Comparative Effectiveness Research Using Meta-Analysis to Evaluate and Summarize Diagnostic Accuracy

Kelly H. Zou, PhD, Ching-Ray Yu, PhD, Ye Tan, PhD, and Martin O. Carlsson, MS
Pfizer Inc, New York, N.Y. 10017, U.S.A.

1. Introduction
- The Agency for Healthcare Research and Quality defines that the Comparative Effectiveness Research (CER) is designed to inform health-care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options.1
- The evidence is generated from research studies that compare drugs, medical devices, tests, surgeries, or ways to deliver health care.1
- Meta-analysis is a quantitative method for combining the results of independent studies, usually drawn from published literature, and for synthesizing summaries and conclusions, which may be used to evaluate therapeutic effectiveness and plan new studies.2

2. Objectives
- To synthesize and combine studies that yield proportions in a two-sample setting according to a Reference Standard (RS).
- To illustrate and compare fixed and random effects methods for:
  - (1) Displaying Sensitivities (Se) and Specificities (Sp);
  - (2) Combining the Diagnostic Odds Ratio (DOR) or log of DOR (LDOR);
  - (3) Generating a Summary Receiver Operating Characteristic (sROC) curve.

3. A Publicly-Available Example on CT Scans of Urolithiasis
- Prospective and retrospective studies from 1995 to 2007 were searched via PubMed, Medline, and Cochrane Library.3
- Low-dose Computed Tomography (CT) scan, with < 3 mSv dose applied for the entire CT examination, was the diagnostic test for the detection of urolithiasis, i.e., a stone located in the ureter.
- Each of the final k = 1, …, 7 studies provided the counts of urolithiasis from low-dose CT to determine urolithiasis (see Table 1).

4. Methods
- Notations and Assumptions
  - A two-by-two table (Table 2) is formed per study, first by stratifying the diagnostic results according to the binary RS (healthy vs. diseased).
  - Within the k-th (k = 1, …, K) study, for the healthy sample of size mk among subjects with RS = 0, the k-th subject-level diagnostic (Dxk) is generated by an independent and identical (i.i.d.) distribution, Xk ~ i.i.d. F(xk), k = 1, …, mk.
  - Similarly and independently, for the diseased sample of size nk among subjects whose RS = 1, the j-th subject-level Dxk is generated by an i.i.d. distribution, Yk ~ i.i.d. F(yk), j = 1, …, nk.

Table 1. Classifications in Each of the 7 Studies on Low-Dose CT to Detect Urolithiasis.

<table>
<thead>
<tr>
<th>Study</th>
<th>Low-dose CT Scan</th>
<th>Diseased Sample</th>
<th>Healthy Sample</th>
<th>Sensitivity (Se)</th>
<th>Specificity (Sp)</th>
<th>True Positive (TP)</th>
<th>False Negative (FN)</th>
<th>True Negative (TN)</th>
<th>False Positive (FP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
</tr>
<tr>
<td>2</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
</tr>
<tr>
<td>3</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
</tr>
<tr>
<td>4</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
</tr>
<tr>
<td>5</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
</tr>
<tr>
<td>6</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
<td>0.957 0.973 792.000</td>
<td>0.600 0.800 1.792</td>
<td>0.008 0.021 0.968</td>
</tr>
</tbody>
</table>

Table 2. A Two-by-Two Table of Counts within Study k.

- The study-level true positive rate is $Sp = TNk/mk$.
- The study-level true positive rate is $Se = TPk/nk$.
- The 95% Confidence Interval (CI) for these estimates may be constructed in a logit space first before being transformed back to the [0,1] interval.

- The Diagnostic Odds Ratio and Log of DOR
  - The ratio of (the odds of the test being positive if the subject has a disease) against (the odds of the test being positive if the subject does not have the disease) is $DOR_k = (TPk/FNk)/(FPk/TNk)$.
  - After a log transformation, $LDOR_k = log(DOR_k) = log((TPk/FNk)/(FPk/TNk))$.
  - The Standard Error (SE) of the estimated $LDOR_k$ is straightforward, $\sqrt{\frac{1}{(TP_k+1)(FN_k+1)(TN_k+1)(FP_k+1)}}, \sqrt{\frac{1}{2}}$.

- Forest Plot: Heterogeneity and Test ROC Curve
  - In a forest plot, the results of K individual studies are displayed as squares centered on the point estimate of the result of each study.4
  - A horizontal line runs through the square to show each 95% CI.5
  - The f statistic measures the heterogeneity, with low, moderate, and high correspond to the benchmark values, f=25%, 50%, and 75%, respectively.6
  - To synthesize across all K studies, both the fixed effects Mantel–Haenszel (MH) and the random effects DerSimonian-Laird (DSL) methods are used.7,8
  - The sROC curve plots (1-Sp, Se) and assumes that, overall, Se = f(Sp), where f is an accuracy parameter within a Lehman family for fitting.9,10

5. Results
- The heterogeneity measure is very low, with $I^2 \leq 20\%$ across all studies.
- Figs. 1 & 2 are the forest plots of $Sp_k$, $Se_k$, and LDOR_k, respectively.
- Fig. 1. Forest Plot of $Sp_k$ and $Se_k$.
- Fig. 2. Forest Plot of LDOR_k via MH and DSL.
- Table 3 give the MH and DSL results, Fig. 3 is the sROC curve: AUC=0.991.

6. Monte-Carlo Simulations
- Assume homogeneity across K=10 studies; thus, the true accuracy is pre-determined to compare fixed (MH) and random (DSL) effects methods.
- Generate k = 1, …, K sets of study-level data with $mk, nk = $25; 50.
- $X_k \sim Binomial(m_k, Sp_k)$, with $Sp_k =$0.6; 0.8 for the healthy subjects.
- $Y_k \sim Binomial(m_k, Se_k)$, with $Se_k =$0.6; 0.8 for the diseased subjects.
- With $MC=10000$ replicates, Table 3 shows the mean bias, Mean Squared Errors (MSE), and coverage probability (95% as the nominal level).

7. Conclusions
- The random effects model yields higher coverage with comparable MSE.
- The choice of method may depend on heterogeneity across all studies.

8. References
- 1. AHRQ. What is Comparative Effectiveness Research? 2013.