separate occasions to invite them to contribute to the analysis. In addition, we contacted other groups, including Médecins Sans Frontières, the United States Centers for Disease Control and Prevention, and Partners In Health, to ascertain if other data were available. Two authors lost patient data from previous publications and 60 authors did not respond to our query. Therefore, it is possible that authors did not share unpublished data for our analysis, and it is also possible that negative studies on ART and drug-resistant TB were not published.

[4] We subtracted 1 point for cure, default and adverse event analysis due to heterogeneity in definition of the outcome. Regarding cure, 6/8 MDR-TB studies defined cure according to the Laserson criteria, while 2 MDR-TB studies, and studies with non-MDR-TB/XDR-TB drug resistance defined cure as 2 negative cultures in the final month of therapy with clinical response. Default and adverse events were variably defined.

[5] When adjusting for initial CD4-cell count HR for cure was 2.93 (0.98, 8.69).

[6] We upgraded the quality of the evidence by 1 if the HR estimate was <0.5 or >2.0.

[7] Few events and participants limited ability to calculate a HR estimate of risk (i.e. no events observed in the ART or no ART group).

[8] When adjusting for initial CD4 count, HR for death was 0.23 (0.12,0.46).

[9] 95% CI for the HR crosses 1.

Table 2: Quality assessment of timing of ART							No of patients		Effect	Quality	Importance
No of studies	Design	Limitations[1]	Inconsistency	Indirectness	Imprecision	Publication bias[2]	Early ART use	Later ART use	Relative (95% CI)		
Prompt initiation of appropriate treatment – timing of initiation of ART and death											
4	Observational studies	No serious limitations[3]	No serious inconsistency	No serious indirectness	Very serious[4,5]	Possible	1/8[6,7] (13%)	2/10 (20%)	HR 0.67 (0.06, 7.17)	⊕○○○	Critical
Prompt initiation of appropriate treatment – timing of initiation of ART and adverse events											
4	Observational studies	Serious[8]	No serious inconsistency	No serious indirectness	Very serious[4,5]	Possible	1/8[6,7] (13%)	1/10 (10%)	HR 0.82 (0.07, 9.59)	⊕○○○	Critical
Prompt initiation of appropriate treatment – timing of initiation of ART and death											
4	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Very serious[4,5]	Possible	2/14[7,9] (14%)	1/4 (25%)	HR 0.69 (0.08, 5.97)	⊕○○○	Critical

ART = antiretroviral therapy; CI = confidence interval; DR-TB = drug-resistant tuberculosis; HR = hazard ratio; MDR-TB = multidrug-resistant tuberculosis; XDR-TB = extensively drug-resistant tuberculosis

[1] Only one study reported timing of ART and second-line TB regimens to allow us to evaluate whether timing of ART initiation is associated with a difference in outcomes of interest.

[2] We did not restrict our search to English language articles and contacted authors of manuscripts published in French, German, Italian, Russian and Spanish. In all cases, we contacted authors on at least three separate occasions to invite them to contribute to the analysis. In addition, we contacted other groups, including Médecins Sans Frontières, the United States Centers for Disease Control and Prevention, and Partners In Health, to ascertain if other data were available. Two authors lost patient data from previous publications and 60 authors did not respond to our query. Therefore, it is possible that authors did not share unpublished data for our analysis, and it is also possible that negative studies on ART and drug-resistant TB were not published.

[3] We subtracted 1 point for cure, default and adverse event analysis due to heterogeneity in definition of the outcome. Regarding cure, 6/8 MDR-TB studies defined cure according to the Laserson criteria, while 2 MDR-TB studies, and studies with non-MDR-TB/XDR-TB drug resistance defined cure as 2 negative cultures in the final month of therapy with clinical response. Default and adverse events were variably defined.

[4] 95% CI for the HR crosses 1.

[5] because of the low number of events and inability to evaluate timing of ART use further we downgraded our imprecision to very serious.

[6] Comparison of ART initiation before or after the first two weeks of TB therapy.

[7] We could not estimate HR for comparison of ART initiation before or after the intensive phase of therapy (i.e. no events observed in the ART or no ART group). Only two studies (12 total patients recorded timing of intensive phase).

[8] Our analysis included patient-level data from 10 separate studies. We adjusted for baseline CD4 count and stratified by drug resistance pattern, and reported these results only when they differed from the unadjusted or un-stratified results. To assess inter-study heterogeneity we conducted a likelihood ratio test to test the hypothesis that the magnitude or direction of a given association depended on which study the patients came from. We downgraded for inconsistency when the result of this test was statistically significant at the 0.05 level. In cases when the number of patients per study was too small for the likelihood test for interaction to be valid, we downgraded for inconsistency only when patients in one comparison group came from an entirely different study than that of the referent group.

[9] Comparison of ART initiation before or after the first 8 weeks of therapy.

Table 3: Quality assessment of ART and number of effective drugs							No of patients		Effect	Quality	Importance
No of studies	Design	Limitations[1]	Inconsistency[2]	Indirectness	Imprecision	Publication bias[3]	≥4 drugs	≤3 drugs	Relative (95% CI)		
A1. Number of effective drugs[4] – cure in ART users											
9	Observational studies	Serious[5]	No serious inconsistency	No serious indirectness	No serious imprecision	Possible	4/13 (31%)	29/59 (49%)	HR 0.53 (0.27, 1.02)	⊕○○○[6]	Other
A2. Number of effective drugs[4] – cure in ART non-users											
9	Observational studies	Serious[5]	No serious inconsistency	No serious indirectness	Serious[7]	Possible	2/18 (11%)	5/35 (14%)	HR 0.26 (0.06, 1.10)	⊕○○○	Other
A3. Number of effective drugs[4] – death in ART users											
10	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[7]	Possible	5/13 (38%)	29/111 (26%)	1.38 (0.6, 3.16)	⊕○○○	Other
A4. Number of effective drugs[4] – death in ART non-users											
10	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[7]	Possible	11/18 (61%)	37/65 (57%)	0.81 (0.42, 1.55)	⊕○○○	Other
A5. Number of effective drugs[4] – adverse events in ART users											
9	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	No serious imprecision	Possible	6/8 (75%)	13/53 (25%)	2.55 (1.27. 5.11)	⊕○○○[8]	Other
A6. Number of effective drugs[4] – adverse events in ART non-users											
9	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	Serious[7]	Possible	4/17 (24%)	11/42 (26%)	0.55 (0.17, 1.65)	⊕○○○	Other

ART = antiretroviral therapy; CI = confidence interval; DST = drug susceptibility testing; HR = hazard ratio; MDR-TB = multidrug-resistant tuberculosis; XDR-TB = extensively drug-resistant tuberculosis

[1] All studies were longitudinal cohort studies in which use of second-line drugs was based on drug susceptibility testing. Heterogeneity in the measurement of exposures and outcomes (including CD4-cell count, definition of cure and record of adverse events) may have limited this analysis.

[2] Our analysis included patient-level data from 10 separate studies. We adjusted for baseline CD4 count and stratified by drug resistance pattern, and reported these results only when they differed from the unadjusted or un-stratified results. To assess inter-study heterogeneity we conducted a likelihood ratio test to test the hypothesis that the magnitude or direction of a given association depended on which study the patients came from. We downgraded for inconsistency when the result of this test was statistically significant at the 0.05 level. In cases where the number of patients per study was too small for the likelihood test for interaction to be valid, we downgraded for inconsistency only when patients in one comparison group came from an entirely different study than that of the referent group. We did not downgrade for heterogeneity in outcomes of cure, adverse events and default, because we already downgraded for variation in definition as noted in 5.

[3] We did not restrict our search to English language articles and contacted authors of manuscripts published in French, German, Italian, Russian and Spanish. In all cases, we contacted authors on at least three separate occasions to invite them to contribute to the analysis. In addition, we contacted other groups, including Médecins Sans Frontières, the United States Centers for Disease Control and Prevention, and Partners In Health, to ascertain if other data were available. Two authors lost patient data from previous publications and 60 authors did not respond to our query. Therefore, it is possible that authors did not share unpublished data for our analysis, and it is also possible that negative studies on ART and drug-resistant TB were not published.

[4] Number of effective drugs was defined as number of drugs in the TB treatment regimen to which DST confirmed sensitivity.

[5] We subtracted 1 point for cure, default and adverse event analysis due to heterogeneity in definition of the outcome. Regarding cure, 6/8 MDR-TB studies defined cure according to the Laserson criteria, while 2 MDR-TB studies, and studies with non-MDR-TB/XDR-TB drug resistance defined cure as 2 negative cultures in the final month of therapy with clinical response. Default and adverse events were variably defined.

[6] Thought to be affected by drug susceptibility. When restricting this analysis to MDR-TB patients, cure was no longer associated with <3 effective drugs – HR 0.83 (0.38, 1.8).

[7] 95% CI for the HR crosses 1.

[8] When restricted to MDR-TB patients, adverse events were still significantly associated with >4 drugs – HR 2.49 (1.28, 4.87). However, this association was no longer significant when looking at ART and total number of drugs (for >5 total drugs in ART users, HR for adverse events were HR 1.2, 0.56, 2.58). We therefore downgraded this observation to very low.

Table 3: Quality assessment of ART use and specific drug groups							No of patients		Effect	Quality	Importance
No of studies	Design	Limitations[1]	Inconsistency[2]	Indirectness	Imprecision	Other	Drug group exposure	No exposure to drug group	Relative (95% CI)		
B1. Group 1 drugs[3] – cure in ART users											
8	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	No serious imprecision	None	29/57 (51%)	4/15 (27%)	HR 3.32 (1.01, 10.88)	⊕⊕○○[5]	Other
B3. Group 1 drugs – death in ART users											
9	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[6]	None	26/107 (24%)	8/17 (47%)	HR 0.57 (0.23, 1.48)	⊕○○○	Other
B5. Group 1 drugs – adverse events in ART users											
6	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	Serious[6]	None	14/47 (30%)	5/12 (42%)	HR 1.11 (0.56, 2.22)	⊕○○○	Other
B1. Group 2 drugs[7] – cure in ART users											
8	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	Serious[6]	None	7/23 (30%)	26/49 (53%)	HR 0.43 (0.18, 1.0)[8]	⊕○○○	Other
B3. Group 2 drugs[7] – death in ART users											
9	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[6]	None	17/72 (24%)	17/52 (33%)	HR 0.75 (0.39, 1.45)[8]	⊕○○○	Other
B6. Group 2 drugs[7] – adverse events in ART users											
6	Observational studies	Serious[4]	No serious inconsistency	No serious indirectness	Serious[6]	None	3/12 (25%)	16/47 (34%)	HR 0.93 (0.35, 2.5)[8]	⊕○○○	Other
B1. Group 3 drugs[9] – cure in ART users											
8	Observational studies	Serious[4]	Serious	No serious indirectness	Serious[6]	None	31/65 (48%)	2/7 (29%)	HR 1.11 (0.71, 1.71)	⊕○○○	Other
B3. Group 3 drugs[9] – death in ART users											
9	Observational studies	No serious limitations	No serious inconsistency	No serious indirectness	Serious[6]	None	22/80 (28%)	12/44 (27%)	HR 0.81 (0.4, 1.66)	⊕○○○	Other
B6. Group 3 drugs[9] – adverse events in ART users											
6	Observational studies	Serious[4]	Serious	No serious indirectness	very serious[10]	None	19/55 (35%)	0/4 (0%)	–	⊕○○○	Other

ART = antiretroviral therapy; CI = confidence interval; DST = drug susceptibility testing; HR = hazard ratio; MDR-TB = multidrug-resistant tuberculosis; XDR-TB = extensively drug-resistant tuberculosis

1 All studies were longitudinal cohort studies in which use of second line drugs was based on drug sensitivity testing. Heterogeneity in the measurement of exposures and outcomes (including CD4-cell count, definition of cure and record of adverse events) may have limited this analysis.

2 Our analysis included patient level data from 10 separate studies. We adjusted for baseline CD4-cell count and stratified by drug resistance pattern and reported these results only when they differed from the unadjusted or unstratified results. To assess inter-study heterogeneity we conducted a likelihood ratio test to test the hypothesis that the magnitude or direction of a given association depended on which study the patients came from. We downgraded for inconsistency when the result of this test was statistically significant at the 0.05 level. In cases when the number of patients per study was too small for the likelihood test for interaction to be valid, we downgraded for inconsistency only when patients in one comparison group came from an entirely different study than that of the referent group. We did not downgrade for heterogeneity in outcomes of cure, adverse events and default, because we already downgraded for variation in definition as noted in 4.

3 Group 1 drugs defined as, rifampicin, isoniazid, ethambutol and pyrazinamide

4 We subtracted 1 point for cure, default, and adverse event analysis due to heterogeneity in definition of the outcome. Regarding cure, 6/8 MDR-TB studies defined cure according to the Laserson criteria, while 2 MDR-TB studies, and studies with non-MDR-TB/XDR-TB drug resistance defined cure as 2 negative cultures in the final month of therapy with clinical response. Default and adverse events were variably defined.

5 Group 1 drugs were associated with a HR of 2.37 (0.66, 8.54) for cure and 0.44 (0.16, 1.19) in MDR-TB patients receiving ART. In XDR-TB patients, no deaths were observed in patients receiving ART but no group 1 drugs, so HR could not be calculated in this group. Cure (HR 0.77 (0.08, 7.5)) was not different in patients receiving group 1 drugs but no ART, but death was less common and approached significance at 95% CI (HR 0.53 (0.27, 1.01).

6 95% CI for the HR crosses 1.

7 Group 2 drugs defined as kanamycin, amikacin, capreomycin and viomycin.

8 We did not observe a difference in cure rates in patients receiving group 2 drugs without ART (HR 1.79 (0.34, 9.37)). In addition, we saw similar trends when excluding kanamycin (HR 0.47 for cure with ART and a group 2 drug, HR 1.76 for cure with ART but no group 2 drug) and when looking only at MDR-TB patients (HR 0.55 (0.21, 1.46) for cure in MDR-TB patients with ART and a group 2 drug vs ART and no group 2 drug and HR 0.46 (0.17, 1.3) for death with ART and a group 2 drug vs ART and no group 2 drug). Default showed a trend toward increased occurrence if receiving ART and a group 2 drug vs ART and no group 2 drug (HR 4.43 (0.88, 22.31)).

9 Group 3 drugs defined as levofloxacin, ciprofloxacin, gatifloxacin, moxifloxacin and ofloxacin.

10 Few events and participants limited ability to calculate a HR ratio estimate of risk (i.e. no events observed in the ART or no ART group).

Question 7: Among MDR-TB patients, is ambulatory therapy compared to inpatient treatment more or less likely to lead to the outcomes of interest?

							Indirect comparison of generalized cost-effectiveness results[1]								
							Outpatient model of care[2]			Control: Inpatient model of care					
	Design	Limitations	Inconsistency	Indirectness	Imprecision	Other considerations	Number of studies [patients]	Resistance profile (# drugs:% patients)	Resource use/ cost (2005 I$)[3,4]	Number of studies [patients]	Resistance profile (# drugs: % patients)	Resource use/ cost (2005 I$)[3,4]	Absolute effect/ difference[3]	Relative effect/ difference[3]	Quality
Viewpoint: health system															
Resource use per patient[5]	Observational	No serious limitations	No serious inconsistency	No serious indirectness	No serious imprecision	None	2 [415]	2:8 3:26 4:38 ≥5:28	bed-days: 0–7 hospital visits: 0–18 clinic visits: 253–450	2 [249]	2:1 3:14 4:26 ≥5:59	Bed-days: 192–321 Hospital visits: 0–250 Clinic visits: 85–171	Bed-days: outpatient 185–321 lower	Bed-days: outpatient 96–100% lower	⊕⊕○○[6]
Cost per patient	Observational	No serious limitations	No serious inconsistency	Serious indirectness[7]	No serious imprecision	None	2 [415]	2:8 3:26 4:38 ≥5:28	Diagnosis:[8] 125 Drugs: 1914 GHS:[9] 3400 Other: 5687 Total: 11126 (3201–29556)	2 [249]	2:1 3:14 4:26 ≥5:59	Diagnosis:[8] 251 Drugs: 4838 GHS:[9] 27068 Other: 3882 Total: 36039 (8349–103127)	Outpatient 24912 (4152–79315) better	Outpatient 63% (33–85%) better	⊕○○○[10]
Cost per compliant[11] patient	Observational	No serious limitations	No serious inconsistency	Serious indirectness[7]	No serious imprecision	None	2 [415]	2:8 3:26 4:38 ≥5:28	12854 (3843–34037)	2 [249]	2:1 3:14 4:26 ≥5:59	40834 (9475–116820)	Outpatient 28119 (4616–89758) better	Outpatient 63% (33–85%) better	⊕○○○
Cost per death averted[12]	Observational	No serious limitations	No serious inconsistency	Serious indirectness[13]	No serious imprecision	None	2 [415]	2:8 3:26 4:38 ≥5:28	17105 (4431–48540)	2 [249]	2:1 3:14 4:26 ≥5:59	48458 (10722–143102)	Outpatient 33099 (3821–109169) better	Outpatient 62% (22–86%) better	⊕○○○
Cost per DALY[14] averted	Observational	No serious limitations	No serious inconsistency	Serious indirectness[13]	No serious imprecision	None	2 [415]	2:8 3:26 4:38 ≥5:28	589 (137–1689)	2 [249]	2:1 3:14 4:26 ≥5:59	1859 (401–5445)	Outpatient 1271 (146–4173) better	Outpatient 62% (22–86%) better	⊕○○○
Viewpoint: patient[15]															
Resource use per patient[5]	Observational	Serious limitations[16]	No serious inconsistency	No serious indirectness	No serious imprecision	None	2 [415]	2:8 3:26 4:38 ≥5:28	Hours: 365–468	2 [249]	2:1 3:14 4:26 ≥5:59	Hours: 3158–5429	Outpatient 2690–5064 better	Outpatient 85–93% better	⊕○○○

1. No two models of MDR-TB care are directly compared in the included studies and no two alternatives are the same. In order to (indirectly) compare cost per death averted and cost per DALY averted across the studies, we modelled a standard alternative of no intervention based on a standard distribution of death rate in the absence of second-line treatment and an assumption of zero cost. We re-calculate cost-effectiveness with regard to this null set for each of the studies. The results are then (partially) generalized for setting, using a standard distribution of DALYs averted per death averted and a global distribution of unit costs [adjusted for inflation, purchasing power parity (PPP), and Gross Domestic Product (GDP) per capita, as appropriate]. The results are not corrected for differences in the basic demography and epidemiology of disease across settings (See Footnote 13). The indirect comparison therefore assumes that effect sizes (death rates) achieved in one setting can be replicated in any other given setting by exactly reproducing the model of care–at local costs.
2. For the purposes of this review, the model of care described by a study is classified as "outpatient" if the average duration of hospitalization among the cohort of patients is no more than seven days. Three of the four included studies had some mix of inpatient and outpatient care; only in one study was the model of care entirely outpatient-based. Within the outpatient models of care, there were no studies looking at community-based care.
3. Numbers in parentheses are the 5th and 95th percentiles, representing the plausible range of values obtained in probabilistic, multivariate uncertainty analyses.
4. A 2005 international dollar (I$) is worth in any given country what 1 US$ could have bought in the United States of America in 2005.
5. Ranges in resource use per patient are lowest and highest cohort averages (mean or median) from across all of the included studies.
6. Low quality: Further research is likely to have an impact on the estimate of effect.
7. Results for the outpatient model of care represent a mix of standardized (298 patients) and individualized regimens (117 patients); whereas results for the inpatient model of care represent individualized regimens only (all 249 patients). The standardized regimen would today be considered substandard. The standardized regimen described by Suarez et al. (2002) is a 18-month daily regimen consisting of kanamycin (1 g injectable) for the first three months, ciprofloxacin (1 g orally), ethionamide (750 mg orally), pyrazinamide (1500 mg orally), and ethambutol (1200 mg orally). If we assumed the cost of an individualized regimen, the cost per patient under the outpatient model of care would increase by 19% (10%-38%), but the relative effect would still be 54% (13%-82%) less than the inpatient-based models of care.
8. Diagnosis costs include smear microscopy, culture, and drug-susceptibility testing using culture; none of the included studies were conducted in sites where or at a time when molecular or genetic testing for MDR-TB was available.
9. General Health-care Services (GHS): the cost associated with utilization of general health-care services (bed-days, hospital visits and clinic visits).
10. Very low quality: We are very uncertain about the estimates of effect.
11. Includes all patient outcomes except default.
12. Cost per death averted per index case and cost per DALY averted include transmission benefits (i.e. reductions in the number of deaths and DALYs from secondary cases infected by the index cases), and well as long-term deaths among defaults and relapses.
13. We know that there are differences between the study settings in terms of basic demography and epidemiology of disease, not least with respect to the resistance profile (see column "Resistance profile"). The fact that there is a higher proportion of patients showing resistance to more than five drugs in the studies of inpatient models of care may confound the results in favor of outpatient-based models. At the same time, the results may be confounded in favor of inpatient-based models, since the outcomes of the outpatient-based models reflect (in part) a substandard regimen. See Footnote 7. It is unclear which confounder predominates.
14. Disability-adjusted life-year (DALY).
15. Only costs of resources used to access the health intervention are included (e.g. transportation, nutrition); within these access costs, time losses are described, but not costed. Productivity losses due to illness are not considered.
16. None of the studies describes losses times in units. We estimate time losses at 16 hours per bed-day, 1 hour per hospital visit and 0.5 hours per clinic visit.

WHO/HTM/TB/2011.6b