



**GENERALIZED EVIDENCE SYNTHESIS IN COMPARATIVE EFFECTIVENESS RESEARCH:
COULD THE EVIDENCE BASE BE BROADENED IN MIXED TREATMENT COMPARISONS?**



 David J. Vanness, PhD
 Ágnes Benedict, MA, MSc

Scientific Consulting & Communications

Discussion Leaders



- David J. Vanness, PhD
Assistant Professor, Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health
- Ágnes Benedict, MA, MS
European Director, Economic Modeling and Simulations, United BioSource Corporation

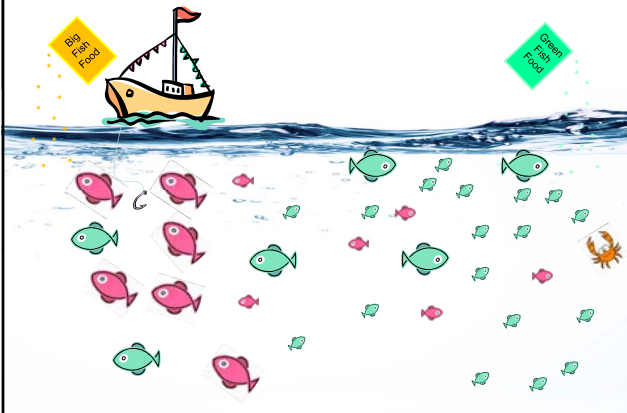



Scientific Consulting & Communications

Workshop Outline

- Background
- Section 1: Study Considerations in Evidence Synthesis in CER
- Section 2: Case Studies in Generalized Evidence Synthesis
- Section 3: A Simulated Teaching Example
- Discussion



Fishing expeditions...



- Two approaches to evidence synthesis:
 - Restrictive inclusion criteria (traditional meta-analysis)
 - » RCT only (possibly restricted further on quality)
 - » Closely matched sample characteristics and interventions
 - » Goal is to reduce heterogeneity and improve internal validity
 - Expansive inclusion criteria (generalized evidence synthesis)
 - » RCTs and observational studies
 - » Wide tolerance for sample characteristics/interventions
 - » Goal is to control for heterogeneity/improve generalizability
- Which is most useful for CER/Patient centered outcomes research?



Needs for CER

- CER is pragmatic: goal is to inform real-world decisions
 - Heterogeneous clinical practice
 - » Treatment switches, discontinuation
 - » Wider variety of patient and clinician behaviors than in controlled trial environment
 - Heterogeneous treatment effects
 - » Effect modifiers may imply definable subgroups
 - Desire to use CER for “personalized guidelines” (see, e.g., Basu 2011)



Methods for CER

- Evidence generation
- Evidence synthesis
 - Decision analysis
 - Meta-analysis methods
 - » Classical frequentist meta-analyses
 - » Bayesian meta-analyses / Bayesian hierarchical models
 - » Meta-regressions – with either of the above
 - Meta-analysis evidence base:
 - » Aggregate data
 - » Combination of IPD and aggregate study level results
- Generalized evidence synthesis
 - Combining RCTs & non-RCTs, various outcome types (e.g. hazard ratio and count data for survival outcomes (Wood et al. 2010); individual patient data with aggregate data (Lambert 2007), data on humans / other species – with adequate methods



Evidence base considerations in Generalized Evidence Synthesis

- Impact of heterogeneity
 - A lot: can increase external validity
- Biases
 - Observational studies:
 - » Confounding – threat to internal validity
 - » Selection bias – threat to internal validity
 - RCTs:
 - » Publication Bias – threat to external validity
 - » “Sponsorship Bias” – threat to external validity



Examples of Generalized Evidence Synthesis

- Decision Modeling
- HTA of sequential use of anti-TNF therapies in RA in patients who failed 1st line anti-TNF therapy
- Bayesian hierarchical models of observational and RCT evidence



Comprehensive Decision Modeling

- Decision analysis models already combine multiple sources of evidence (and are really structured evidence syntheses)
- So, if a meta-analysis is to be used together with a decision analysis for CER, is there a need to restrict the evidence base?
- Use meta-analysis to summarize evidence base and build results directly into decision analysis models using Bayesian method (Cooper, Sutton, Abrams et al; 2004 and Cooper, Sutton and Abrams 2002)



GES: HTA of sequential use of anti-TNF therapies in RA

- Objective: to estimate ACR20, 50 and 70 responses on various treatments for the population of interest: patients with rheumatoid arthritis who failed 1st line anti-TNF therapy
- Treatments considered: adalimumab, etanercept, infliximab, abatacept, rituximab
- Evidence base: RCTs in anti-TNF failure population: REFLEX (rituximab), ATTAIN (abatacept), RADIATE (tocilizumab), GO-AFTER (golimumab) – with the latter two not in the scope of the NICE HTA
- RCTs in anti-TNF naïve patient population for all comparators



GES: HTA of sequential use of anti-TNF therapies

- Other evidence: 5 studies of various designs, providing comparative evidence on ACR responses (i.e. not only one “arm”)
 - ReAct study – prospective open-label, multicentre study on adalimumab; including 6610 patients, 899 had a history of previous etanercept and/or infliximab therapy; 12 wks
 - Buch 2005: (n=207, 12 wks) – prospective study of consecutive patients, treated with infliximab, qualified as nonresponders, staying on infliximab or switched to etanercept
 - Wick 2005 (n=27, 24 wks) – retrospective analyses of patients in STURE registry, switching between anti-TNFs
 - Nikas 2006 (n=24, 52 wks) – open label, single-center comparative study of switchers from infliximab to adalimumab to adalimumab only controls
 - Furst 2007 (n=27, 15 wks) – randomised, open-label, clinical trial of 28 patients with an inadequate response to etanercept, receiving background methotrexate and randomised 1:1 to discontinue etanercept and receive infliximab or to continue etanercept.



GES: HTA of Sequential use of anti-TNF therapies

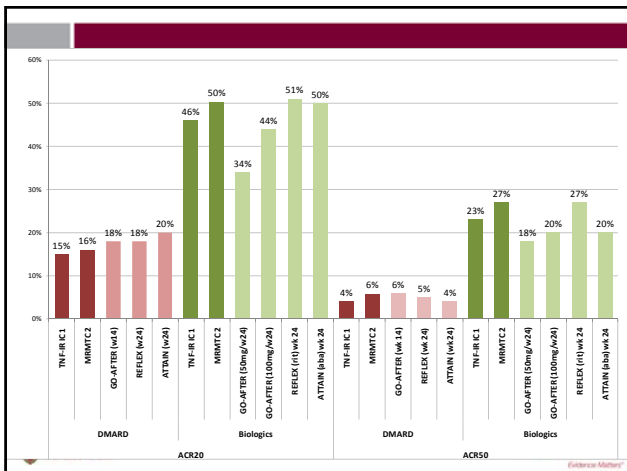
Manufacturer:	Abbott	SP	Roche	BMS
Included studies	29 RCTs + 5 non RCT comparative study	34 RCTs	3 RCTs and 18 RCTs	4 RCTs
Population covered	Late RA TNF-naïve and TNF- IR	Early & Late RA TNF-naïve and TNF- IR	TNF-IR population DMARD-IR population	TNF-IR population
# within/out of scope (per NICE)	2 / 34	2 / 32	2 / 1 18/0	2 / 2
Method of Analyses	Bayesian hierarchical model, with meta-regression	Network meta-analysis, adjusted for disease duration	Bayesian Indirect comparison	Bayesian Indirect comparison



GES: HTA of Sequential use of anti-TNF therapies

▪ Criticisms by NICE:

- NICE "could not quantify relative effect of a second TNF inhibitor in comparison with either conventional DMARDs or alternative biological DMARDs."
- use of data from populations beyond the scope of the appraisal [...] was inappropriate because of variability of studies from which the data were taken
- exchangeability of relative treatment effects between the included studies could not be assumed and thus the validity of the results [by manufacturers] was questionable
- NICE CE model employed various arbitrary assumptions



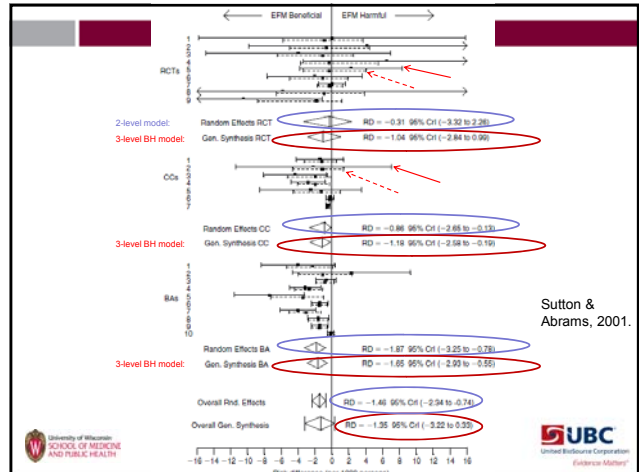
Examples of Published Generalized Evidence Syntheses

Authors, year	Disease area/treatment	Type of studies	Type of Model	Outcome	Covariate or bias adjustment
Grines et al 2008	AngioJet thromboectomy vs PCI	2 RCT + 9 non-RCT	3-level hierarchical model	Odds ratio, short term mortality rate	no
McCarron et al. 2010	Abdominal Aortic Aneurysms: EVAR vs OSR	4 RCT / 40 observational (75 with covariate imputation)	3-level hierarchical model + various mortality	Odds ratio, peri-operative mortality	age, gender, CVD study or study arm level
Prevost et al. 2000	Breast cancer screening	4 RCT / 4 "observational"	2 and 3-level hierarchical model	Relative risk, BC mortality	study level, age
Sampath et al. 2007	Loop diuretics	5 RCT / 8 non-RCT	3-level hierarchical model	Risk ratio, mortality	age, % males, control arm risk, quality score, study termination date
Sutton & Abrams 2001	Electronic fetal heart rate monitoring	9 RCT / 7 comparative cohort / 10 before-after	3-level hierarchical model + various mortality	Risk difference, perinatal mortality	publication year



GES: Sutton and Abrams 2001

- Background: widespread use of electronic fetal heart rate monitoring (EFM) in early 1970s in UK coincided with a decrease in overall perinatal mortality rate
- Evidence-based reviews conclude that EFM has not been shown to reduce perinatal mortality
- Conclusion ignores observational evidence: 17 studies
- Objective to include all evidence in a Bayesian three level hierarchical model with random effect for study type



McCarron et al. 2010: EVAR vs OSR

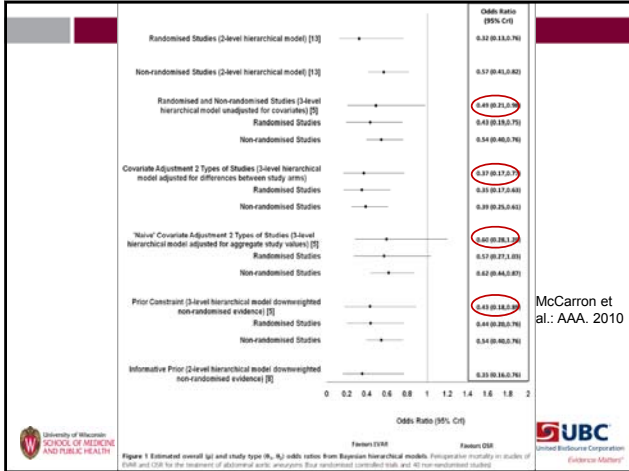
- Applied Bayesian hierarchical model to a subset of studies from a systematic review of endovascular (EVAR) and open surgical repair (OSR) in the treatment of abdominal aortic aneurysms (AAA).
- Examined odds ratios of peri-operative mortality
- Tested covariates for aggregate values for patient characteristics (e.g., mean age) within studies and two approaches for down-weighting biased evidence

McCarron et al. 2010: AAA

Table 1 Covariate Data: Average Imbalance between Study Arms

Study Type	Average Difference	
	(EVAR-OSR)	(EVAR-OSR)
Non-randomised		
Male (proportion)	0.09	0.10
Age (years)	2.60	2.53
Cardiac disease (proportion)	0.13	0.14
Pulmonary disease (proportion)	not considered as missing in 43% of the 75 non-randomised studies	0.10
Renal disease (proportion)	not considered as missing in 54% of the 75 non-randomised studies	0.05
Randomised		
Male (proportion)	0.05	0.05
Age (years)	0.82	0.82
Cardiac disease (proportion)	0.05	0.05
Pulmonary disease (proportion)	not considered as missing in 25% of the 4 randomised studies	0.13
Renal disease (proportion)	not considered as missing in 50% of the 4 randomised studies	0.07

a.male, age, cardiac disease, b.male, age, cardiac disease, pulmonary disease, renal disease.



Generalized Evidence Synthesis – Findings

- Change in estimate of mean effects (overall, and by study) with 3-level model depends on level of disparity between results by study type
 - May result in wider credible interval for mean effect if very different results by study type
- Priors for means and variances can be adjusted to give greater weight to RCTs or assert that observational studies will be more heterogeneous / biased
- No standard method for choice of priors
- Impact of priors varied across studies
 - Priors for between study heterogeneity seem influential
- No adjustment for publication bias, selection bias

Covariate Adjustment

- Limited by data availability
- Selection requires prior knowledge of key covariates
- Covariate adjustment needs to be done at the appropriate level: at aggregate study level does not account for covariate imbalance in non-randomized studies
- Imputation of additional covariates may have large impact

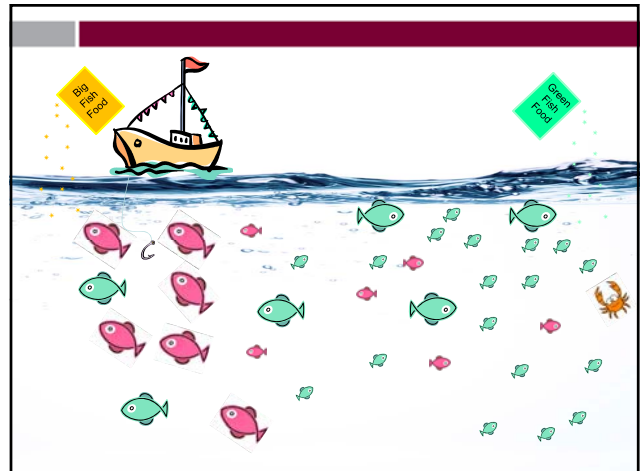


Illustration by Simulation

- Major Goal: use simulated environment to show potential benefits of using a broad evidence base
- Secondary Goal: walk through simple meta-regression MTC WinBUGS code
- Overall message: restricting synthesis to only evidence of high internal validity (i.e., RCTs) is no guarantee of suitability for CER when there are threats to generalizability
- Meant to illustrate – does not prove including observational evidence is always a good idea



Context

- Monte Carlo methods used to simulate the process of evidence generation
 - Three hypothetical treatments: 1, 2 and 3
 - Observable (C) and unobservable (U) individual characteristics affect outcomes differently for the 3 treatments
 - Goal is to compare effectiveness (success rate) of all 3 treatments for a subgroup defined by C

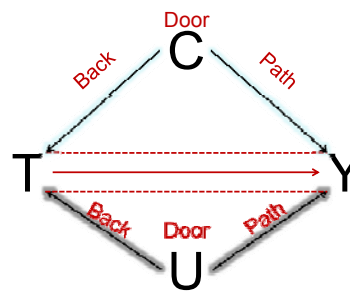


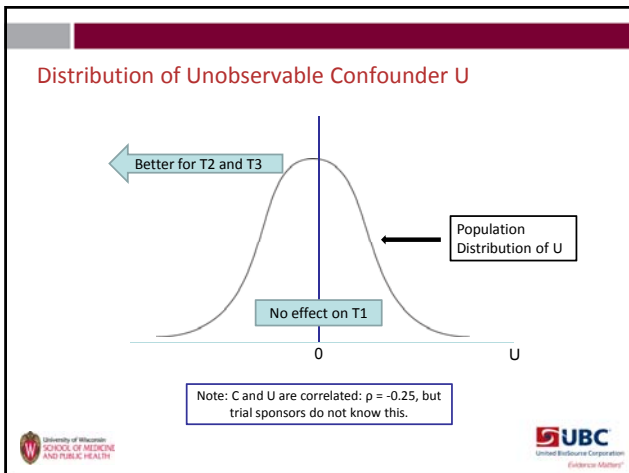
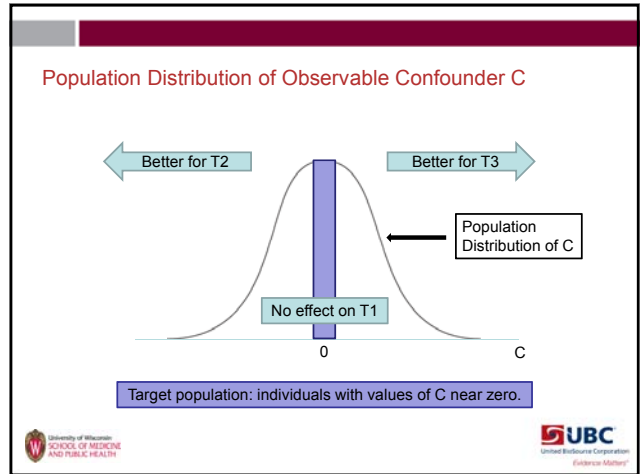
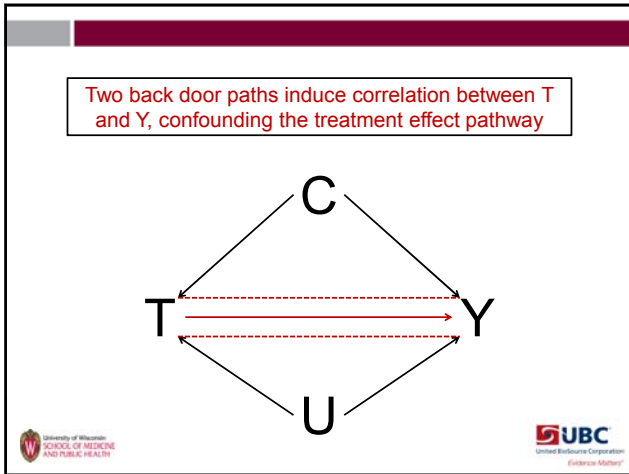
Simulated Observational Studies

- Our simulated observational studies suffer from endogenous treatment selection:
 - Individual patients more likely to select treatment with the highest probability of success for them
 - Selection probability depends both on observable (C) and unobservable (U) factors
 - Both of these factors also affect outcomes – creates treatment selection bias



Directed Acyclic Graph (DAG) (Pearl 1995; Greenland, Pearl and Robins 2002)





- Simulated Randomized Controlled Trials
- In our simulated RCTs, treatments are assigned randomly: no treatment selection bias
 - We also assume the trial is double-blinded; no attrition or other threats to internal validity
 - However, we simulate two threats to external validity (generalizability):
 - Publication Bias - Statistically significant and/or favorable results are more likely to be published:

Significant Negative: 0%	Insignificant Negative: 25%
Significant Positive: 100%	Insignificant Positive: 50%
 - “Sponsorship Bias” - Propensity for sponsors to conduct trials with high “assurance” (sponsored intervention statistically significantly better)
- The University of Wisconsin School of Medicine and Public Health logo and UBC logo are at the bottom.

Simulating Sponsorship Bias

- Assume Treatment 1 is “placebo” and that Treatments 2 and 3 have sponsors who decide:
 - Whether their trial will be head to head (2 vs 3)
 - or placebo controlled (2 vs 1 or 3 vs 1)
 - or whether to do any trial at all!
- Assume sponsors of treatments 2 and 3 have a limited research budget:
 - Each can afford 10 2-arm trials of size 50-150 per arm



Targeting Trials to Subpopulations

- Define a sample population with mean $C = \mu_C$
- Sponsors believe effectiveness is:
 - T1: $\Pr(\text{Success}) = \Phi(-0.4)$
 - T2: $\Pr(\text{Success}) = \Phi(0.2 - 0.6 \mu_C)$
 - T3: $\Pr(\text{Success}) = \Phi(0 + 0.4 \mu_C)$
- Note: at desired population $\mu_C=0$, actual rates:
 - T1: $\Pr(\text{Success}) = 0.34$
 - T2: $\Pr(\text{Success}) = 0.58$
 - T3: $\Pr(\text{Success}) = 0.50$

With odds ratios:

- 2 versus 1: = 2.62
- 3 versus 1: = 1.90
- 3 versus 2: = 0.73



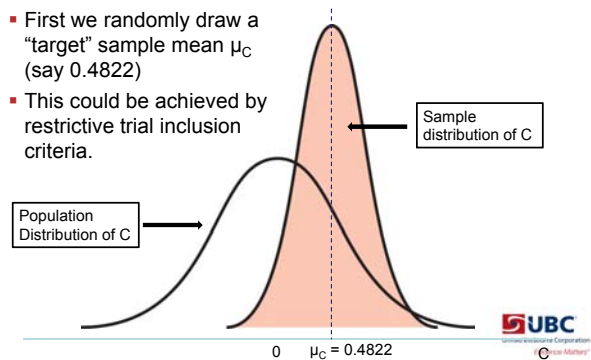
Deciding Which RCTs Get Done

- Sponsor 2:
 - Given μ_C and a fixed (budget determined) sample size and assuming 5% Type I error rate, calculate the power for trials of 2 vs. 1 and 2 vs. 3.
 - Choose the trial type with the greatest power.
 - Choose no trial if power < 80%
- Sponsor 3:
 - Do the analogous exercise for trials of 3 vs. 1 and 3 vs. 2
- Repeat using different randomly generated values of μ_C until 10 trials have been conducted by each sponsor.



An example...

- First we randomly draw a “target” sample mean μ_C (say 0.4822)
- This could be achieved by restrictive trial inclusion criteria.



Example (continued)

$$T1: \Pr(\text{Success}) = \Phi(-0.4) = 0.345$$

$$T2: \Pr(\text{Success}) = \Phi(0.2 - 0.6 \times 0.4822) = 0.464$$

$$T3: \Pr(\text{Success}) = \Phi(0 + 0.4 \times 0.4822) = 0.576$$

- Suppose (perhaps due to budget constraints), the sponsors consider trials with sample size of 150 patients per arm.
- From Sponsor 2's point of view, a 2 vs. 3 trial would not be considered and a 2 vs. 1 trial would have low assurance (power of 55.9%); so no trial is done by Sponsor 2.



Example (continued)

- But from Sponsor 3's point of view, in the subpopulation with $\mu_c = 0.4822$:

3 vs. 1 trial has 98.2% power 3 vs. 2 trial has 49.0% power

- So, a simulated 3 vs. 1 trial is done:
 - Arm 1: Actual $\mu_c = 0.405$, 57/150 successes
 - Arm 2: Actual $\mu_c = 0.368$, 84/150 successes
 - 95% CI for difference in proportion: (0.250 - 0.475, $P < 0.001$)
- The favorable results are published with probability 1.



Simulating Observational Studies

- Conduct 3 observational studies with sample size 500-2,000
- Sample drawn from population distribution of C and U and no publication or sponsorship bias
- Biased treatment selection with probability proportional to actual probability of success for each patient i .

$$T1: \Pr(\text{Success}) = \Phi(-0.4)$$

$$T2: \Pr(\text{Success}) = \Phi(0.2 - 0.6 C_i - 0.2 U_i)$$

$$T3: \Pr(\text{Success}) = \Phi(0 + 0.4 C_i - 0.2 U_i)$$

Implies patients and physicians have "private knowledge" of U_i and are more likely to choose treatments with the best chance of success.



Three Approaches to Evidence Synthesis



- After 10 RCTs from each sponsor and 3 observational studies are complete, we synthesize the evidence using one of 3 approaches:

1. Full MTC meta-regression evidence synthesis

- Use all 20 RCTs and 3 observational studies
- Unconstrained baseline (weakest assumption)
- "Mixed effect" model for treatment effects
 - Random effect reflects between trial variation due to unobserved factors
 - Fixed effect adjustment of treatment effect for trial arm-level μ_c
 - Implies "conditional exchangeability" – studies are exchangeable conditional on arm-level μ_c



WinBUGS code



Unconstrained Baseline

Random treatment effect

Meta regression: treatment effect interaction with observed arm-level covariate C[i]



```

model{
  #Likelihood
  for(i in 1:N_ARMS){
    SUCCESS[i] ~ dbin(p[i],N_PATIENTS[i])
    logit(p[i]) <- mu[STUDY[i]] + delta[i]*I>equals(TREATMENT[i],1)
    delta[i] ~ dnorm(mu_delta[i],prec)
    mu_delta[i] <- b[1]*C[i]*equals(TREATMENT[i],2) +
      b[2]*C[i]*equals(TREATMENT[i],3) +
      r[TREATMENT[i]] - r[CONTROL[i]]
  }
  #Priors
  prec <- pow(sd,2)          #Random Effect Variance
  sd ~ dunif(0,10)
  for(j in 1:N_STUDIES){
    mu[j] ~ dnorm(0,0.001)   #Study Baselines
  }
  r[1] <- 0
  for (k in 2:N_TREATMENTS){
    r[k] ~ dnorm(0,0.001)    #Treatment Effects on Log-Odds Scale
  }
  #Prediction at C=0
  log(or21) <- r[2]
  log(or31) <- r[3]
  log(or32) <- r[3] - r[2]
}
  
```



Three Approaches to Evidence Synthesis

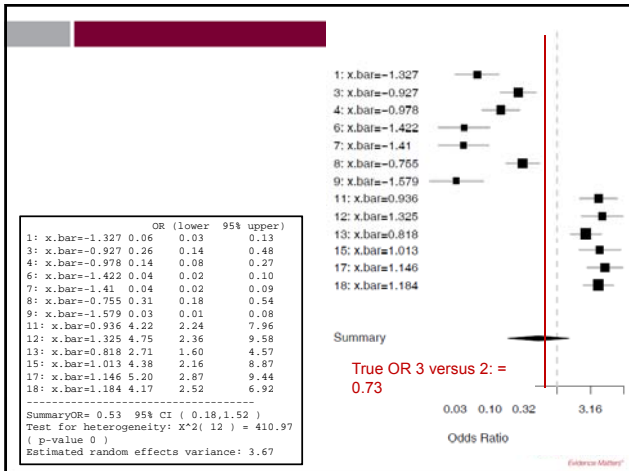
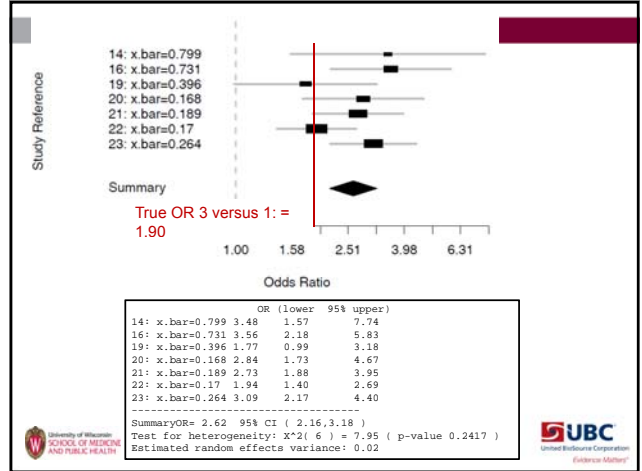
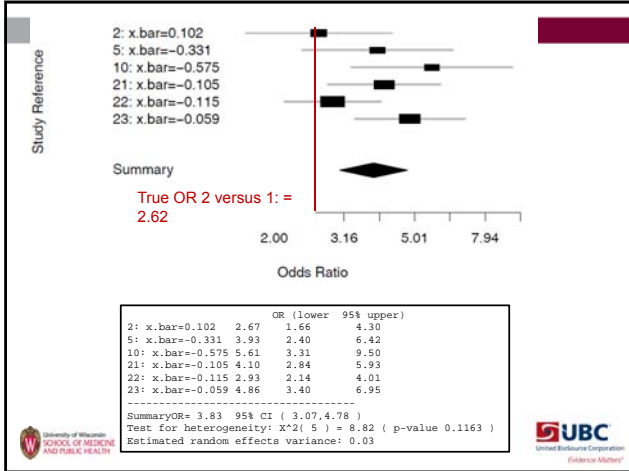
2. MTC meta-regression
 - Same as MTC Mixed Effect Meta-Regression generalized evidence synthesis, except...
 - Exclude observational studies
3. Narrow-Scope MTC
 - Exclude observational studies
 - Exclude RCTs with $\mu_C < -1$ or $\mu_C > 1$
 - No meta-regression to control for μ_C

Comparison to Traditional Pairwise Meta-Analysis

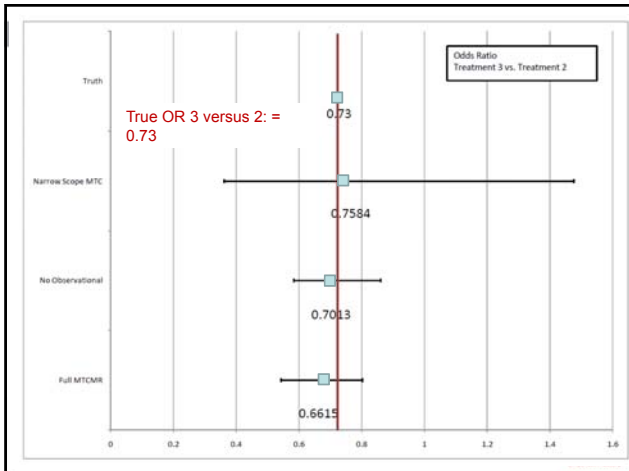
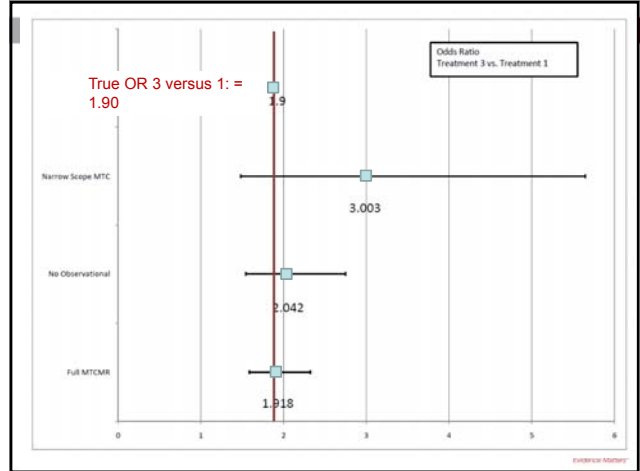
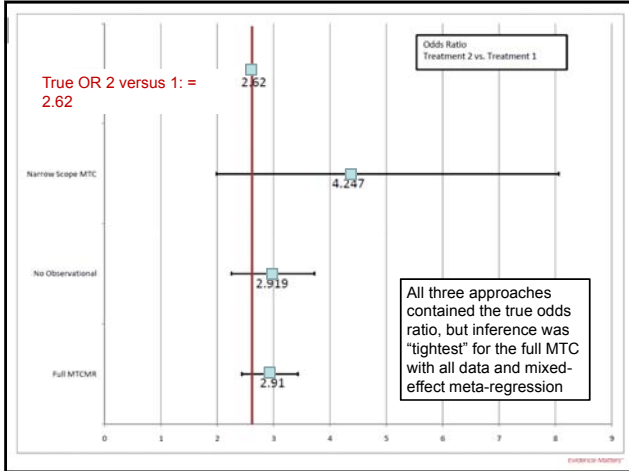
- To connect to classical meta-analysis, consider the following results using the method of DerSimonian and Laird...



Now, the 3 Bayesian Evidence Synthesis Approaches...

UBC
United Business Corporation
Evidence Matters

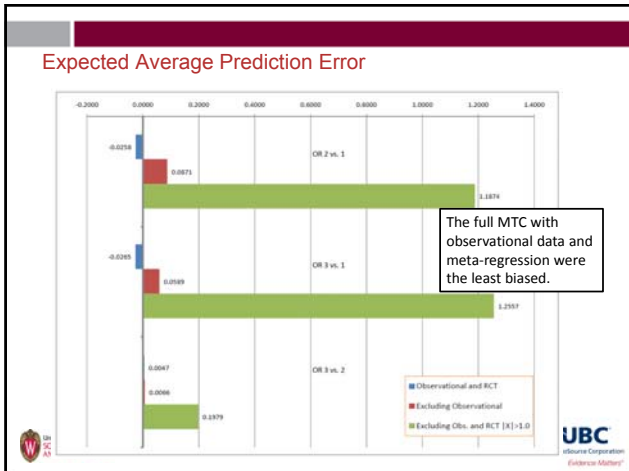
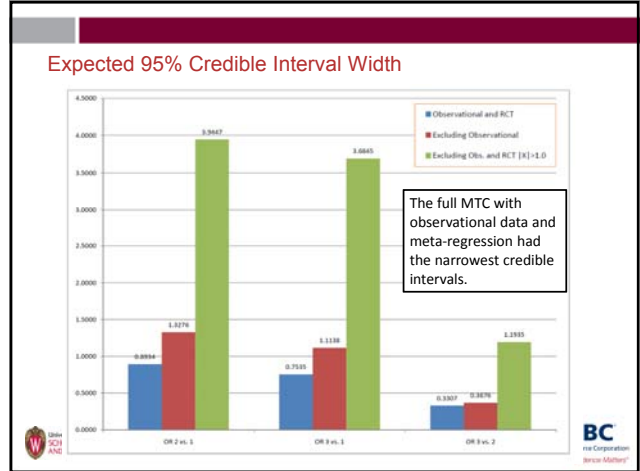
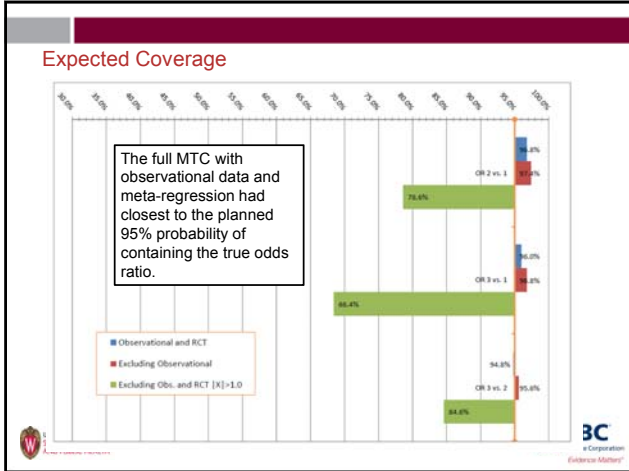


Rinse and Repeat...

- Repeat the exercise a large number of times (say, 500) to investigate the following:
 - Coverage:** what percent of the time do the 95% credible intervals as constructed actually contain the true odds ratio for each method?
 - Credible Interval Width:** which method provides the most precise inference
 - Average Prediction Error:** which method gets closest to the true odds ratios on average

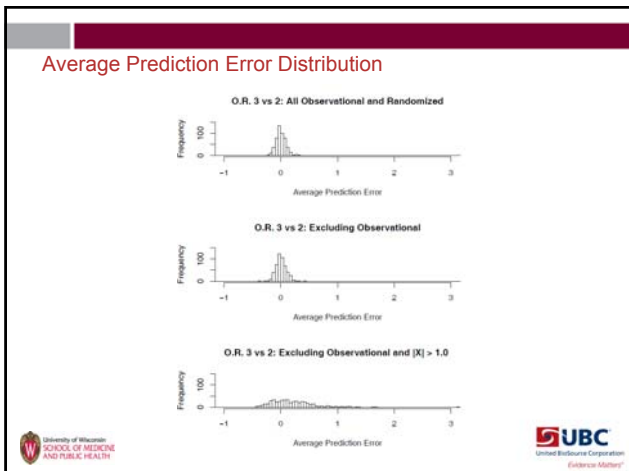
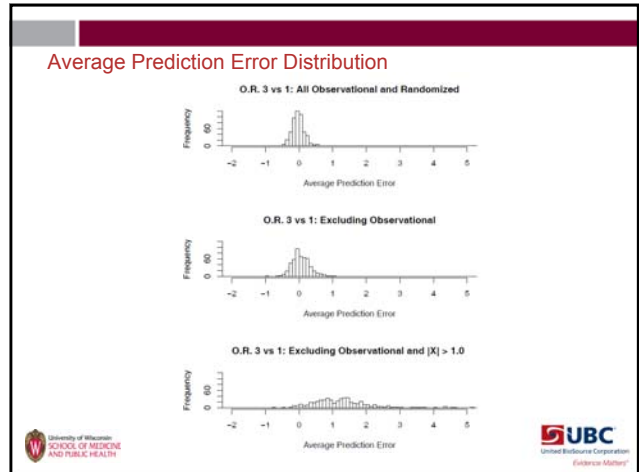
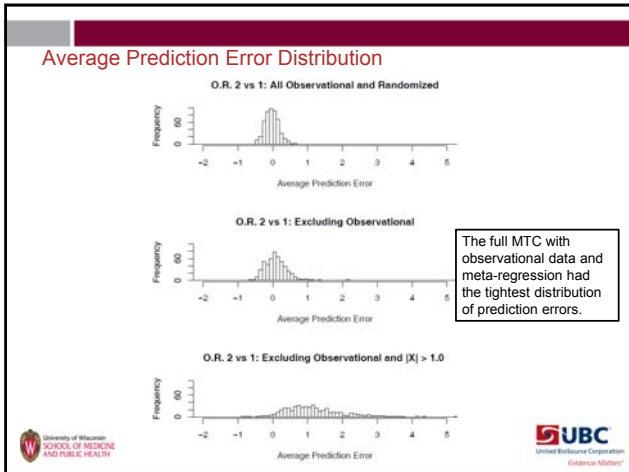
University of Wisconsin SCHOOL OF MEDICINE AND PUBLIC HEALTH

UBC United BioSource Corporation Evidence Maturity



How reliable are the 3 methods?

- Expected Average Prediction Error tells us whether we tend to get the answer right "on average"
- But we also care about minimizing how far off we are likely to be in any given analysis.
- Looking at the Distribution of Average Prediction Error is somewhat analogous to the standard error of the mean.



Summary: Lessons Learned from This Simulation

- Automatically excluding observational evidence because of threats to internal validity may not necessarily improve inference given an entire body of evidence... and indeed can make it worse.
- Even if all evidence is internally valid, evidence synthesis may not present good inference for CER if publication and sponsorship bias are present.
- As in all things, a "rule of reason" applies.

University of Wisconsin SCHOOL OF MEDICINE AND PUBLIC HEALTH

UBC United BioSource Corporation Evidence Matters™

Discussion



Helpful References

- Basu A. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics* 2011; In Press.
- Cooper NJ, Sutton AJ, Abrams KR, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: A Bayesian approach. *Health Economics* 2004; 13(3) 203-226
- Cooper NJ, Sutton AJ, Abrams KR. Decision analytical economic modeling within a Bayesian framework: Application to prophylactic antibiotics use for caesarean section. *Statistical Methods in Medical Research* 2002; 11: 491-512.
- Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in medicine*. 2009;28(14):1861–1881.
- Nixon RM, Bansback N, Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Statistics in medicine*. 2007;26(6):1237–1254.

