Promise versus Reality: Optimism Bias in Package Inserts for Tuberculosis Diagnostics

Claudia M. Denkinger,a Jasmine Grenier,b Jessica Minion,c and Madhukar Pailde

Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA; Faculty of Medicine, McGill University, Montreal, Quebec, Canada; Department of Medical Microbiology & Immunology, University of Alberta, Edmonton, Alberta, Canada; Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada; and Respiratory Epidemiology & Clinical Research Unit, Montreal Chest Institute, Montreal, Quebec, Canada

Laboratorians and clinicians often rely on package inserts of diagnostic tests to assess their accuracy. We compared test accuracy for tuberculosis diagnostics reported in 19 package inserts against estimates in published meta-analyses and found that package inserts generally report overoptimistic accuracy estimates. However, package inserts of most tests approved by the U.S. Food and Drug Administration (FDA) or endorsed by the World Health Organization provide more realistic estimates that agree with meta-analyses.

Most laboratory professionals anticipate that test performance when applied to patient care may be less impressive than what is reported in package inserts by test manufacturers. However, this gap between promise and reality has not been examined in a systematic way.

One approach for studying this gap is to compare sensitivity and specificity estimates in package inserts with pooled estimates from published systematic reviews and meta-analyses, which often include studies from diverse settings and may provide a more realistic assessment of test accuracy.

We conducted such a comparison using commercial tuberculosis (TB) diagnostics as a case study. The TB diagnostic pipeline has rapidly expanded, and a large number of meta-analyses have been published on various TB tests (24). Several tests are now endorsed by the World Health Organization (WHO). At the same time, there are examples of suboptimal and inaccurate TB diagnostic tests, and their use has been discouraged by the WHO (32). In fact, in vitro tests are often poorly regulated in many countries (13).

We searched PubMed, the Cochrane Library, and the Evidence-Based TB Diagnosis website (www.tbevidence.org) for systematic reviews on the accuracy of diagnostics for TB published through March 2012 (search terms are available from the authors upon request). We excluded meta-analyses that reported only performance characteristics other than sensitivity and specificity (e.g., reproducibility, likelihood ratios), as they were not comparable to information provided in package inserts.

We searched company websites for package inserts and contacted test manufacturers if package inserts were not available online. Diagnostic tests were not considered if they were not commercially available. Diagnostic tests for latent TB were not considered, because no good reference standard is available to determine accuracy.

A total of 19 TB tests were included in the final analysis, because they met our eligibility criteria and package inserts as well as meta-analyses where available. These included gamma interferon (IFN-γ) release assays (IGRAs) for active TB, antibody-based serological tests, antigen detection tests, nucleic acid amplification tests (NAAT), and culture-based tests.

As seen in Table 1, the quality of information provided in package inserts varies widely. Eighteen out of 19 package inserts in total overestimated the test accuracy. In particular, technologies that were neither recommended by the WHO nor approved by the U.S. Food and Drug Administration (FDA) reported higher accuracy (i.e., serological tests for active TB, IGRAs for active TB, bacteriophage-based tests for active TB and drug-susceptibility testing, and urine lipoarabinomannan [LAM] antigen assays) (30, 32, 34). Claims made in package inserts ranged on average from 20 to 30% higher for sensitivity estimates than meta-analysis (comparing a range of findings in meta-analyses to estimates in package inserts). A direct comparison of absolute estimates of test accuracy was not feasible for most nonapproved tests, because the systematic reviews were unable to compute pooled sensitivity and specificity due to heterogeneity across studies.

In contrast, there is a better match between sensitivity and specificity estimates in package inserts and meta-analyses for tests that are approved by the FDA (i.e., Gen-Probe amplified MTB) or endorsed by the WHO (line probe assays, Xpert MTB/RIF, and MODS) (4, 31, 33, 34). For the WHO-endorsed tests that allowed a comparison of test accuracy between meta-analyses and package inserts (i.e., all except for the line probe assays), the package inserts overestimated sensitivity and specificity by at most 5%. This was also true for FDA-approved tests. IGRAs are FDA approved for latent TB but not active TB. The sensitivity for IGRAs for active TB was overestimated in package inserts by up to 20%. The comparison of the specificity of IGRAs was limited by the fact that meta-analyses of active TB often used TB suspects as controls, while package inserts reported specificity for latent TB infection among healthy, low-risk populations (6, 9, 19, 26).

We also found that test accuracy estimates in package inserts are often derived from unpublished, in-house, case-control studies with small numbers of specimens. Confirmed TB cases and healthy controls are often used, which can introduce significant
<table>
<thead>
<tr>
<th>Indication</th>
<th>Specimen</th>
<th>Test, yr of package insert publication (if reported)</th>
<th>WHO endorsed</th>
<th>FDA approved</th>
<th>Package insert No. of samples included</th>
<th>Comment on package insert</th>
<th>Meta-analysis Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>No. of studies no. of participants included (reference)</th>
<th>Comment on meta-analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFN-γ release assays</td>
<td>Blood</td>
<td>TB-SPOT, TB assay (Oxford Immunotec, Abingdon, United Kingdom), 2010</td>
<td>No</td>
<td>No</td>
<td>96 97</td>
<td>189 for sensitivity, 311 for specificity</td>
<td>Retrospective analysis of samples</td>
<td>92 59</td>
<td>27/3,886 (26)</td>
<td>Sensitivity reported for culture-confirmed cases; specificity reported for TB suspects, which may in part explain the reduced estimate compared to the PI</td>
</tr>
<tr>
<td></td>
<td>Blood</td>
<td>QuantiFERON-TB Gold In-Tube (QFT-GIT, Cellestis, Chadstone, Australia), 2006</td>
<td>No</td>
<td>No</td>
<td>89 99</td>
<td>54 for sensitivity, 581 for specificity</td>
<td>Prospective analysis; unpublished, not stratified by clinically relevant subgroups for active TB (i.e., HIV+, TB prevalent)</td>
<td>81 79</td>
<td>27/3,886 (26)</td>
<td>Sensitivity reported for culture-confirmed cases; specificity reported for TB suspects, which may in part explain the reduced estimate compared to the PI</td>
</tr>
<tr>
<td>Active PTB</td>
<td>Blood</td>
<td>TB-SPOT</td>
<td>No</td>
<td>No</td>
<td>96 97</td>
<td>189 for sensitivity, 311 for specificity</td>
<td>Retrospective analysis of samples</td>
<td>83 (68 HIV+, 88 HIV-), 61 (52 HIV+, 98 HIV-)</td>
<td>2/1,434 (19)</td>
<td>Stratified by HIV; sensitivity reported for culture-confirmed cases; specificity reported for TB suspects</td>
</tr>
<tr>
<td>Active PTB</td>
<td>Blood</td>
<td>QFT-IT</td>
<td>No</td>
<td>No</td>
<td>89 99</td>
<td>54 for sensitivity, 581 for specificity</td>
<td>Prospective analysis; unpublished, not stratified by clinically relevant subgroups for active TB (i.e., HIV+, TB prevalent)</td>
<td>69 (65 HIV+, 84 HIV-), 52 (50 HIV+, 98 HIV-)</td>
<td>2/1,434 (19)</td>
<td>Stratified by HIV; sensitivity reported for culture-confirmed cases; specificity reported for TB suspects</td>
</tr>
<tr>
<td>Active PTB and EPTB</td>
<td>Blood</td>
<td>TB-SPOT</td>
<td>No</td>
<td>No</td>
<td>96 97</td>
<td>189 for sensitivity, 311 for specificity</td>
<td>Retrospective analysis of samples</td>
<td>88–91 86–88</td>
<td>2 MAs: 16/1,638; 36/1,825 for sensitivity and 7/767 for specificity (6, 9)</td>
<td>Sensitivity reported in healthy controls; specificity lower if assessed in TB suspects</td>
</tr>
<tr>
<td></td>
<td>Blood</td>
<td>QFT-IT</td>
<td>No</td>
<td>No</td>
<td>89 99</td>
<td>54 for sensitivity, 581 for specificity</td>
<td>Prospective analysis; unpublished, not stratified by clinically relevant subgroups for active TB (i.e., HIV+, TB prevalent)</td>
<td>79–81 93–99</td>
<td>2 MAs: 16/1,638; 36/1,825 for sensitivity and 7/767 for specificity (6, 9)</td>
<td>Sensitivity reported in healthy controls; specificity lower if assessed in TB suspects</td>
</tr>
<tr>
<td>Antigen-based tests</td>
<td>Urine</td>
<td>LAM-ELISA (Chemogen, Clearview, Now Alere TB-LAM ELISA, Waltham, MA), 2011</td>
<td>No</td>
<td>No</td>
<td>73–81 (for HIV+ only) 70–88</td>
<td>469 2 unpublished, prospective studies in HIV+ patients; varying gold standard; specificity in 96 healthy controls (93–96%)</td>
<td>47–51 94–96</td>
<td>2 MAs: 5/845; 7/2,583 (11, 21)</td>
<td>Evaluated precommercial and commercial tests; concerns about methodology in majority of studies; specificity mostly tested on TB suspects but healthy controls included too; sensitivity approaches no. in package inserts in patients with advanced HIV</td>
<td></td>
</tr>
<tr>
<td>Active PTB, HIV+</td>
<td>Urine</td>
<td>LAM-ELISA (Chemogen, Clearview), 2008</td>
<td>No</td>
<td>No</td>
<td>Estimates not reported for HIV+</td>
<td>Estimates not reported for HIV+</td>
<td>No samples from HIV+ patients included</td>
<td>14 97</td>
<td>2 MAs: 5/845; 7/2,583 (11, 21)</td>
<td>Evaluated precommercial and commercial tests; concerns about methodology in majority of studies; specificity mostly tested on TB suspects but healthy controls included too; sensitivity approaches no. in package inserts in patients with advanced HIV</td>
</tr>
<tr>
<td>Test Type</td>
<td>Blood</td>
<td>Antibody Type</td>
<td>WHO Recommendation</td>
<td>Sensitivity (%)</td>
<td>Specificity (%)</td>
<td>Control No.</td>
<td>Source of Information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------------</td>
<td>---------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active SA PTB</td>
<td>Blood</td>
<td>Ansa-TB IgG (Ansa Biologicals, Strasbourg, France)</td>
<td>No recommended against</td>
<td>71–100</td>
<td>Not reported</td>
<td>76</td>
<td>92</td>
<td>11/1,570 (28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active SA PTB</td>
<td>Blood</td>
<td>Ansa IgG</td>
<td>No recommended against</td>
<td>71–100</td>
<td>Not reported</td>
<td>59</td>
<td>91</td>
<td>11/1,570 (28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active EPTB</td>
<td>Blood</td>
<td>Ansa IgG</td>
<td>No recommended against</td>
<td>71–100</td>
<td>Not reported</td>
<td>81</td>
<td>85</td>
<td>11/1,570 (28)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Active PTB.

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Blood</th>
<th>Antibody Type</th>
<th>WHO Recommendation</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Control No.</th>
<th>Source of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercially available ELISA + ICT (Mycodot, Mosesman Blackstone, MA; Mycobacterium tuberculosis IgG, IBL, Hamburg, Germany), 2011; ActiveTBDetect (Inbio International, Seattle WA), 2008; SEYA (Mahatma Gandhi Institute of Medical Sciences, India; Pathosysyme, Omega Diagnostics, Abra, Scotland), 2009; Hexagon (Human Gesellschaft Biochemica und Diagnostica, Wiesbaden, Germany), 2011; Serocheck-MTB (Zephyr, Biomedicals, Goa, India)</td>
<td>No recommended against</td>
<td>70–100 (all tests but Ansa)</td>
<td>90–100 (all tests but Ansa)</td>
<td>No, often not reported or &lt;50</td>
<td>Many tests are no longer commercially available; reported information in PIs very limited; often only small, unpublished, retrospective analyses using stored samples; positive exception is Mycodot with large prospective studies</td>
<td>60–88</td>
<td>50–98</td>
</tr>
<tr>
<td>Indicator</td>
<td>Specimen</td>
<td>Test, yr of package insert publication (if reported)</td>
<td>WHO endorsed</td>
<td>FDA approved</td>
<td>Sensitivity (%)</td>
<td>Specificity (%)</td>
<td>No. of samples included</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-----------------------------------------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Active EPTB Blood</td>
<td>Commercially available ELISA + ICT (Mycoblot, Moxosan Blackstone, MA; Mycobacterium tuberculosis IgG, IBL, Hamburg, Germany), 2011; ActiveTBdiast (Indico International, Seattle WA), 2008; SEVA (Mahatma Gandhi Institute of Medical Sciences, India; Patholyme, Omega Diagnostics, Alva, Scotland), 2009; Hexagon (Human Gesellschaft Biochemica und Diagnostica, Wiesbaden, Germany), 2011; Serocheck-MTB (Zephyr, Bumedicals, Goa, India)</td>
<td>WHO recommended against</td>
<td>No</td>
<td>No</td>
<td>48–100 (S&lt;sup&gt;+&lt;/sup&gt;, 97–99, S&lt;sup&gt;−&lt;/sup&gt;, 83–88)</td>
<td>73–100</td>
<td>No, often not reported or &lt;50</td>
</tr>
<tr>
<td>Active PTB Sput direct</td>
<td>FASTPlaque-TB (Biotec, Kentford, United Kingdom), 2004</td>
<td>No</td>
<td>No</td>
<td>73–82 (S&lt;sup&gt;+&lt;/sup&gt;, 49–67, S&lt;sup&gt;−&lt;/sup&gt;, 87)</td>
<td>98–99 (S&lt;sup&gt;+&lt;/sup&gt;, 98–100, S&lt;sup&gt;−&lt;/sup&gt;, 83–88)</td>
<td>&gt;2,000</td>
<td>2 published, prospective studies</td>
</tr>
<tr>
<td>Active PTB, R Rif Sput S&lt;sup&gt;+&lt;/sup&gt;</td>
<td>FASTPlaque Rif, FASTPlaque Response (Biotec, Kentford, United Kingdom), 2003</td>
<td>No</td>
<td>No</td>
<td>96–100 (only S&lt;sup&gt;+&lt;/sup&gt; cases)</td>
<td>98–100</td>
<td>374</td>
<td>Fid only available for FASTPlaque Response; 2 published studies; 17–27% with uninterpretable results</td>
</tr>
<tr>
<td>Nucleic acid-based test MTB detection Active PTB Sput direct</td>
<td>Xpert MTB/Rif (Cepheid, Sunnyvale, CA), 2011</td>
<td>Yes</td>
<td>No</td>
<td>92 (S&lt;sup&gt;+&lt;/sup&gt;, 98, S&lt;sup&gt;−&lt;/sup&gt;, 73)</td>
<td>99</td>
<td>1335</td>
<td>No information on sample collection, controls, other; Seminal papers cited</td>
</tr>
<tr>
<td>Active PTB Sput direct</td>
<td>Amplified MTD (Gen-Probe, San Diego, CA), 2001</td>
<td>No</td>
<td>Yes</td>
<td>86 (S&lt;sup&gt;+&lt;/sup&gt;, 97, S&lt;sup&gt;−&lt;/sup&gt;, 72)</td>
<td>99 (S&lt;sup&gt;+&lt;/sup&gt;, 100, S&lt;sup&gt;−&lt;/sup&gt;, 99)</td>
<td>206</td>
<td>Prospective, unpublished data from 7 study sites</td>
</tr>
<tr>
<td>Sput direct</td>
<td>Probe Tec ET (BD, Franklin Lakes, NJ), 2010</td>
<td>Yes</td>
<td>No</td>
<td>91 (S&lt;sup&gt;+&lt;/sup&gt;, 99, S&lt;sup&gt;−&lt;/sup&gt;, 75)</td>
<td>97</td>
<td>986</td>
<td>Prospective, unpublished data from 3 study sites</td>
</tr>
<tr>
<td>Test Method</td>
<td>Sputum Type</td>
<td>Nucleic Acid Test</td>
<td>Resistance</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Publication Details</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------------</td>
<td></td>
</tr>
</tbody>
</table>
| Active PTB, R
RIF | Sputum direct | Xpert MTB/Rif | Yes | 97 | 98 | 567 | No further information on sample collection, controls; seminal papers cited |
| Active PTB, R
RIF | Sputum direct and indirect | GenoType MTBDRplus (Hain Lifescience GmbH, Nehren, Germany), 2009 | Yes | No | No | 98–99 | 99 | 2 MAs: 5/767, 4/951 (2, 17) |
| Active PTB, R
INH | Sputum direct and indirect | Inno LIPA RIF/TB (Innogenetics NV, Gent, Belgium), 2011 | Yes | No | No | 89–96 | 99–100 | 2 MAs: 5/767, 5/981 (2, 17) |
| Microscopic
observation drug susceptibility testing | Sputum direct | TB MODS kit (Hardy Diagnostics, Santa Maria, CA), 2011 | Yes | No | No | 98 for MTB detection (R 100% for RIF, 97% for INH) | 100 for MTB detection | 92 | 96 | 14/3,711 for sensitivity; 12/7,226 for specificity (15) |
| Active PTB, R
RIF | Sputum direct | MODS | Yes | No | No | 98 for MTB detection (R 100% for RIF, 97% for INH) | 100 for MTB detection | 96–98 | 96–99 | 2 MAs: 6/1187 and 9/1474 (2, 20) |
| Active PTB, R
INH | Sputum direct | MODS | Yes | No | No | 98 for MTB detection (R 100% for RIF, 97% for INH) | 100 for MTB detection | 92 | 96 | 2 MAs: 6/1187 and 9/1474 (2, 20) |

---

*PI, package insert; R, resistance; RIF, rifampin; INH, isoniazid; PTB, pulmonary TB; EPTB, extrapulmonary TB; S⁺, smear positive; S⁻, smear negative; SS, sample size; MAs, meta-analyses; MTB, Mycobacterium tuberculosis.*
selection (spectrum) bias (25). The data in the package inserts often are not stratified based on important predictors of performance, including prevalence of TB or HIV and adults versus children, which may contribute to the overestimation of accuracy (3, 8, 19). In contrast, meta-analyses were often based on a fairly large number of studies that used cross-sectional or prospective designs and often were conducted in clinical settings with TB suspects that had a confirmed alternative final diagnosis serving as controls. Results were often stratified based on clinically relevant subgroups.

In general, involvement of industry and test developers in diagnostic evaluations has been associated with an overestimation of test accuracy (1). With TB tests, this has been documented with bacteriophage-based tests and urine lipoarabinomannan assays (11, 14, 21, 22). Users in real-world clinical settings may lack the same degree of expertise and skill as test developers. Also, quality control and assurance in routine clinical and laboratory settings may not match that of the industry. While data included in package inserts are almost always funded by industry, a proportion of studies included in the meta-analyses also are industry supported or conducted by test developers. This may then spuriously narrow the gap between package insert and meta-analyses estimates.

Our study has limitations. We were unable to compute numeric differences in the estimates of meta-analyses versus package inserts because pooling of data was often not possible due to heterogeneity between studies and the presence of several meta-analyses that included partially overlapping studies. We acknowledge that real-world performance of tests, especially when tests are scaled up in public health programs, may be worse than those reported in research studies, including meta-analyses (27). Thus, the real gap between package insert estimates and real-world performance may be even wider than what we document here. Pragmatic trials and implementation research are needed to overcome this problem (18). We also acknowledge that tests that measure the immune response to TB (i.e., serology, IGRAs) rather than products of Mycobacterium tuberculosis (i.e., Xpert MTB/RIF) might be more prone to variability in the results; however, this underlines the fact that accuracy data should always be stratified based on clinically relevant subgroups (i.e., HIV positive).

In summary, this case study of TB diagnostics suggests that package inserts often report overoptimistic estimates of test accuracy, especially if the products are not FDA approved (provided that approval was solicited) or WHO endorsed. These data provide some reassurance that independent review by credible agencies such as the FDA and WHO may serve as a yardstick for judging new TB technologies. However, not all TB tests are reviewed by the FDA or WHO, and most developing countries have weak regulatory systems for diagnostics. It is important that these countries create systems for in-country validation of all TB tests, guided by their national TB programs. Also, an expansion of the WHO prequalification of diagnostic programs to TB diagnostics will help countries procure quality-assured TB tests.

To overcome the problem of optimism bias, studies evaluating diagnostics under routine clinical and programmatic conditions, independent of industry sponsorship or test developers, are needed, as they provide more useful and realistic evidence to guide laboratorians, clinicians, and decision-makers. Furthermore, studies must go beyond accuracy and assess clinical impact of tests on decision-making and patient outcomes and collect operational and cost-effectiveness data in programmatic settings (7, 18).

ACKNOWLEDGMENTS

This study was supported by grants from the Canadian Institutes of Health Research (CIHR MOP-88918) and European and Developing Countries Clinical Trials Partnership (EDCTP) (TB-NEAT). Madhukar Pai is supported by a CIHR New Investigator Award and a career award from the Fonds de Recherche du Québec—Santé. None of these agencies were involved in the conduct or review of this study or the decision to publish its results.

We thank Daphne Ling for her contributions to this project. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed.

REFERENCES

29. Reference deleted.