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 Publically-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 uroliathsis (see Table 1).

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

Study (k)	#Healthy (m_k)	True Negative (TN_k)	False Positive (FP_k)	#Diseased (n_k)	True Positive (TP_k)	False Negative (FN_k)	Specificity (Sp_k)	Sensitivity (Se_k)	Diagnostic Odds Ratio (DOR_k)	Log-DOR ($LDOR_k$)
1	23	22	1	37	36	1	0.957	0.973	792.000	6.675
2	29	28	1	80	77	3	0.966	0.962	718.667	6.577
3	62	61	1	147	142	5	0.984	0.966	1732.400	7.457
4	14	12	2	102	96	6	0.857	0.941	96.000	4.564
5	40	38	2	102	99	3	0.950	0.971	627.000	6.441
6	24	23	1	101	98	3	0.958	0.970	751.333	6.622
7	142	133	9	158	154	4	0.937	0.975	568.944	6.344

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 m_k among subjects with RS_k=0, the i-th subject-level diagnosis (Dx_k) is generated by an independent and identical ($i.i.d.$) distribution, $X_{k,i}$~$i.i.d.$ $F(x_k)$, i=1, ..., m_k.
- Similarly and independently, for the diseased sample of size n_k among subjects whose RS_k=1, the j-th subject-level Dx_k is generated by an $i.i.d.$ distribution, $Y_{k,j}$~$i.i.d.$ $F(y_k)$, j=1, ..., n_k.

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

Binary Diagnosis (Dx_k)	Binary Reference Standard (RS_k)		Marginal Count
	RS_k=0 (Healthy)	RS_k=1 (Diseased)	
Dx_k=0 (Negative)	True Negative (TN_k)	False Negative (FN_k)	TP_k+FN_k
Dx_k=1 (Positive)	False Positive (FP_k)	True Positive (TP_k)	FP_k+TP_k
Marginal Count	m_k=TN_k+FP_k	n_k=FN_k+TP_k	N_k=m_k+n_k

Specificity and Sensitivity

- The study-level true negative rate is Sp_k=TN_k/m_k.[4]
- The study-level true positive rate is Se_k=TP_k/n_k.[4]
- 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.

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=(TP_k/FN_k)/(FP_k/TN_k).
- After a log transformation, $LDOR_k$=ln(DOR_k)=ln(TP_k)-ln(FN_k)+ln(TN_k)-ln(FP_k).
- The Standard Error (SE) of the estimated $LDOR_k$ is straightforward, $[1/(TP_k)+1/(FN_k)+1/(TN_k)+1/(FP_k)]^{1/2}$, with the associated 95% CI constructed.

Forest Plot, Heterogeneity Test, and Summary 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.[5]
- A horizontal line runs through the square to show each 95% CI.[5]
- The I^2 statistic measures the heterogeneity, with low, moderate, and high correspond to the benchmark values, I^2=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_k, Se_k) and assumes that, overall, Se=(1-Sp)$^{\theta}$, where θ is an accuracy parameter within a Lehman family for fitting.[9, 10]

5. Results

- The heterogeneity measure is very low, with I^2≈0 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 and Se.

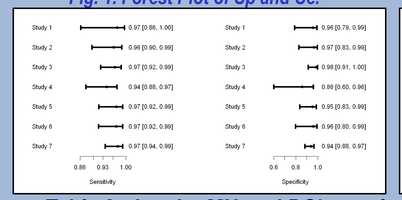
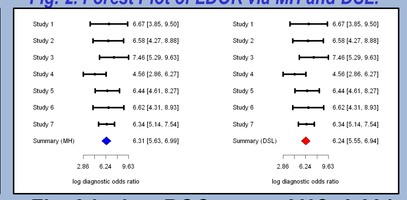


Fig. 2. Forest Plot of LDOR via MH and DSL.

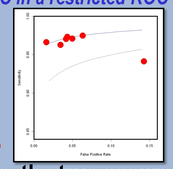


- Table 3 give the MH and DSL results; Fig. 3 is the sROC curve: AUC=0.991.

Table 3. Synthesized Results by MH and DSL.

Method	Combined $LDOR$	95% CI
Fixed Effects (MN)	6.312	(5.629, 6.995)
Random Effects (DSL)	6.244	(5.545, 6.942)

Fig. 3. The sROC in a restricted ROC space.



6. Monte-Carlo Simulations

- Assume homogeneity across K=10 studies; therefore, 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 m_k=n_k={25; 50}.
- X_k~ $Binomial$(m_k, Sp_k), with Sp_k={0.6; 0.8} for the health subjects.
- Y_k~ $Binomial$(m_k, Se_k), with Se_k ={0.6; 0.8} for the diseased subjects.
- With MC=10000 replicates, Table 4 shows the mean bias, Mean Squared Errors (MSE), and coverage probability (with 95% as the nominal level).

Table 4. Monte-Carlo Simulation Results in K=10 Studies with 10000 replicates.

# Total	Underlying Accuracy			Fixed Effects (MH) Method			Random Effects (DSL) Method		
(m_k+n_k)	Sp_k	Se_k	$LDOR_k$	Mean Bias	MSE	Coverage	Mean Bias	MSE	Coverage
25+25 =50	0.600	0.600	0.811	-0.005	0.034	0.922	-0.008	0.034	0.962
	0.600	0.800	1.792	-0.001	0.042	0.936	0.002	0.041	0.968
	0.800	0.600	1.792	-0.002	0.043	0.933	0.000	0.041	0.966
	0.800	0.800	2.773	0.006	0.054	0.939	-0.007	0.051	0.968
50+50 =100	0.600	0.600	0.811	-0.004	0.017	0.921	-0.006	0.017	0.962
	0.600	0.800	1.792	-0.008	0.021	0.932	-0.008	0.021	0.962
	0.800	0.600	1.792	-0.006	0.021	0.932	-0.006	0.021	0.965
	0.800	0.800	2.773	-0.010	0.026	0.937	-0.018	0.026	0.961

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.
2. Zou KH et al. *Acad Radiol.* 2004; 11: 127-133.
3. Niemann T et al. *Am J Roentgenol.* 2008; 191: 396-401.
4. Zou KH et al. *Biom J.* 2012;54: 249-263.
5. Lewis S & Clarke M. *BMJ.* 2001; 322: 1479-1480.
6. Higgins JP et al. *BMJ.* 2003; 327: 557-560.
7. Robins J et al. *Am J Epidemiol.* 1986; 124: 719-723.
8. DerSimonian R & Laird N. *Control Clin Trials.*1986; 7: 177-188.
9. Holling H et al. *Statistical Modelling.* 2012; 12: 347-375.
10. Doebler P. *R Package 'mada'.* 2013.