Impact on patient outcomes: Diagnostic RCTs in TB

J. Lucian (Luke) Davis, MD, MAS
Division of Pulmonary & Critical Care Medicine
San Francisco General Hospital
July 9, 2013
Outline

• Defining elements of impact on patient outcomes

• Role of diagnostic RCTs

• Selected examples of different types of diagnostic RCTs & study design challenges

• Limitations of diagnostic RCTs
• Defining elements of impact on patient outcomes

• Role of diagnostic RCTs

• Selected examples of different types of diagnostic RCTs & study design challenges

• Limitations of diagnostic RCTs
How do you define impact?
Accuracy vs. impact

RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies

The GRADE system can be used to grade the quality of evidence and strength of recommendations for diagnostic tests or strategies. This article explains how patient-important outcomes are taken into account in this process.


87% TB diagnostic literature early accuracy studies ≤5% cover patient / public health outcomes or cost

One goal: correct classification

<table>
<thead>
<tr>
<th>Diagnostic Accuracy</th>
<th>TB</th>
<th>Not TB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test+</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Test-</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
Another goal: correct treatment

<table>
<thead>
<tr>
<th>Therapeutic Accuracy</th>
<th>TB</th>
<th>Not TB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test+</strong> <strong>Therapy +</strong></td>
<td>True Positive Early rule-in</td>
<td>False Positive Over-management</td>
</tr>
<tr>
<td><strong>Test-</strong> <strong>Therapy –</strong></td>
<td>False Negative Under-management</td>
<td>True Negative Early rule-out</td>
</tr>
</tbody>
</table>
A pragmatic definition

What processes and outcomes are important to patients, clinicians, and public health practitioners?
What are the consequences of misclassification?

<table>
<thead>
<tr>
<th>Treatment</th>
<th>TB</th>
<th>Not TB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No Treatment</strong></td>
<td><strong>Patient</strong></td>
<td><strong>Public Health</strong></td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td></td>
<td><strong>Patient</strong></td>
</tr>
</tbody>
</table>

- Drug toxicities/interactions
- Anxiety/stigma
- Missed true diagnosis
- Misallocated meds/labs/x-rays
- Misallocated health worker time
- Progression of 1° cases
- Added anxiety/delay/cost
- TB transmission
- Progression of 2° cases
- Added diagnostic visits
Outline

• Defining elements of impact on patient outcomes

• **Role of diagnostic RCTs**

• Selected examples of different types of diagnostic RCTs & study design challenges

• Limitations of diagnostic RCTs
Each management step can affect health outcomes
Structure of diagnostic RCT

Schünemann et al. BMJ 2008
• Defining elements of impact on patient outcomes

• Role of diagnostic RCTs

• **Selected examples of different types of diagnostic RCTs & study design challenges**

• Limitations of diagnostic RCTs
Individually randomized diagnostic trials

- Individuals randomized to intervention or control
- Not pragmatic if intervention planned to be used in entire lab / clinic / region rather than for specific individuals
- Cannot observe post-implementation effects on lab / clinic / regional level, including costs, and changes in demand, delays, and outcomes
Example: individual diagnostic RCT

Improvement of tuberculosis case detection and reduction of discrepancies between men and women by simple sputum-submission instructions: a pragmatic randomised controlled trial

Mishal Sameer Khan, Osman Dar, Charalambos Sismanidis, Karam Shah, Peter Godfrey-Faussett

Summary

Background In several settings, women with suspected tuberculosis are less likely to test smear positive than are men. Submission of poor-quality sputum specimens by women might be one reason for the difference between the sexes. We did a pragmatic randomised controlled trial to assess the effect of sputum-submission instructions on female patients.

Methods 1494 women and 1561 men with suspected tuberculosis attending the Federal Tuberculosis Centre in Rawalpindi, Pakistan, were randomly assigned between May and July, 2005 either to receive sputum-submission guidance before specimen submission or to submit specimens without specific guidance, according to prevailing practice. Of enrolled patients, 133 (4%) declined to participate. The primary outcome measure was the proportion of instructed and non-instructed women testing smear positive. Intention-to-treat analysis was undertaken on the basis of treatment allocation. This study is registered with the International Standard Randomised Controlled Trial number 34123170.

Findings Instructed women were more likely to test smear positive than were controls (Risk ratio 1.63 [95% CI 1.19–2.22]). Instructions were associated with a higher rate of smear-positive case detection (58 [8%] in controls vs 95 [13%] in the intervention group; p=0.002), a decrease in spot-saliva submission (p=0.003), and an increase in the number of women returning with an early-morning specimen (p=0.02). In men, instructions did not have a significant effect on the proportion testing smear positive or specimen quality.

Interpretation In the Federal Tuberculosis Centre in Rawalpindi, lower smear positivity in women than in men was mainly a function of poor-quality specimen submission. Smear positivity in women was increased substantially by provision of brief instructions. Sputum-submission guidance might be a highly cost-effective intervention to improve smear-positive case detection and reduce the disparity between the sexes in tuberculosis control in low-income countries.

Khan et al. Lancet 2007
Trial synopsis: *Khan et al. Lancet 2007*

- **Patients:** Adults attending Rawalpindi TB Centre
- **Intervention:** Standardized sputum collection instruction
- **Comparison:** No instruction
- **Outcome:** Smear-positive TB diagnosis

Planned stratified analysis by gender

(No culture to confirm incremental cases true positive)

- **Advantages of RCT design**
  - Avoids challenges of dividing sputum samples
  - Avoids individual contamination effect of instruction
Individual randomization: *Khan et al. Lancet 2007*

3055 randomisation
1561 men
1494 women

**Intervention group**
- 1520 allocated
  - 771 male patients received allocated intervention
  - 749 female patients received allocated intervention

- 1520 outcome determined
  - 771 men
  - 749 women

- 1520 data analysis
  - 771 men
  - 749 women
  - 2819 specimens
  - 1417 men
  - 1402 women

**Control group**
- 1535 allocated
  - 790 male patients received allocated intervention
  - 745 female patients received allocated intervention

- 1535 outcome determined
  - 790 men
  - 745 women

- 1535 data analysis
  - 790 men
  - 745 women
  - 2813 specimens
  - 1449 men
  - 1364 women
Diagnostic RCT: small effect sizes

diRuffano et al BMJ 2012
### Results: *Khan et al. Lancet 2007*

<table>
<thead>
<tr>
<th></th>
<th>Control (n=745)</th>
<th>Instructed (n=749)</th>
<th>Risk ratio instructed vs control (95% CI)</th>
<th>Risk difference: instructed minus control (95% CI)</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients positive†</td>
<td>58 (8%)</td>
<td>95 (13%)</td>
<td>1.63 (1.19 to 2.22)</td>
<td>4.9 (1.8 to 8.0)</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients positive†</td>
<td>78 (9.9)</td>
<td>90 (11.7)</td>
<td>1.18 (0.89 to 1.57)</td>
<td>1.8 (−1.3 to 4.9)</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- Contamination effects on staff collecting sputum?
- Variations in feasibility in different clinics / populations?
Cluster randomized trial designs

- **Parallel**
  - Site randomized to control or intervention
  - Directly compare clusters, which may be matched based on site characteristics

- **Cross-over**
  - Randomized order of control and intervention
  - Fewer clusters but takes longer
  - Wash-out between periods

Slide adapted from Susan van den Hof
Comparison of two active case-finding strategies for community-based diagnosis of symptomatic smear-positive tuberculosis and control of infectious tuberculosis in Harare, Zimbabwe (DETECTB): a cluster-randomised trial

Elizabeth L Corbett, Tsietsi Bandason, Trinh Ducong, Ethel Dauya, Beauty Makumure, Gavin J Churchyard, Brian G Williams, Shungu S Munyuati, Anthony E Butterworth, Peter R Mason, Stanley Munroga, Richard J Hayes

Summary

Background Control of tuberculosis in settings with high HIV prevalence is a pressing public health priority. We tested two active case-finding strategies to target long periods of infectiousness before diagnosis, which is typical of HIV-negative tuberculosis and is a key driver of transmission.

Methods Clusters of neighbourhoods in the high-density residential suburbs of Harare, Zimbabwe, were randomised to receive six rounds of active case finding at 6-monthly intervals by either mobile van or door-to-door visits. Randomisation was done by selection of discs of two colours from an opaque bag, with one disc to represent every cluster, and one colour allocated to each intervention group before selection began. In both groups, adult (≥16 years) residents volunteering chronic cough (≥2 weeks) had two sputum specimens collected for fluorescence microscopy. Community health workers and cluster residents were not masked to intervention allocation, but investigators and laboratory staff were masked to allocation until final analysis. The primary outcome was the cumulative yield of smear-positive tuberculosis per 1000 adults per group, compared between intervention groups; analysis was by intention to treat. The secondary outcome was change in prevalence of culture-positive tuberculosis from before intervention to before round six of intervention in 12% of randomly selected households from the two intervention groups combined; analysis was based on participants who provided sputum in the two prevalence surveys. This trial is registered, number ISRCTN84352452.

Findings 46 study clusters were identified and randomly allocated equally between intervention groups, with 55 741 adults in the mobile van group and 54 691 in the door-to-door group at baseline. HIV prevalence was 21% (1916/9060) and in the 6 months before intervention the smear-positive case notification rate was 2.8 per 1000 adults per year. The trial was completed as planned with no adverse events. The mobile van detected 255 smear-positive patients from 5466 participants submitting sputum compared with 137 of 4711 participants identified through door-to-door visits (adjusted risk ratio 1.48, 95% CI 1.11–1.96, p = 0.0087). The overall prevalence of culture-positive tuberculosis declined from 6.5 per 1000 adults (95% CI 5.1–8.3) to 3.7 per 1000 adults (2.6–5.0; adjusted risk ratio 0.59, 95% CI 0.40–0.89, p = 0.0112).

Interpretation Wide implementation of active case finding, particularly with a mobile van approach, could have rapid effects on tuberculosis transmission and disease.

Corbett et al. Lancet 2010
Trial synopsis: Corbett et al. Lancet 2010

- **Population:** Adults with cough ≥ 2 weeks in Harare
- **Intervention:** Active TB case finding via mobile van
- **Comparison:** Active TB case finding via household visits
- **Outcome:** AFB+ TB per 1000 subjects screened

**Advantages of RCT design**
- Implementation design
- More pragmatic than individual RCT – assess feasibility & efficacy
- Minimize cross-cluster contamination
- Randomization at group level appropriate if exposure is a group level exposure (e.g., public health or social intervention)
Results: Corbett et al. Lancet 2010

- Mobile van: 255/5466 (4.7%)
- Household screening: 137/4711 (2.9%)
- RR 1.48, 95% CI 1.11–1.96, p<0.01, favoring mobile van

- Intention-to-treat analysis

- ~115,000 individuals in 46 clusters
Cluster RCTs: intra-class correlation

• Reduced statistical power a major challenge
  – Members of a group (cluster) typically resemble each other more than they resemble members of other groups (clusters)
  – **FEWER independent observations**

• Intra-class Correlation Coefficient (**ICC**), also known as $\rho$
  – Ratio of variances: between-cluster to total (between + within)
    • ICC=0 : no correlation (independent observations)
    • ICC=1 : perfect correlation (observations dependent on site)
    • ICC typically = 0.001 - 0.05
Effective sample size = sample size / design effect

- Design effect is the sample size penalty for using a clustered design

- Design effect = 1 + (m-1) * ICC

- k = 46 clusters
- m = 2500 people/cluster
- Total n = 115,000

- Hard to predict ICC in advance!

<table>
<thead>
<tr>
<th>ICC</th>
<th>Design effect</th>
<th>Effective sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>115,000</td>
</tr>
<tr>
<td>0.001</td>
<td>3.5</td>
<td>33,000</td>
</tr>
<tr>
<td>0.01</td>
<td>26</td>
<td>4,400</td>
</tr>
<tr>
<td>0.1</td>
<td>251</td>
<td>460</td>
</tr>
<tr>
<td>0.5</td>
<td>1251</td>
<td>92</td>
</tr>
<tr>
<td>1</td>
<td>2500</td>
<td>46</td>
</tr>
</tbody>
</table>

Slide adapted from John Metcalfe
Cluster RCTs: challenges of clustering

- Ignoring clustering at the design stage $\rightarrow$ type II error
  - Standard sample size calculations will be underpowered for CRT’s

- Ignoring correlation at the analysis phase $\rightarrow$ type I error
  - Not accounting for clustering gives falsely narrow confidence intervals
Other challenges of cluster RCTs

- Five additional challenges
  1) Assigning control clusters representatively
  2) Blinding
  3) Minimizing contamination between clusters
  4) Implementing interventions simultaneously
  5) Piggy-backing new interventions on rapidly and differentially changing practices
Cluster randomized trial designs

**Parallel**
- Site randomized to control or intervention
- Directly compare clusters, which may be matched based on site characteristics

**Cross-over**
- Randomized order of control and intervention
- Fewer clusters but takes longer
- Wash-out between periods
A Multi-Country Non-Inferiority Cluster Randomized Trial of Frontloaded Smear Microscopy for the Diagnosis of Pulmonary Tuberculosis

Luis Eduardo Cuevas¹,²*, Mohammed Ahmed Yassin¹, Najla Al-Sonboli³, Lovett Lawson⁴, Isabel Arbide⁵, Nasher Al-Aghbari⁶, Jeevan Bahadur Sherchand⁷, Amin Al-Absi⁶, Emmanuel Nnamdi Emenyonu⁴, Yared Merid⁸, Mosis Ifenyi Okobi⁹, Juliana Olubunmi Onuoha⁴, Melkamsew Aschalew⁸, Abraham Aseffa¹⁰, Greg Harper¹, Rachel Mary Anderson de Cuevas¹, Kristin Kremer¹¹, Dick van Soolingen¹¹, Carl-Michael Nathanson², Jean Joly², Brian Faragher¹, Stephen Bertel Squire¹, Andrew Ramsay²

¹ Liverpool School of Tropical Medicine, Liverpool, United Kingdom, ² United Nations Children’s Fund/United Nations Development Programme/World Bank/World Health Organization Special Programme for Research and Training in Tropical Diseases, Geneva, Switzerland, ³ Medical Faculty, Sana’a University, Sana’a, Yemen, ⁴ Zankli Medical Center, Abuja, Nigeria, ⁵ Bushulo Major Health Centre, Awassa, Ethiopia, ⁶ National Tuberculosis Institute, Sana’a, Yemen, ⁷ Tribhuvan University Institute of Medicine, Kathmandu, Nepal, ⁸ Southern Region Health Bureau, Awassa, Ethiopia, ⁹ Wuse General Hospital, Abuja, Nigeria, ¹⁰ Armauer Hansen Research Institute, Addis Ababa, Ethiopia, ¹¹ Tuberculosis Reference Laboratory, National Institute for Public Health and the Environment, Bilthoven, The Netherlands


- **Population:** 6627 adults with cough ≥ 2 weeks, 8 sites
- **Intervention:** Front-loaded sputum collection
- **Comparison:** Standard sputum collection
- **Outcomes:**
  1. Diagnostic accuracy (non-inferiority)
  2. Collection completion rate (superiority)

*Cuevas et al, PLoS Med 2011*
## Non-inferiority

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superiority</td>
<td>A clinically relevant (&gt; Δ) and statistically significant (p&lt;0.05) difference</td>
</tr>
<tr>
<td>Equivalence</td>
<td>No statistically significant difference at pre-specified clinically relevant margin</td>
</tr>
<tr>
<td>Non-inferiority</td>
<td>New test no worse than existing test at pre-specified clinically relevant margin</td>
</tr>
</tbody>
</table>

![Graph showing the concept of non-inferiority](image)

Jones, BMJ, 2006
Several challenges for non-inferiority studies

• “Margin of non-inferiority” (Δ) must be established *a priori*

• Looking for a generally smaller effect than in superiority trials – usually require larger sample sizes

• Natural incentives against bias lacking since common sources of error will bias towards the null, e.g.
  – Non-completion of testing and study drop-outs
  – Missing data
  – Cross-overs

• Report both intention-to-treat (intention-to-test, ITT) and per protocol (PPA) analyses

*Slide adapted from John Metcalfe*

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ITT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>63.6% (59.7%–67.5%)</td>
<td>97.4% (93.5%–99.9%)</td>
</tr>
<tr>
<td>SM</td>
<td>64.8% (61.3%–68.3%)</td>
<td>97.8% (94.3%–99.9%)</td>
</tr>
<tr>
<td>SSM</td>
<td>70.2% (66.5%–73.9%)</td>
<td>96.9% (93.2%–99.9%)</td>
</tr>
<tr>
<td>SMS</td>
<td>65.9% (62.3%–69.5%)</td>
<td>97.6% (94.0%–99.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PPA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>63.6% (59.6%–67.6%)</td>
<td>97.6% (93.6%–99.9%)</td>
</tr>
<tr>
<td>SM</td>
<td>65% (61.6%–68.4%)</td>
<td>97.8% (94.4%–99.9%)</td>
</tr>
<tr>
<td>SSM</td>
<td>70.6% (66.7%–74.5%)</td>
<td>97% (93.1%–99.9%)</td>
</tr>
<tr>
<td>SMS</td>
<td>66.4% (62.9%–69.9%)</td>
<td>97.5% (94.0%–99.9%)</td>
</tr>
</tbody>
</table>

3-smear difference under ITT: 4.3% (95% CI 0.6% to 9.0%)

2-smear difference under ITT: −1.2% (95% CI −3.9% to 6.4%)

Differences under PPA: "Similar"

P < 0.001

Percentage

Number of sputum samples submitted

One
First two
All three

Spot-Spot-Morning
Spot-Morning-Spot
## Diagnostic RCTs: not always efficient?

<table>
<thead>
<tr>
<th>Setting</th>
<th>Prevalence of tuberculosis</th>
<th>Number of patients enrolled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subgroup A (ZN)</td>
<td>Subgroup B (LED FM)</td>
</tr>
<tr>
<td><strong>Paired observational studies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ramsay et al (2009)¹⁸</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nepal</td>
<td>Primary health clinic</td>
<td>26%</td>
</tr>
<tr>
<td>Nigeria</td>
<td>District hospital clinic</td>
<td>37%</td>
</tr>
<tr>
<td>Yemen</td>
<td>Subspecialty clinic</td>
<td>29%</td>
</tr>
<tr>
<td>Cattamanchi et al (2011)²⁴</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uganda</td>
<td>Referral hospital</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Randomised controlled trial</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cuevas et al (2011)²⁵²⁶ياة</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethiopia</td>
<td>Primary health clinic</td>
<td>33%</td>
</tr>
<tr>
<td>Nepal</td>
<td>Subspecialty clinic</td>
<td>13%</td>
</tr>
<tr>
<td>Nigeria</td>
<td>District hospital clinic</td>
<td>21%</td>
</tr>
<tr>
<td>Yemen</td>
<td>Subspecialty clinic</td>
<td>26%</td>
</tr>
</tbody>
</table>

Davis et al Lancet Infect Dis 2013
Outline

• Defining elements of impact on patient outcomes

• Role of diagnostic RCTs

• Selected examples of different types of diagnostic RCTs & study design challenges

• Limitations of diagnostic RCTs
Diagnostic RCTs: not always ethical?

The Ethics of Testing a Test: Randomized Trials of the Health Impact of Diagnostic Tests for Infectious Diseases

1. **Equipoise** – Evidence of accuracy vs. evidence of impact?
2. **Equity** – Impact in central vs. peripheral settings?
3. **Informed consent** – Participants in cluster RCTs should be informed why they don’t have access to the standard test

Dowdy et al Clin Infect Dis 2012
The ultimate goal of medical care, including diagnostic testing, is to improve patient outcome. Accordingly, it has been advocated widely that when establishing a test’s diagnostic accuracy, the impact of the test on patient outcome subsequently must be quantified. When studying patient outcome in medical research, the use of randomized comparisons comes into perspective. In our view, randomized studies often are not necessary to validly estimate the effect of the diagnostic test on patient outcome. Results of cross-sectional diagnostic studies, combined with results from therapeutic studies, often will suffice.

Ann Epidemiol 2006;16:540–544. © 2006 Elsevier Inc. All rights reserved.

When performing a randomized trial to determine the impact of a diagnostic test or strategy on patient outcome, an initially diagnostic research question is transformed into therapeutic research question (with the goal of establishing causality) with corresponding consequences for the design of the study. A disadvantage of a randomized approach to directly quantify the contribution of a diagnostic test and treatment on patient outcome is that it often addresses diagnosis and treatment as a single combined strategy, a “package deal.” This makes it impossible to determine afterwards whether a positive effect on patient outcome was attributed solely to the improved diagnosis by using the test under study or to the chosen (new) treatment strategies.
When Is Measuring Sensitivity and Specificity Sufficient To Evaluate a Diagnostic Test, and When Do We Need Randomized Trials?

Sarah J. Lord, MBBS, MS; Les Irwig, MBBCh, PhD; and R. John Simes, MBBS, MS, MD

The clinical value of using a new diagnostic test depends on whether it improves patient outcomes beyond the outcomes achieved using an old diagnostic test. When can studies of diagnostic test accuracy provide sufficient information to infer clinical value, and when do clinicians need to wait for results from randomized trials? The authors argue that accuracy studies suffice if a new diagnostic test is safer or more specific than, but of similar sensitivity to, an old test. However, if a new test is more sensitive than an old test, it leads to the detection of extra cases of disease. Results from treatment trials that enrolled only patients detected by the old test may not apply to these extra cases. Clinicians need to wait for results from randomized trials assessing treatment efficacy in cases detected by the new diagnostic test, unless they can be satisfied that the new test detects the same spectrum and subtype of disease as the old test or that treatment response is similar across the spectrum of disease.


For author affiliations, see end of text.
*Patients diagnosed as diseased with the new test may not respond in the same way as patients diagnosed with the old test.

Lord, Annals Int Med, 2006
In summary

• **No RCT needed**
  – When new test has similar sensitivity to old test
  – When new test has better specificity than old test

• **RCT needed**
  – When new test has a better sensitivity than old test
  – When new test detects different disease spectrum from old test

• **RCT may be needed, but comparing tradeoffs may suffice**
  – Better sensitivity / worse specificity
  – Worse sensitivity / better specificity
  – Similar sensitivity / specificity, different feasibility/cost characteristics