# Structured approaches for the screening and diagnosis of childhood tuberculosis in a high prevalence region of South Africa

Mark Hatherill,[a] Monique Hanslo,[a] Tony Hawkridge,[b] Francesca Little,[c] Lesley Workman,[a] Hassan Mahomed,[a] Michele Tameris,[a] Sizulu Moyo,[a] Hennie Geldenhuys,[a] Willem Hanekom,[a] Lawrence Geiter[d] & Gregory Hussey[a]

**Objective** To measure agreement between nine structured approaches for diagnosing childhood tuberculosis; to quantify differences in the number of tuberculosis cases diagnosed with the different approaches, and to determine the distribution of cases in different categories of diagnostic certainty.

**Methods** We investigated 1445 children aged < 2 years during a vaccine trial (2001–2006) in a rural South African community. Clinical, radiological and microbiological data were collected prospectively. Tuberculosis case status was determined using each of the nine diagnostic approaches. We calculated differences in case frequency and categorical agreement for binary (tuberculosis/not tuberculosis) outcomes using McNemar's test (with 95% confidence intervals, CIs) and Cohen's kappa coefficient ($K$).

**Findings** Tuberculosis case frequency ranged from 6.9% to 89.2% (median: 41.7). Significant differences in case frequency ($P < 0.05$) occurred in 34 of the 36 pair-wise comparisons between structured diagnostic approaches (range of absolute differences: 1.5–82.3%). Kappa ranged from 0.02 to 0.71 (median: 0.18). The two systems that yielded the highest case frequencies (89.2% and 70.0%) showed fair agreement ($K$: 0.33); the two that yielded the lowest case frequencies (6.9% and 10.0%) showed slight agreement ($K$: 0.18).

**Conclusion** There is only slight agreement between structured approaches for the screening and diagnosis of childhood tuberculosis and high variability between them in terms of case yield. Diagnostic systems that yield similarly low case frequencies may be identifying different subpopulations of children. The study findings do not support the routine clinical use of structured approaches for the definitive diagnosis of childhood tuberculosis, although high-yielding systems may be useful screening tools.

*Une traduction en français de ce résumé figure à la fin de l'article.* *Al final del artículo se facilita una traducción al español.* الترجمة العربية لهذه الخلاصة في نهاية النص الكامل لهذه المقالة.

## Introduction

Despite the scale of the worldwide tuberculosis epidemic, the disease remains very difficult to diagnose in children, especially in regions with limited resources.[1] Childhood tuberculosis is often paucibacillary and the diagnosis rests on interpretation of chest radiograph findings and non-specific symptoms and signs.[1] Improving diagnostic accuracy and reliability is key to integrating childhood tuberculosis into national control programmes, and the World Health Organization (WHO) has thus prioritized diagnostic criteria for childhood tuberculosis.[2] Objective, reproducible tuberculosis diagnosis will also be pivotal for defining end-points in trials of new tuberculosis vaccines.[3] The need for accurate diagnosis is felt most acutely among younger children, who contribute substantially to the burden of tuberculosis in high prevalence regions.[4–7]

Routine clinical use of a structured diagnostic approach that is unsuited to a particular setting can result in systematic errors in estimating the burden of tuberculosis and in patient management. It follows that regional guidelines for screening and diagnosis of childhood tuberculosis should be tailored to their epidemiological context.

The relative merits of existing structured diagnostic approaches are debatable.[5,8–17] Hesseling et al. reviewed 16 such approaches and noted that few of the scoring systems, algorithms and classifications for the screening and diagnosis of childhood tuberculosis have been validated against a gold standard. Most have been developed for hospital-based studies and their usefulness in community settings is relatively unknown.[5,18–21] Some have suggested that structured diagnostic approaches should be used only as screening tools to select children for further investigation,[9,10] while others have proposed a simplified case definition of childhood tuberculosis, based on cardinal symptoms, as an alternative to complex diagnostic systems.[1,22]

Existing structured approaches to childhood tuberculosis provide a logical and reproducible basis for diagnosis based on clinical acumen, which Cundall termed "the art of the possible".[23] However, we hypothesized that commonly used, structured approaches for screening and diagnosing childhood tuberculosis may show poor agreement and yield highly variable case frequency results. The objectives of this paper were to quantify the tuberculosis case frequencies obtained by means of nine different diagnostic systems, to assess agreement between systems, and to offer possible explanations for discordant findings.

## Methods

This analysis is based on data collected during a bacille Calmette-Guérin (BCG) vaccine trial conducted by the South African Tuberculosis Vaccine Initiative (SATVI) from March

[a] School of Child and Adolescent Health, University of Cape Town, Anzio Road, Cape Town, 7925, South Africa.
[b] Aeras Global TB Vaccine Foundation, Rockville, United States of America (USA).
[c] Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa.
[d] Otsuka Pharmaceutical Development and Commercialization Inc., Rockville, USA.
Correspondence to Mark Hatherill (e-mail: mark.hatherill@uct.ac.za).
(*Submitted: 9 January 2009 – Revised version received: 11 September 2009 – Accepted: 7 October 2009 – Published online: 29 December 2009*)

Table 1. **Results of chest radiograph assessment by three independent paediatric reviewers, grouped by certainty of tuberculosis diagnosis, South Africa, 2001–2006**

| Diagnostic certainty[a] | Reviewer 1 | | Reviewer 2 | | Reviewer 3 | | Final classification | |
|---|---|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % | No. | % |
| Highly likely to have tuberculosis | 16 | 1.1 | 29 | 2.0 | 171 | 11.8 | | |
| Likely to have tuberculosis | 20 | 1.4 | 38 | 2.6 | 323 | 22.4 | | |
| Suspected of having tuberculosis | 124 | 8.6 | 145 | 10.0 | 242 | 16.7 | | |
| **Positive** | **160** | **11.1** | **212** | **14.6** | **736** | **50.9** | **271** | **18.8** |
| Inconclusive | 45 | 3.1 | 35 | 2.4 | 82 | 5.7 | | |
| Abnormal but not tuberculosis | 102 | 7.1 | 139 | 9.6 | 312 | 21.6 | | |
| Normal | 1038 | 71.8 | 778 | 53.9 | 59 | 4.1 | | |
| **Negative** | **1185** | **82.0** | **952** | **65.9** | **453** | **31.4** | **1174** | **81.2** |
| Not read | 100 | 6.9 | 281 | 19.5 | 256 | 17.7 | | |
| **Total** | **1445** | **100** | **1445** | **100** | **1445** | **100** | **1445** | **100** |

[a] "Highly likely to have tuberculosis", "likely to have tuberculosis" and "suspected of having tuberculosis" were classified as positive; "inconclusive", "abnormal but not tuberculosis" and "normal" were classified as negative. Final chest radiograph classification was determined by agreement of at least two reviewers.

2001 to August 2006 near Cape Town, South Africa (clinical trials identifier: NCT00242047).[8] In the Boland-Overberg region of South Africa, tuberculosis incidence among children aged < 2 years was estimated as > 3000 cases per 100 000 in 2006.[6,8,24] In the trial, which compared the vaccine efficacy obtained with percutaneous versus intradermal Tokyo-172 BCG, 11 680 neonates were followed up for a minimum of 2 years after vaccination.[8]

Children in the community suspected of having tuberculosis due to a history of contact with an adult case or to the presence of symptoms compatible with the disease were identified by a regional surveillance system. All such children underwent comprehensive radiological and bacteriological investigation, even if they had no symptoms. The presence and duration of cough, wheezing, fever or weight loss; the response to antibiotics; and the proximity of contact with an adult having tuberculosis (mother, other person within the household, person outside of the household), were recorded. Human immunodeficiency virus (HIV) status was determined by a rapid antibody test and, if the result was positive, confirmatory polymerase chain reaction (PCR) was performed as well. Tuberculin skin tests included both Mantoux and Tine. Chest radiographs (anteroposterior and lateral) were reviewed by three paediatricians and classified in terms of the likelihood of tuberculosis (Table 1). Two consecutive, paired gastric lavages and induced sputum samples were obtained for smear microscopy and culture of *Mycobacterium tuberculosis* using mycobacteria growth indicator tubes (Becton Dickinson and Co., Sparks, MD, United States of America). A diagnostic algorithm was developed, based on approaches described by Cundall and WHO, for objective post hoc determination of tuberculosis status as the trial end-point.[21,23] The decision to start tuberculosis treatment was made on discharge by the attending clinician on the basis of all available results, independent of the assigned trial end-point.

A protocol-specified objective was to compare the structured approaches used to diagnose childhood tuberculosis in developing countries with a high prevalence of tuberculosis and limited resources. Diagnostic approaches relevant to sub-Saharan Africa, dating from 1990 onwards, were selected by literature review and expert consultation. Recent modifications were preferred over versions predating the HIV era. Eight structured approaches were compared with the SATVI trial algorithm for tuberculosis case frequency.[8] The country of origin, lineage and type of approach are summarized in Table 2.

Structured diagnostic approaches were categorized as follows:

i) binary, with the diagnosis being simply positive or negative (yes = tuberculosis; no = not tuberculosis);[12,15]
ii) hierarchical, with stratification into categories of diagnostic certainty, such as "definite", "probable", "possible", "unlikely" or "not tuberculosis";[8,14,16] or
iii) numerical, with a score obtained by adding the weighted values assigned to each variable (score ≥ $x$ = tuberculosis).[9–11,13]

Data for the variables used in these diagnostic approaches were collected prospectively during the trial. Missing variables were assigned a zero value. Referenced threshold values were used for the analysis unless cut-off thresholds were unspecified, and trial algorithm values were used as the default.[8] To standardize reporting, the terms for the hierarchical categories of diagnostic certainty were "unlikely/not", "possible", "probable", and "definite" tuberculosis.[11,13,14,16] Details of the various diagnostic approaches are provided in Appendix A (Available at: http://vacfa.com/index.php?option=com_content&view=section&layout=blog&id=10&Itemid=10).

The variables required by each system to compute a tuberculosis outcome for each child were programmed using STATA version 10 (StataCorp, Inc., College Station, TX, USA). Tuberculosis cases were defined by:

i) "positive" classification for binary (tuberculosis/not tuberculosis) systems;
ii) "definite", "probable" or "possible" classification for hierarchical systems; or
iii) score ≥ the specified cut-off for numerical scoring systems.

The analysis of binary outcomes compared the nine diagnostic approaches in terms of the number and percentage of tuberculosis cases diagnosed among the children investigated. McNemar's test was used to compare the paired proportions of tuberculosis cases diagnosed with each system. *P*-values were not manipulated to adjust for multiple comparisons. Cohen's kappa coefficient ($K$) was used to examine agreement between individual observations for each system. Weighted $K$

Table 2. **Nine structured approaches for diagnosing childhood tuberculosis**

| Approach | Year | Origin | Source data | Classification | Purpose | Lineage |
|---|---|---|---|---|---|---|
| WHO–Harries[10] | 1996 | International | Clinical | Numerical | Diagnosis | Based on Keith Edwards criteria[18] (Papua New Guinea) |
| Fourie[9] | 1998 | International | Clinical | Numerical | Screening | High tuberculosis prevalence areas[9] |
| Osborne[14] | 1995 | Zambia | Clinical Radiological Bacteriological | Hierarchical | Screening | Adapted from Cundall[23] (Kenya) and WHO[21] |
| Migliori[12] | 1992 | Uganda | Clinical Radiological Bacteriological | Binary | Diagnosis | Derived from Ghidey and Habte[19] (Ethiopia) |
| Stegen–Toledo[13] | 2003 | Peru | Clinical Radiological Bacteriological | Numerical | Diagnosis | Adapted from Stegen–Jones[20] (Chile) |
| MASA[15] | 1996 | South Africa | Clinical Radiological | Binary | Diagnosis | Clinical practice guideline, MASA |
| Stoltz–Donald[16] | 1990 | South Africa | Clinical Radiological Bacteriological | Hierarchical | Screening | Adapted from Cundall[23] (Kenya) and WHO[21] |
| Kibel[11] | 1999 | South Africa | Clinical Radiological Bacteriological | Numerical | Diagnosis | Clinical practice guideline, UCT |
| SATVI vaccine trial algorithm[8] | 2006 | South Africa | Clinical Radiological Bacteriological | Hierarchical | Diagnosis | Adapted from Cundall[23] (Kenya) and WHO[21] |

MASA, Medical Association of South Africa; SATVI, South African Tuberculosis Vaccine Initiative; UCT, University of Cape Town; WHO, World Health Organization.

statistics were calculated for systems with hierarchical classifications. The degree of agreement was defined by the following values of $K$: 0–0.2 = slight; 0.2–0.4 = fair; 0.4–0.6 = moderate; 0.6–0.8 = substantial; and 0.8–1.0 = nearly perfect.[25]

In total, 1869 case episodes involving 1654 children were investigated, and one case episode was selected for each child. Since children older than 2 years were excluded, 1445 children were included in this analysis.

## Results

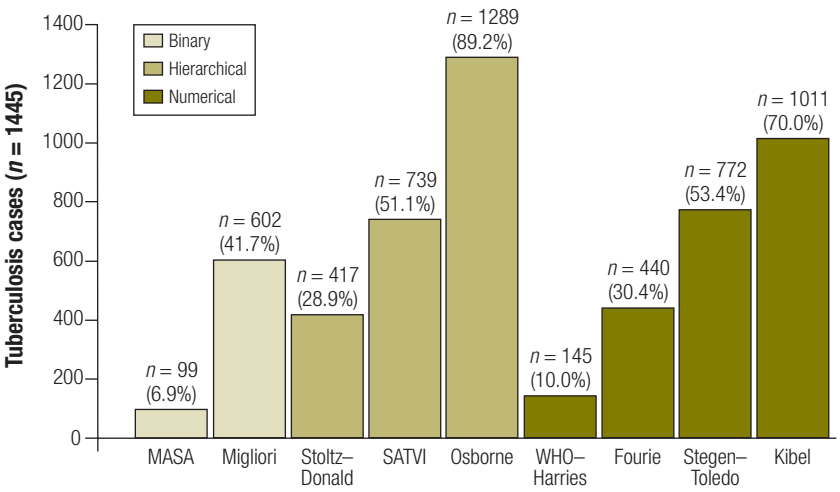
The median age at investigation was 11.4 months (interquartile range: 6.0–17.4). Contact with an adult with tuberculosis was reported for 952 children (65.9%), and 628 children (43.5%) had cough lasting > 2 weeks. Weight was recorded as being 60–80% of expected weight–for–age in 316 (21.9%) children and as being < 60% of expected weight–for–age in 29 children (2.0%). Of the 1445 children studied, 54 (3.7%) tested positive for HIV with enzyme-linked immunosorbent assay, and 28 of these children (1.9%) were confirmed positive for HIV by polymerase chain reaction (PCR) assay. The chest radiograph was compatible with tuberculosis in 271 children (18.8%) and *Mycobacterium tuberculosis* was cultured from induced sputum or gastric lavage in 172 children (11.9%). Treatment for tuberculosis was started by the attending clinician in 611 children (42.3%).

### Comparison of binary outcomes

Fig. 1 illustrates the number and percentage of tuberculosis cases diagnosed with each system. The median tuberculosis case frequency was 41.7% (602 of the 1445 children investigated).

Differences in tuberculosis case frequency are shown in Table 3. The differences were significant ($P < 0.05$) in 34 of 36 possible pair-wise comparisons between the various structured diagnostic approaches. Only the comparisons between the Stegen–Toledo and SATVI approaches and between the Stoltz–Donald and Fourie approaches yielded non-significant differences. The pair-wise

Fig. 1. **Frequency of cases classified as tuberculosis with various scoring systems, with hierarchical and numerical outcomes condensed to a binary "tuberculosis/not tuberculosis" output , South Africa, 2001–2006**




MASA, Medical Association of South Africa; SATVI, South African Tuberculosis Vaccine Initiative; WHO, World Health Organization.

differences in tuberculosis case frequency ranged from 1.5% to 82.3%.

Table 4 summarizes the observed agreement between all structured diagnostic approaches and shows the *K* statistics for binary "tuberculosis/not tuberculosis" outcomes. For the 36 pairwise comparisons, *K* ranged from 0.02 to 0.71 (median *K*: 0.18).

Two systems based on clinical, radiological and bacteriological source data (Osborne and Kibel) generated the highest tuberculosis case frequencies, yet showed only fair agreement. Four systems – MASA, Osborne, Fourie and WHO–Harries – demonstrated poor to fair agreement with all of the structured diagnostic approaches analysed. Notably, two numerical systems – MASA and WHO–Harries– classified the fewest case episodes as tuberculosis, but showed only slight agreement.

### Comparison of hierarchical outcomes

The distribution of diagnoses in categories of ascending diagnostic certainty is illustrated for three hierarchical and two numerical-hierarchical scoring systems (Fig. 2). The distribution of the diagnostic categories assigned by the Osborne and Kibel systems was similar: a bell-shaped curve with most diagnoses grouped in the "possible" and "probable" categories. By contrast, the Stegen–Toledo and Stoltz–Donald systems yielded results with opposite distributions, with most cases in the "not"/"unlikely" or "definite" categories.

Table 5 summarizes the observed agreement and weighted *K* for hierarchical and numerical-hierarchical systems across categories of increasing diagnostic certainty. Hierarchical agreement was nearly perfect between SATVI and Stoltz–Donald, and substantial between Kibel and Osborne.

### Comparison of numerical outcomes

Tuberculosis case frequency ranged from 10.0% to 70.0% across four numerical scoring systems (Kibel, Fourie, WHO–Harries and Stegen–Toledo) when set at the pre-specified threshold (Fig. 3). Relative to the observed distribution of scores, two of the numerical systems (Kibel and Stegen–Toledo) used a low threshold for tuberculosis diagnosis, resulting in case frequencies of 70.0% and 53.4%, respectively. The other two systems (Fourie and WHO–Harries) used a relatively high diagnostic threshold, resulting in case frequencies of only 30.4% and 10.0%.

## Discussion

The most striking finding of this study was the wide variation (6.9–89.2%) in the frequency of tuberculosis cases diagnosed with the nine structured diagnostic systems. The fact that the differences in tuberculosis case frequency were statistically significant for all but two of 36 possible paired comparisons between systems suggests that the burden of childhood tuberculosis in a given population could be under- or overestimated by as much as 82%. The risk of systematic clinical error is clearly high, and excess morbidity or unnecessary treatment may result if an inappropriate diagnostic system is used for routine management. The variability in tuberculosis case frequency also underscores the importance of accurate phenotyping for interpretation of clinical trial end-points; genotypic studies, and studies of immune correlates.

The second major finding is that the systems that yielded the highest and lowest tuberculosis case frequencies, namely the Osborne (89.2%) and Kibel (70.0%) and the MASA (6.9%) and WHO–Harries (10.0%) systems, demonstrated only fair or slight agreement with each other. Although the two outlier systems that generated the lowest results yielded similar tuberculosis case frequencies, the slight agreement suggests that they may be identifying different subpopulations.

In this study, the variation in tuberculosis case frequency observed when different structured diagnostic approaches were used and the relatively poor agreement between systems were more pronounced than previously reported. Edwards et al. retrospectively assessed agreement between clinical scoring systems used to diagnose tuberculosis among 91 children at a hospital in Kinshasa, Democratic Republic of the Congo. The four approaches (Fourie, WHO provisional guidelines, Stegen–Kaplan, and Ghidey–Habte) generated tuberculosis case frequencies ranging from 87% to 96%.[9,19–21] Agreement between systems ranged from fair (*K*: < 0.4) to moderate (*K*: 0.4–0.6).[26] The reason Edwards et al. found less variation in case frequency may be that the study was hospital-based and all children had been diagnosed with tuberculosis on the original Edwards scale.[18,26]

We have also shown marked variation between hierarchical systems in the certainty of the diagnosis of tuberculosis.[13,14] The evaluation of related hierarchical approaches with similar distributions (SATVI and Stoltz–Donald) by weighting *K* for concordant and discordant categories resulted in better agreement than for binary outcomes.[8,16] Although hierarchical and numerical systems that share key variables, such as a positive tuberculin skin test, a positive chest radiograph, and a positive sputum culture (Stegen–Toledo, Stoltz–Donald, and SATVI) showed moderate agreement, other systems with the same common variables showed less agreement and outlying case frequencies (Kibel, Osborne).[8,11,13,14,16] It follows that system structure, weighting of variables and the exact order of Boolean decision-making may be as important as the constituent variables in determining the diagnostic output of each system.

There are several other reasons for the observed variation in tuberculosis case frequency and the relatively poor agreement between diagnostic approaches. They include differences in: (i) the purpose for which the systems were developed (as a screening tool or for definitive diagnosis; for clinical management or to obtain a trial end-point); (ii) clinical setting (community or hospital); (iii) disease severity (mild or severe tuberculosis); and (iv) regional prevalence of tuberculosis and/or HIV infection (low or high). Ideally, for clinical trials a low-yielding diagnostic system should be used to minimize false positives at the expense of lower sensitivity.[8] On the other hand, clinicians might prioritize sensitivity to avoid the potentially fatal consequences of underdiagnosis and delayed treatment.[14,27] Therefore, approaches designed for clinical management, especially to serve as screening tools, might yield higher tuberculosis case frequencies.[9,14,27] Although the SATVI trial algorithm lay in the mid-range of case frequency estimates, in the absence of a gold standard it is not possible to determine which of the nine approaches yielded the most accurate rate of tuberculosis.[8] However, the proportion of children treated for tuberculosis on clinical grounds (42.3%) was almost identical to the median tuberculosis

Table 3. **Differences in case frequency[a] yielded by nine structured approaches for the diagnosis of tuberculosis, South Africa, 2001–2006**

| System | MASA | Migliori | SATVI | Osborne | Stoltz–Donald | Kibel | Fourie | WHO–Harries | Stegen–Toledo | No. (%) diagnosed with tuberculosis |
|---|---|---|---|---|---|---|---|---|---|---|
| **MASA** | | 34.8 (32.3–37.3) | 44.2 (41.7–47.0) | 82.3 (80.3–84.3) | 22.0 (19.8–24.2) | 63.1 (60.5–65.7) | 23.5 (21.0–26.2) | 3.1 (1.3–5.1) | 46.5 (43.9–49.2) | **99 (6.9)** |
| **Migliori** | $P < 0.0001$ | | 9.4 (6.7–12.2) | 47.5 (44.8–50.3) | 12.8 (10.0–15.6) | 28.3 (25.4–31.2) | 11.3 (8.3–14.1) | 31.7 (28.9–34.3) | 11.7 (9.8–13.7) | **602 (41.7)** |
| **SATVI** | $P < 0.0001$ | $P < 0.0001$ | | 38.1 (35.3–40.8) | 22.2 (20.1–24.5) | 18.9 (16.0–21.7) | 20.7 (17.5–23.9) | 41.1 (38.3–43.9) | 2.3 (–0.3–4.9) | **739 (51.1)** |
| **Osborne** | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | | 60.3 (57.7–63.0) | 19.2 (16.9–21.6) | 58.8 (56.0–61.5) | 79.2 (76.9–81.4) | 35.8 (33.1–38.4) | **1289 (89.2)** |
| **Stoltz–Donald** | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | | 41.1 (38.2–44.0) | 1.5 (–1.6–4.8) | 18.9 (16.1–21.6) | 24.5 (22.0–27.2) | **417 (28.9)** |
| **Kibel** | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | | 39.6 (36.5–42.5) | 60.0 (57.3–62.6) | 16.6 (13.8–19.3) | **1011 (70.0)** |
| **Fourie** | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P = 0.344$ | $P < 0.0001$ | | 20.4 (17.9–22.9) | 23.0 (19.8–26.1) | **440 (30.4)** |
| **WHO–Harries** | $P = 0.0009$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | | 43.4 (40.6–46.2) | **145 (10.0)** |
| **Stegen–Toledo** | $P < 0.0001$ | $P < 0.0001$ | $P = 0.083$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | $P < 0.0001$ | | **772 (53.4)** |
| **No. (%) diagnosed with tuberculosis** | **99 (6.9)** | **602 (41.7)** | **739 (51.1)** | **1289 (89.2)** | **417 (28.9)** | **1011 (70.0)** | **440 (30.4)** | **145 (10.0)** | **772 (53.4)** | 1445 (100) |

MASA, Medical Association of South Africa; SATVI, South African Tuberculosis Vaccine Initiative; WHO, World Health Organization.

[a] Absolute differences in proportion (denominator =1445) and corresponding 95% CIs are above diagonal spaces; McNemar's test *P*-values for differences in proportion are below diagonal spaces.

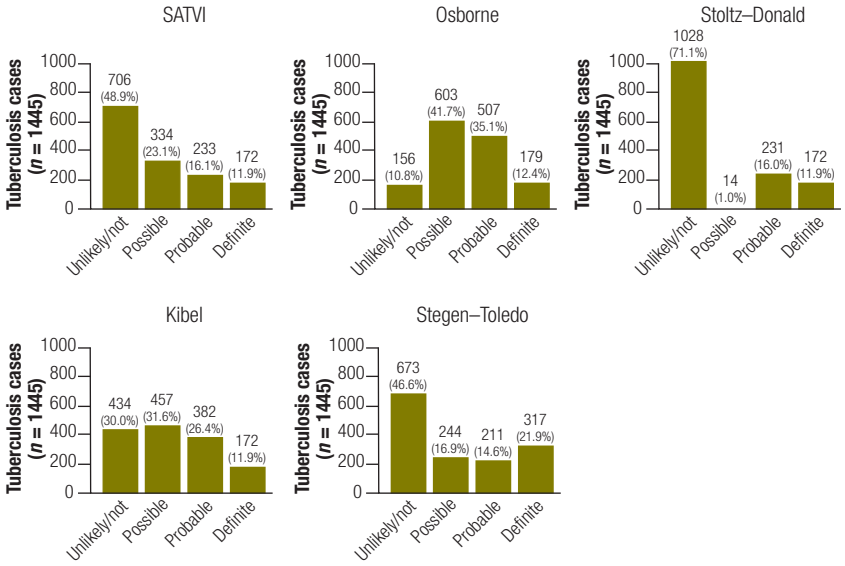
Table 4. **Observed agreement[a] among nine structured approaches for diagnosing tuberculosis, South Africa, 2001–2006**

| System | MASA | Migliori | SATVI | Osborne | Stoltz–Donald | Kibel | Fourie | WHO–Harries | Stegen–Toledo | No. (%) diagnosed with tuberculosis |
|---|---|---|---|---|---|---|---|---|---|---|
| **MASA** | | 65.2 | 55.7 | 17.7 | 78.0 | 36.8 | 69.9 | 87.3 | 53.4 | **99 (6.9)** |
| **Migliori** | 0.19 | | 72.7 | 51.1 | 71.0 | 61.3 | 68.3 | 64.4 | 85.3 | **602 (41.7)** |
| **SATVI** | 0.13 | 0.46 | | 58.3 | 77.7 | 67.3 | 58.8 | 54.9 | 76.4 | **739 (51.1)** |
| **Osborne** | 0.02 | 0.13 | 0.15 | | 37.9 | 76.9 | 38.6 | 20.6 | 61.9 | **1289 (89.2)** |
| **Stoltz–Donald** | 0.31 | 0.38 | 0.56 | 0.07 | | 53.8 | 62.6 | 69.4 | 70 | **417 (28.9)** |
| **Kibel** | 0.06 | 0.27 | 0.34 | 0.33 | 0.21 | | 52.7 | 39.2 | 70.7 | **1011 (70.0)** |
| **Fourie** | 0.09 | 0.32 | 0.18 | 0.06 | 0.10 | 0.18 | | 73.8 | 59.0 | **440 (30.4)** |
| **WHO–Harries** | 0.18 | 0.18 | 0.11 | 0.02 | 0.08 | 0.08 | 0.24 | | 53.4 | **145 (10.0)** |
| **Stegen–Toledo** | 0.12 | 0.71 | 0.53 | 0.19 | 0.42 | 0.38 | 0.20 | 0.12 | | **772 (53.4)** |
| **No. (%) diagnosed with tuberculosis** | **99 (6.9)** | **602 (41.7)** | **739 (51.1)** | **1289 (89.2)** | **417 (28.9)** | **1011 (70.0)** | **440 (30.4)** | **145 (10.0)** | **772 (53.4)** | 1445 (100) |

*K*, kappa statistic; MASA, Medical Association of South Africa; SATVI, South African Tuberculosis Vaccine Initiative; WHO, World Health Organization.

[a] Observed percentage agreement for paired individual observations ($n = 1445$) is above diagonal spaces; $K$ values are below diagonal spaces.

Fig. 2. **Frequency of tuberculosis diagnoses assigned to each category of diagnostic certainty, in order of increasing certainty of tuberculosis, with five hierarchical or hierarchical–numerical systems, South Africa, 2001–2006**




SATVI, South African Tuberculosis Vaccine Initiative

case frequency across all nine diagnostic approaches (41.7%).

## The importance of context

This study was carried out in a community in which children with suspected tuberculosis were identified early, when the disease was probably mild.[8] By contrast, the WHO–Harries system assigns the highest diagnostic weight to chronic illness, severe malnutrition and extrapulmonary tuberculosis, all of which occur more frequently in hospitalized children. It is therefore not surprising that this approach yielded a low tuberculosis case frequency in our context.[10] Similarly, the MASA approach, which requires the presence of the complete triad of symptoms compatible with tuberculosis, as well as a positive tuberculin skin test and a suggestive chest radiograph, is designed as a treatment guideline for hospitalized children.[15] The Osborne approach, which yielded results at the upper extreme of tuberculosis case frequency, was designed in a developing country setting where the index of suspicion for tuberculosis is high. It functions best as a screening tool, since children with suspected or possible tuberculosis are not necessarily treated.[11,14,16] Similarly, the Kibel system is designed to guide initial treatment decisions rather than to establish a definitive diagnosis in resource-limited settings.[11,27] The Fourie system, also designed as a screening tool, yielded one of the lowest tuberculosis case frequencies, which suggests that it may be unsuitable for screening in our epidemiological setting.[9] Some have noted that regional HIV prevalence may affect the performance of a particular diagnostic approach unless HIV infection status is incorporated.[5,8,14] The confounding effect of HIV status on diagnostic decision-making is likely to be greatest in systems that emphasize the non-specific features of malnutrition.[10] Edwards et al. noted that HIV-infected children scored higher on the Keith Edwards scale,[18] a feature that would be common to the WHO–Harries approach. Consequently, the current edition of the WHO's *TB/HIV: a clinical manual* no longer recommends the use of diagnostic scoring systems.[10,26]

Table 5. **Observed agreement[a] among five hierarchical structured approaches for diagnosing tuberculosis, South Africa, 2001–2006**

| System | SATVI | Osborne | Stoltz–Donald | Kibel | Stegen–Toledo |
|---|---|---|---|---|---|
| SATVI | | 79.2 | 92.4 | 81.0 | 81.0 |
| Osborne | 0.48 | | 72.5 | 85.9 | 77.7 |
| Stoltz–Donald | 0.80 | 0.40 | | 76.5 | 78.9 |
| Kibel | 0.51 | 0.60 | 0.43 | | 80.9 |
| Stegen–Toledo | 0.53 | 0.45 | 0.49 | 0.54 | |

*K*, kappa statistic; SATVI, South African Tuberculosis Vaccine Initiative.
[a] Observed percentage agreement for paired individual observations ($n$ =1445) is above diagonal spaces; weighted *K* values are below diagonal spaces.

## Study limitations

This study has several limitations. Investigations were nested within a clinical trial that might not reflect clinical practice in developing regions. Variables were analysed in a standardized fashion that may differ from that used in the original diagnostic systems, and we acknowledge the potential limitations of *K* scores for assessing agreement. Children were younger than 2 years (an age group in which diagnostic imprecision is highest) and the findings may not be applicable to older children with a different disease spectrum. Since the study was community-based and investigations were geared towards pulmonary tuberculosis, there may have been a bias against diagnostic approaches that included features of extra-pulmonary tuberculosis. Furthermore, since all children identified by active case-finding were investigated for tuberculosis, even if they had no symptoms, the discrepancies between clinical, symptom-based and bacteriology-based systems may have been exaggerated. Structured diagnostic approaches were selected on the basis of relevance to the sub-Saharan region. Thus, four of the nine approaches were of South African origin.[8,11,15,16] We acknowledge the existence of other structured approaches for diagnosing childhood tuberculosis, such as the Sant'Anna score, but they were not included in this analysis.[17,28]

## Significance of findings

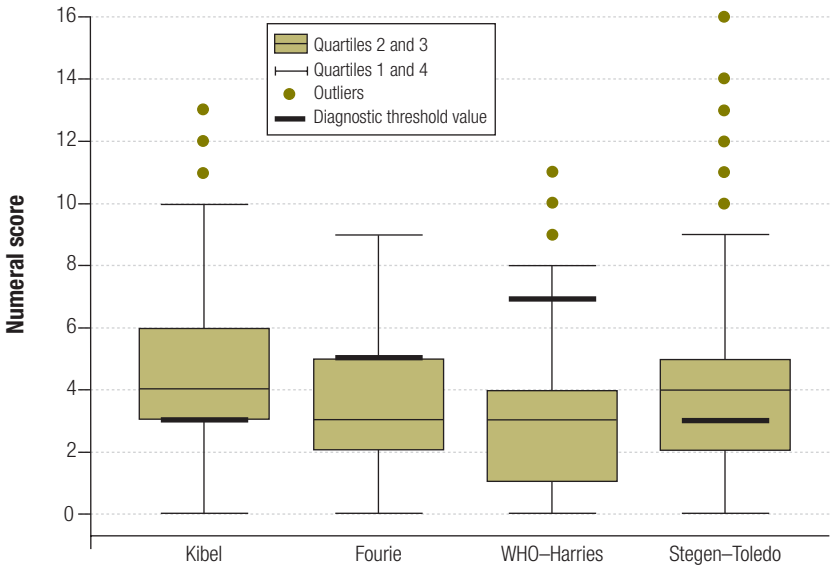
The public health significance of these findings is illustrated by the marked differences in tuberculosis case frequency and the poor agreement between diagnostic systems. Regional tuberculosis control programmes should make an informed decision to advocate a specific approach for the screening and diagnosis of childhood tuberculosis. Clearly, the

study data do not support the routine, uncritical use of any particular diagnostic system for therapeutic decision-making. Some diagnostic approaches may in fact be best suited to specific settings. For example, a high-yielding system, such as Osborne, may be suitable as a screening tool, whereas the low-yielding WHO–Harries system may be most appropriate as a tool for diagnosing severe tuberculosis in regions with a low prevalence of HIV infection.

## Conclusion

Although systems with a moderate case yield are less prone to extreme diagnostic error, the predictive value of any one system cannot be determined in the absence of a gold standard. Any structured approach to estimate tuberculosis case frequency can yield biased results if used in a way that differs from that for which it was originally designed, whether for clinical care or research purposes, screening or definitive diagnosis, mild or severe disease, or in low or high tuberculosis prevalence regions. However, in the absence of validation cohorts, there is limited evidence that these systems would have better diagnostic accuracy in their original settings. The findings of this study should not undermine confidence in existing diagnostic methods. Instead, they should encourage innovative research and critical analysis in the search for improved diagnostics for childhood tuberculosis. ■

Fig. 3. **Distribution of scores ($n$ = 1445) obtained with different numerical scoring systems for the diagnosis of childhood tuberculosis, South Africa, 2001–2006**




### Acknowledgements

We thank the staff of The South African Bacille Calmette-Guérin Trial Team for data collection; Maurice Kibel, John Burgess and Robert Gie for expert radiology review; Suzanne Verver for epidemiological support; and Lyness Matizirofa for statistical support.

**Funding:** The study was supported by the Aeras Global TB Vaccine Foundation, a non-profit organization that aims to develop tuberculosis vaccines.

**Competing interests:** TH and LG are current and previous full time employees of the trial sponsor. The authors have not entered into any agreements that have limited the completion of the research as planned, and they have had full control of all primary data.

## ملخص

### الأساليب المنظمة للتحري عن سل الأطفال وتشخيصه في مناطق انتشاره المرتفع في جنوب أفريقيا

**الغرض** قياس التوافق بين الأساليب المنظمة التسعة لتشخيص سل الأطفال؛ وتحديد كمية الاختلافات في عدد حالات السل المشخصة بهذه الأساليب المختلفة، وتحديد توزيع الحالات في الفئات المختلفة للتأكد من التشخيص.

**الطريقة** استقصى الباحثون 1445 طفلاً أكبر من عمر سنتين أثناء تجربة اللقاح في الأعوام (2001 – 2006) في المجتمعات الريفية في جنوب أفريقيا. جمع الباحثون المعطيات السريرية والشعاعية والمكروبيولوجية. وحدد الباحثون حالات السل باستخدام كل من الأساليب التشخيصية التسعة، وحسبوا الاختلافات في تكرار الحالات والتوافق بين المجموعات في النتائج الثنائية (الإصابة بالسل / وعدم الإصابة بالسل) باستخدام اختبار ماكنيمار McNemar's (مع فاصلات الثقة 95%) ومعامل كوهين كابا Cohen's kappa coefficient.

**الموجودات** تراوح تردد حالات السل من 6.9% إلى 89.2% (الوسيط: 41.7). وكان هناك اختلافات يعتد بها في تكرار الحالات (قوة الاحتمال P أقل من 0.05) في 34 من 36 زوجاً للمقارنة بين الأساليب التشخيصية المنظمة (مدى الاختلافات المطلقة: 1.5 – 82.3%). وتراوح معامل كابا Kappa من 0.02 إلى 0.71 (الوسيط: 0.18). وقد أظهر الأسلوبان اللذان لهما أعلى تكرار للحالات (89.2% و 70.0%) توافقاً جيداً (معامل كابا: 0.33)؛ وأظهر الأسلوبان اللذان لهما أقل تكرار للحالات (6.9% و 10.0%) توافقاً ضئيلاً.

**الاستنتاج** هناك فقط توافق ضئيل بين الأساليب المنظمة للتحري عن سل الأطفال وتشخيصه، وهناك تفاوت كبير بين هذه الأساليب من حيث اكتشاف الحالات. إن النظم التشخيصية التي تتشابه في إنتاج تكرار منخفض للحالات من الممكن أنها تحدد فئات فرعية مختلفة من الأطفال. ولا تدعم نتائج الدراسة الاستخدام السريري الروتيني للأساليب المنظمة للتشخيص النهائي لسل الأطفال، بالرغم من أن النظم التي تؤدي إلى نتائج مرتفعة قد تكون مفيدة كأدوات للتحري.

## Résumé

### Approches structurées pour le dépistage et le diagnostic de la tuberculose chez l'enfant dans une région d'Afrique du Sud où cette maladie est fortement prévalente

**Objectif** Mesurer le degré d'accord entre neuf approches structurées pour le diagnostic de la tuberculose chez l'enfant ; quantifier les différences en termes de nombres de cas de tuberculose diagnostiqués entre ces neuf approches ; et déterminer la répartition des cas dans les différentes catégories de certitude diagnostique.

**Méthodes** Nous avons étudié 1445 enfants de moins de 2 ans appartenant à une communauté rurale d'Afrique du Sud, dans le cadre d'un essai vaccinal (2001-2006). Des données cliniques, radiologiques et microbiologiques ont été collectées prospectivement. Nous avons déterminé quel statut diagnostique (tuberculeux/non tuberculeux) était affecté par chacune des approches aux cas potentiels de tuberculose. Nous avons calculé les différences en termes de fréquence des cas et l'accord concernant la catégorie de certitude pour les résultats binaires (tuberculose/absence de tuberculose) en utilisant le test de McNemar (avec les intervalles de confiance à 95 %, IC) et le coefficient kappa de Cohen (K).

**Résultats** La fréquence des cas de tuberculose se situait entre 6,9 et 89,2 % (médiane : 41,7 %). Des différences significatives sont apparues dans la fréquence des cas ($p < 0,05$) dans 34 des 36 comparaisons par paire entre les approches diagnostiques structurées (plage de différences absolues : 1,5-82,3 %). Le coefficient kappa variait de 0,02 à 0,71 (médiane : 0,18). Les deux systèmes donnant les plus fortes fréquences de cas (89,2 % et 70,0 % respectivement) présentaient un accord satisfaisant (K : 0,33) ; les deux autres systèmes, qui avaient fourni les plus faibles fréquences (6,9 % et 10,0 %, respectivement), n'étaient que faiblement en accord (K : 0,18).

**Conclusion** Il n'existe qu'un faible accord entre les approches structurées du dépistage et du diagnostic de la tuberculose chez l'enfant et il apparaît entre elles une forte variabilité du rendement en cas. Les systèmes diagnostiques ayant fourni de manière similaire des fréquences de cas peu élevées pourraient identifier des sous-populations d'enfants différentes. Les résultats de cette étude ne sont pas en faveur d'un usage clinique systématique de ces approches structurées pour le diagnostic définitif des enfants tuberculeux, mais les systèmes fournissant un rendement élevé en cas pourraient constituer des outils de dépistage utiles.

## Resumen

### Sistemas estructurados de cribado y diagnóstico de la tuberculosis infantil en una región de alta prevalencia de Sudáfrica

**Objetivo** Medir la concordancia entre nueve sistemas estructurados de diagnóstico de la tuberculosis infantil; cuantificar las diferencias en cuanto al número de casos de tuberculosis diagnosticados con los diferentes sistemas, y determinar la distribución de casos en distintas categorías de certeza diagnóstica.

**Métodos** Se estudió a 1445 niños menores de 2 años durante un ensayo de vacunas (2001–2006) llevado a cabo en una comunidad rural de Sudáfrica. Se reunieron de forma prospectiva datos clínicos, radiológicos y microbiológicos, y se determinó si los niños sufrían o no tuberculosis usando cada una de las nueve modalidades de diagnóstico. Para calcular las diferencias en la frecuencia de casos y la concordancia de categorías para resultados binarios (tuberculosis/no tuberculosis), aplicamos la prueba de McNemar (con intervalos de confianza del 95%) y el coeficiente kappa de Cohen ($K$).

**Resultados** La frecuencia de casos de tuberculosis se situó entre 6,9% y 89,2% (mediana: 41,7). Se observaron diferencias significativas en la frecuencia de casos ($P < 0,05$) en 34 de las 36 comparaciones emparejadas entre los sistemas de diagnóstico estructurado (intervalo de diferencias absolutas: 1,5–82,3%). Kappa osciló entre 0,02 y 0,71 (mediana: 0,18). Los dos sistemas que hallaron las frecuencias de casos más altas (89,2% y 70,0%), mostraron una concordancia aceptable ($K$: 0,33); y los dos que hallaron las frecuencias de casos más bajas (6,9% y 10,0%) mostraron una concordancia baja ($K$: 0,18).

**Conclusión** Se observa solo una baja concordancia entre los sistemas estructurados en lo relativo al cribado y diagnóstico de la tuberculosis infantil, y una alta variabilidad entre ellos en términos de detección de casos. Sistemas de diagnóstico que arrojan frecuencias de casos similarmente bajas podrían estar detectando subpoblaciones de niños diferentes. Los resultados del estudio no respaldan el uso clínico sistemático de criterios estructurados para el diagnóstico definitivo de la tuberculosis infantil, pero los sistemas que consiguen valores altos de detección pueden ser un valioso instrumento de cribado.

## References

1. Marais BJ, Gie RP, Hesseling AC, Schaaf HS, Lombard C, Enarson DA et al.A refined symptom-based approach to diagnose pulmonary tuberculosis in children. *Pediatrics* 2006;118:e1350–9. doi:10.1542/peds.2006-0519 PMID:17079536
2. *A research agenda for childhood tuberculosis.* Geneva: World Health Organization; 2007 (WHO/HTM/TB/2007.2381).
3. Skeiky YA, Sadoff JC. Advances in tuberculosis vaccine strategies. *Nat Rev Microbiol* 2006;4:469–76. doi:10.1038/nrmicro1419 PMID:16710326
4. Nicol MPDM, Wood K, Hatherill M, Workman L, Hawkridge A, Eley B et al.A comparison of T-SPOT.TB and tuberculin skin test for the evaluation of young children at high risk for tuberculosis in a community setting. *Pediatrics* 2009;123:38–43. doi:10.1542/peds.2008-0611 PMID:19117858
5. Hesseling AC, Schaaf HS, Gie RP, Starke JR, Beyers N. A critical review of diagnostic approaches used in the diagnosis of childhood tuberculosis. *Int J Tuberc Lung Dis* 2002;6:1038–45. PMID:12546110
6. Groenewald P. *Boland-Overberg region annual health status report 2004.* Worcester: Department of Information Management, Department of Health; 2004 .
7. Houwert KA, Borggreven PA, Schaaf HS, Nel E, Donald PR, Stolk J. Prospective evaluation of World Health Organization criteria to assist diagnosis of tuberculosis in children. *Eur Respir J* 1998;11:1116–20. doi:10.1183/09031936.98.11051116 PMID:9648965
8. Hawkridge A, Hatherill M, Little F, Goetz MA, Barker L, Mahomed H et al.Efficacy of percutaneous versus intradermal BCG in the prevention of tuberculosis in South African infants: randomised trial. *BMJ* 2008;337:a2052. doi:10.1136/bmj.a2052 PMID:19008268
9. Fourie PB, Becker PJ, Festenstein F, Migliori GB, Alcaide J, Antunes M et al.Procedures for developing a simple scoring method based on unsophisticated criteria for screening children for tuberculosis. *Int J Tuberc Lung Dis* 1998;2:116–23. PMID:9562121
10. Harries A, Maher D, Graham S. *TB/HIV: a clinical manual.* 2nd ed. Geneva: World Health Organization; 2004.
11. Kibel M. *A point system for management of childhood tuberculosis.* Cape Town: Institute of Child Health, University of Cape Town; 1999.

12. Migliori GB, Borghesi A, Rossanigo P, Adriko C, Neri M, Santini S et al.Proposal of an improved score method for the diagnosis of pulmonary tuberculosis in childhood in developing countries. *Tuber Lung Dis* 1992;73:145–9. doi:10.1016/0962-8479(92)90148-D PMID:1421347
13. Montenegro SH, Gilman RH, Sheen P, Cama R, Caviedes L, Hopper T et al.Improved detection of Mycobacterium tuberculosis in Peruvian children by use of a heminested IS6110 polymerase chain reaction assay. *Clin Infect Dis* 2003;36:16–23. doi:10.1086/344900 PMID:12491196
14. Osborne CM. The challenge of diagnosing childhood tuberculosis in a developing country. *Arch Dis Child* 1995;72:369–74. doi:10.1136/adc.72.4.369 PMID:7763076
15. Pinkney-Atkinson V. *TB practical guidelines: managed care and quality review*. Cape Town: Medical Association of South Africa; 1996.
16. Stoltz AP, Donald PR, Strebel PM, Talent JM. Criteria for the notification of childhood tuberculosis in a high-incidence area of the western Cape Province. *S Afr Med J* 1990;77:385–6. PMID:2330522
17. Sant'Anna CC, Orfaliais CT, March Mde F, Conde MB. Evaluation of a proposed diagnostic scoring system for pulmonary tuberculosis in Brazilian children. *Int J Tuberc Lung Dis* 2006;10:463–5. PMID:16602415
18. Edwards K. The diagnosis of childhood tuberculosis. *P N G Med J* 1987;30:169–78. PMID:3314246
19. Ghidey Y, Habte D. Tuberculosis in childhood: an analysis of 412 cases. *Ethiop Med J* 1983;21:161–7. PMID:6603973
20. Stegen G, Jones K, Kaplan P. Criteria for guidance in the diagnosis of tuberculosis. *Pediatrics* 1969;43:260–3. PMID:5304285
21. *Provisional guidelines for the diagnosis and classification of the EPI target diseases for primary health care, surveillance and special studies*. Geneva: World Health Organization; 1983 (EPI/GEN/83/84).
22. Marais BJ, Gie RP, Obihara CC, Hesseling AC, Schaaf HS, Beyers N. Well defined symptoms are of value in the diagnosis of childhood pulmonary tuberculosis. *Arch Dis Child* 2005;90:1162–5. doi:10.1136/adc.2004.070797 PMID:16131501
23. Cundall DB. The diagnosis of pulmonary tuberculosis in malnourished Kenyan children. *Ann Trop Paediatr* 1986;6:249–55. PMID:2435230
24. Country profile. South Africa. In: *Global tuberculosis control. WHO report 2008*. Geneva: World Health Organization; 2008. Available from: http://www.who.int/globalatlas [accessed 20 November 2009].
25. McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, For GG. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *CMAJ* 2004;171:1369–73. PMID:15557592
26. Edwards DJ, Kitetele F, Van Rie A. Agreement between clinical scoring systems used for the diagnosis of pediatric tuberculosis in the HIV era. *Int J Tuberc Lung Dis* 2007;11:263–9. PMID:17352090
27. Kibel MA, Hussey G. Problems in the diagnosis of childhood tuberculosis. *S Afr Med J* 1990;77:379–80. PMID:2330519
28. Sant'Anna CC, Santos MA, Franco R. Diagnosis of pulmonary tuberculosis by score system in children and adolescents: a trial in a reference center in Bahia, Brazil. *Braz J Infect Dis* 2004;8:305–10. doi:10.1590/S1413-86702004000400006 PMID:15565261

**Corrigenda**

In volume 88, Number 3, March 2010:

- page 169 the 4th sentence of the 5th paragraph and page 170 photo caption should read "Lee In-sook, whose mother and sister both had cancer, …"
- page 170, the quotation in the final paragraph should be attributed to Yang Boon-min
- page 200, the first sentence of the second paragraph should read: "Table 1 shows the estimated incidence of selected communicable diseases in 2004 in the world and in the region."
- page 201, the title for Table 1 should read: "Estimated incidence of selected communicable diseases worldwide and in the South-East Asia Region of the World Health Organization, 2004[a]" and the column header should read "Estimated incidence (in thousands)"