EvidenceSynthesis

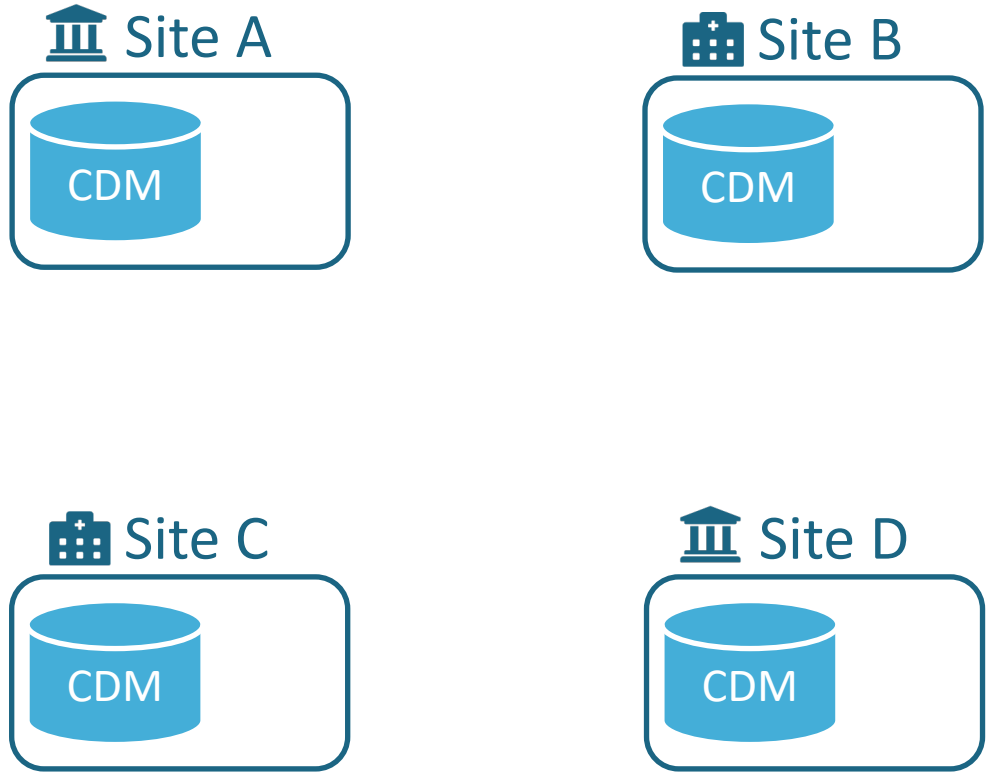
An R package for combining causal effect estimates without sharing individual person data

Martijn Schuemie

Yong Chen

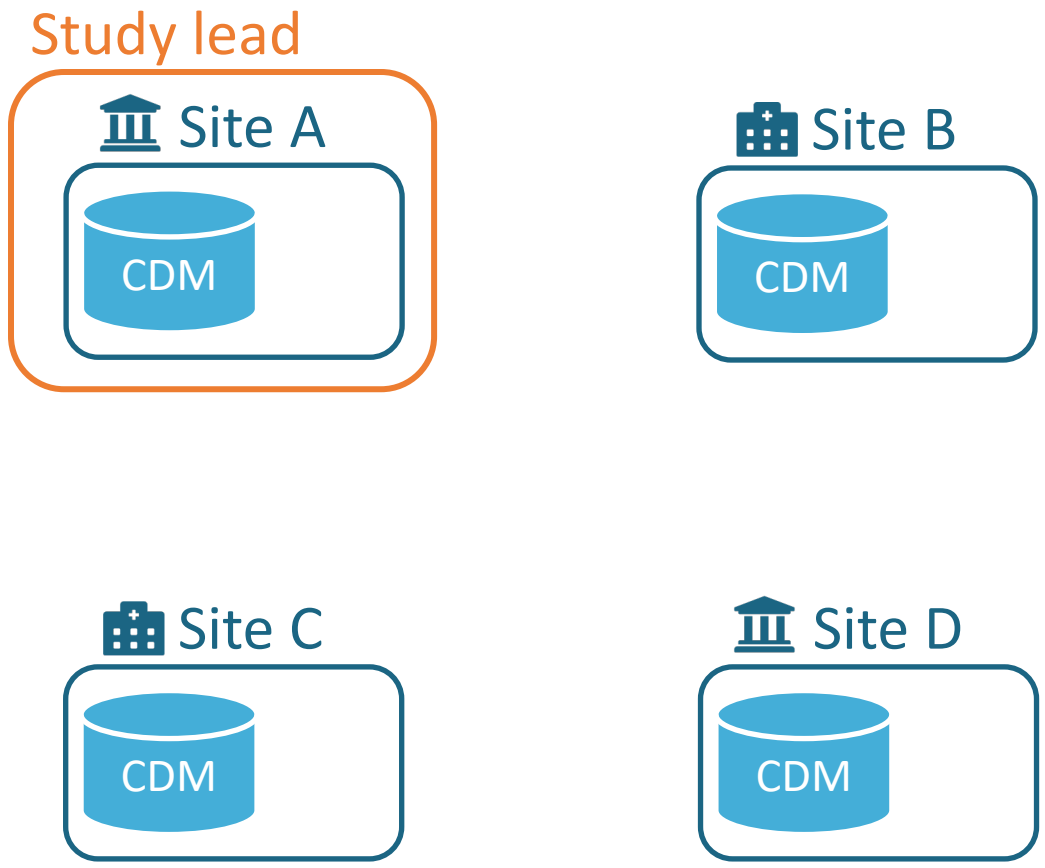
Distributed Research Network

- Multiple sites with data
  - Hospital EHRs (Electronic Health Records)
  - Administrative Claims
- Patient-level data cannot be shared
- Each site uses the Common Data Model (CDM)



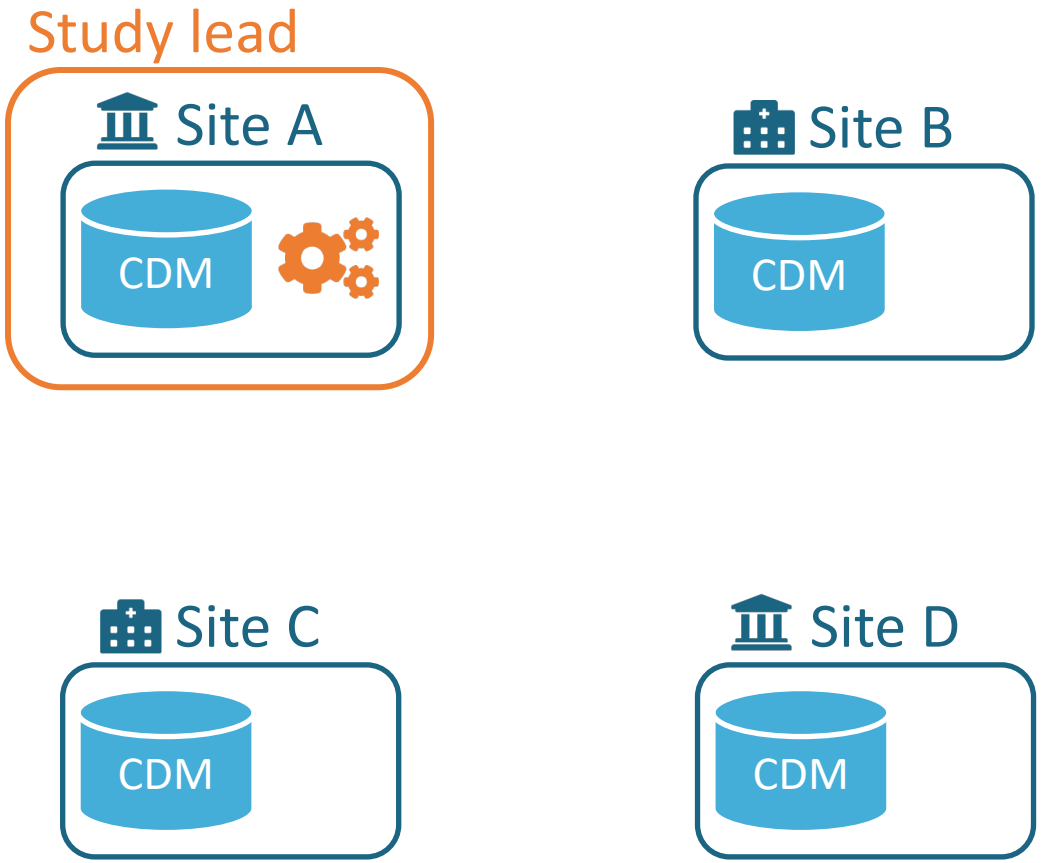
Distributed Research Network

- A site can lead a study



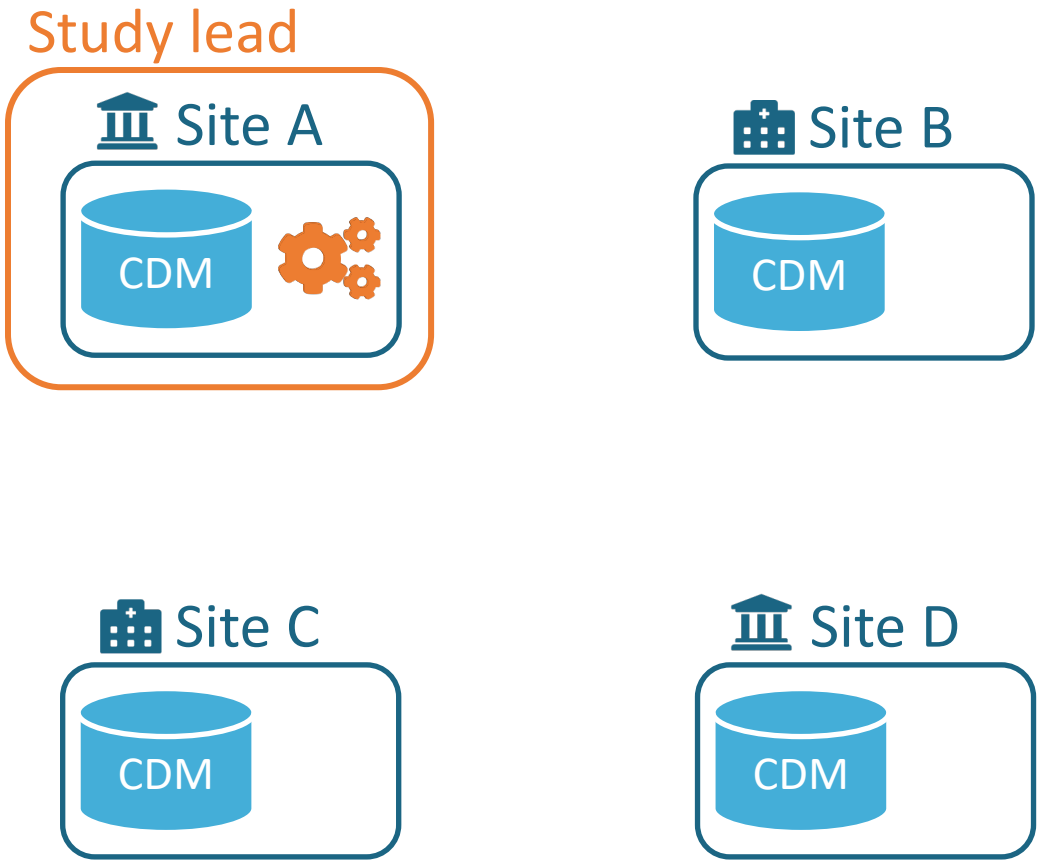
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally



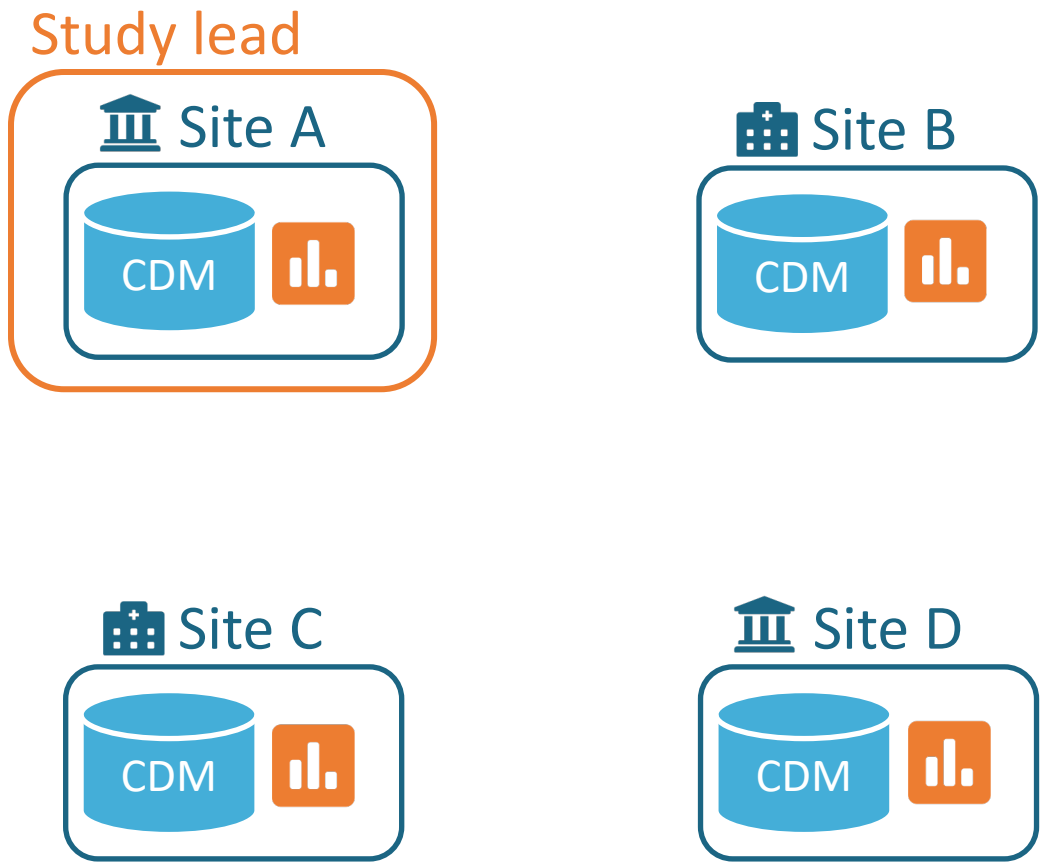
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants



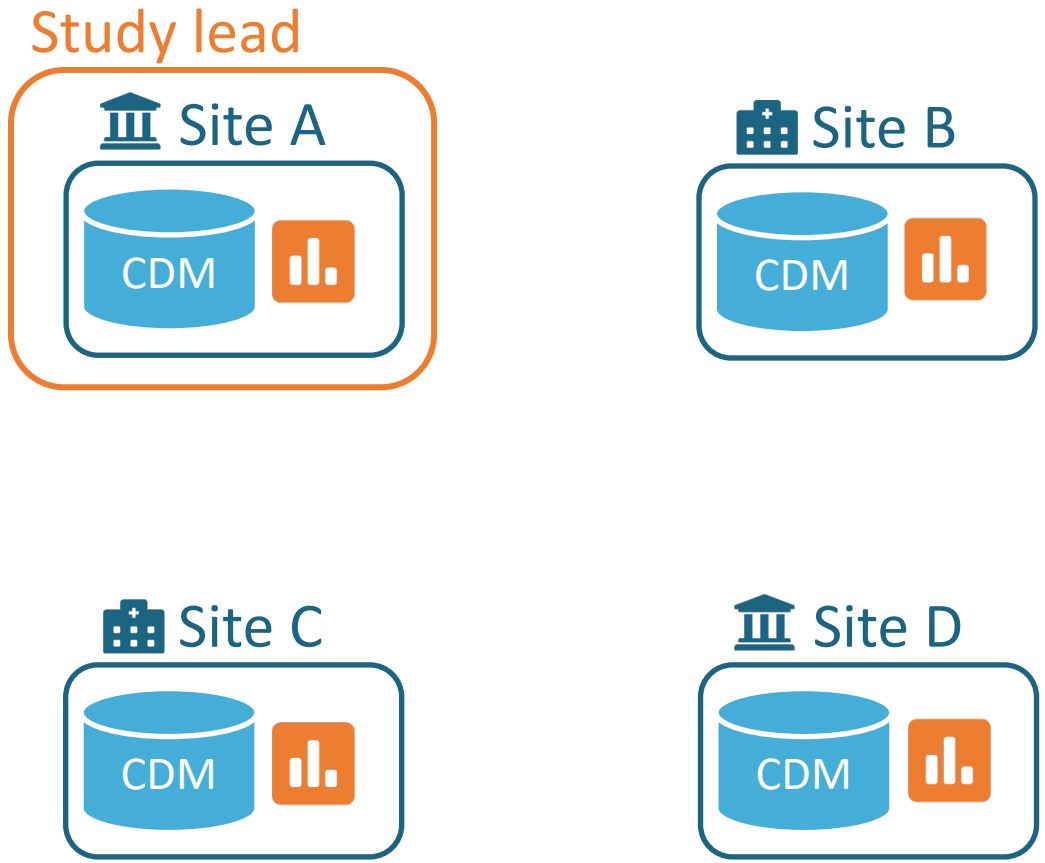
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)



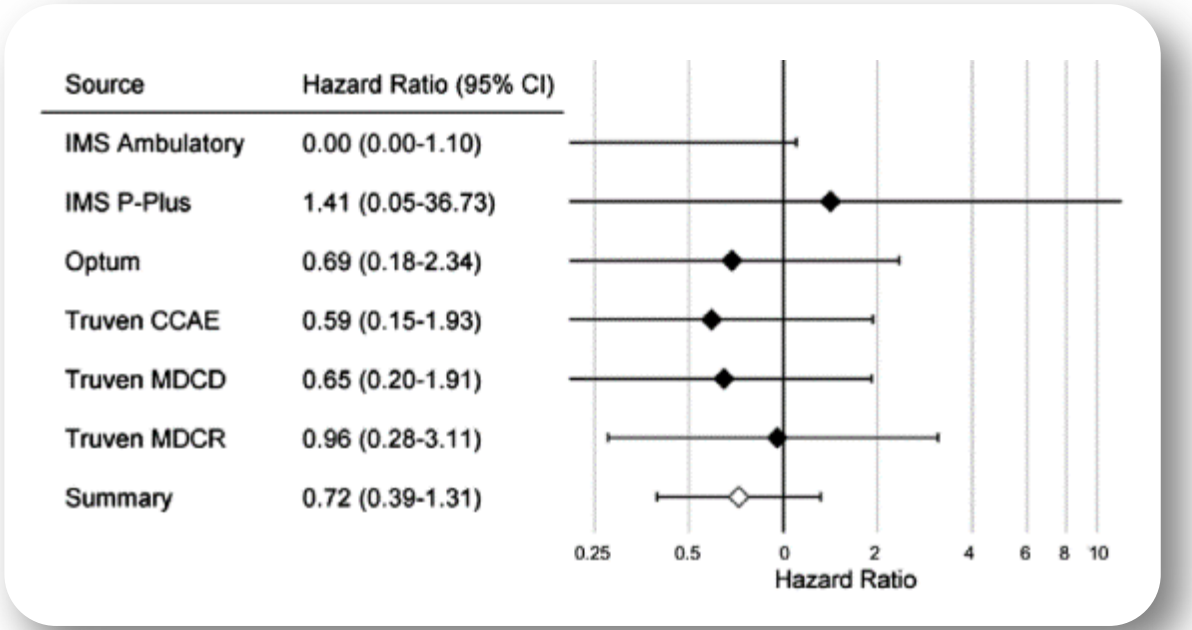
Distributed Research Network

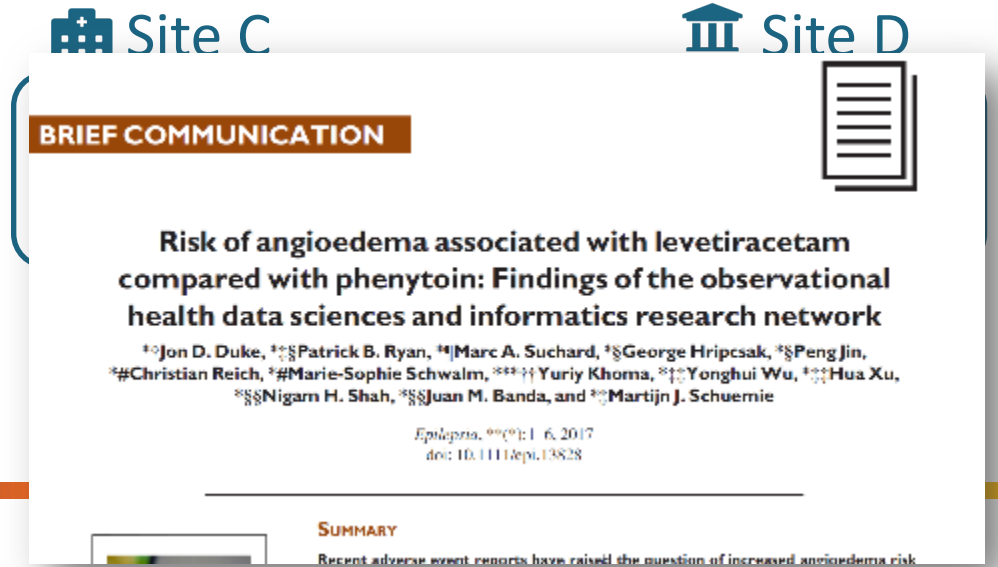
- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)
- Results are send back to lead site



Distributed Research Network

- A site can lead a study
- Analysis code is developed locally





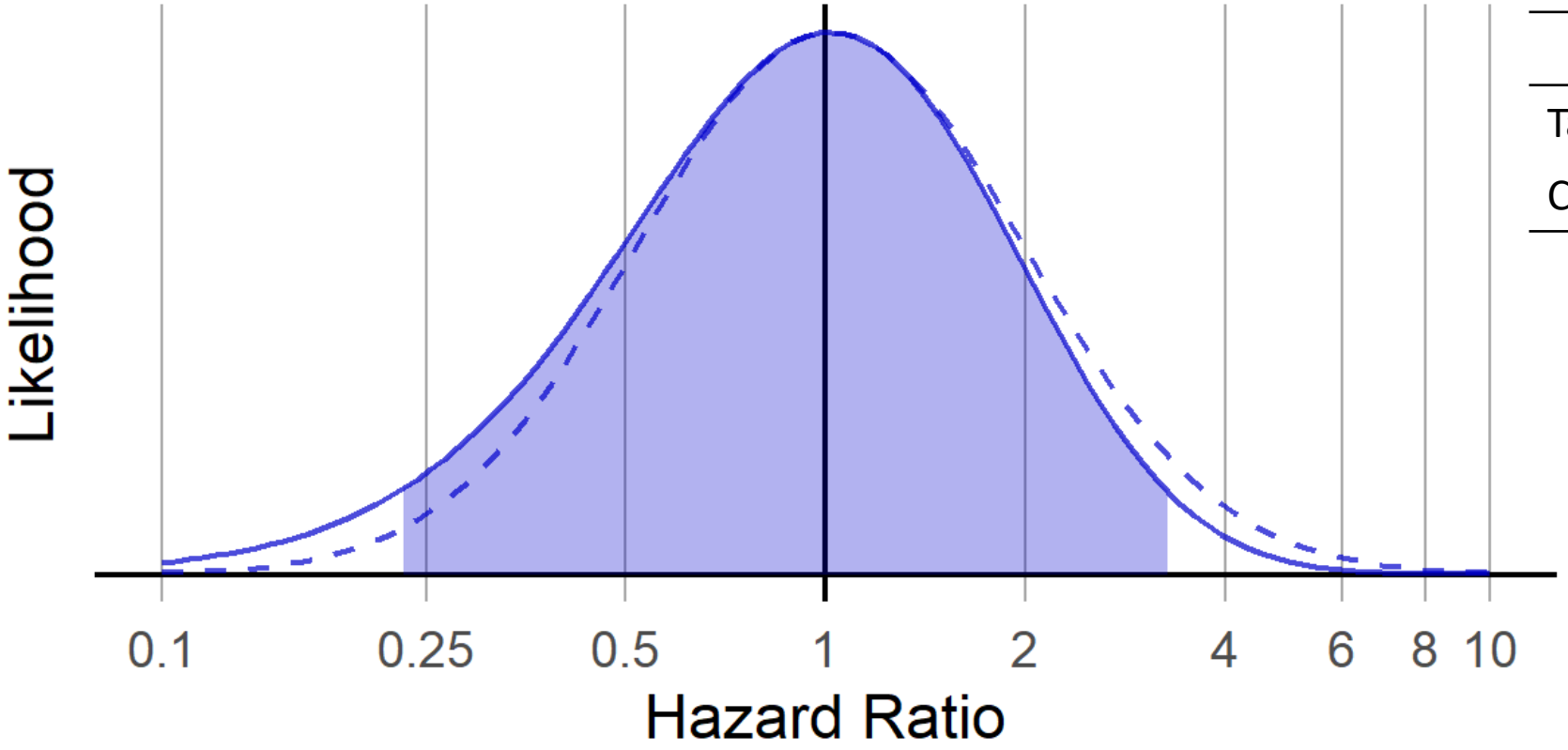
What to share when estimating causal effects?

- Cannot share patient-level data
- Usually **propensity-score stratified time-to-event** or **conditional Poisson regression**: no 2-by-2 tables
- Point-estimates + standard errors?

Normal assumption violated when counts are low

Hazard Ratio = 1.02 (0.27 – 3.78) Assuming normal distribution

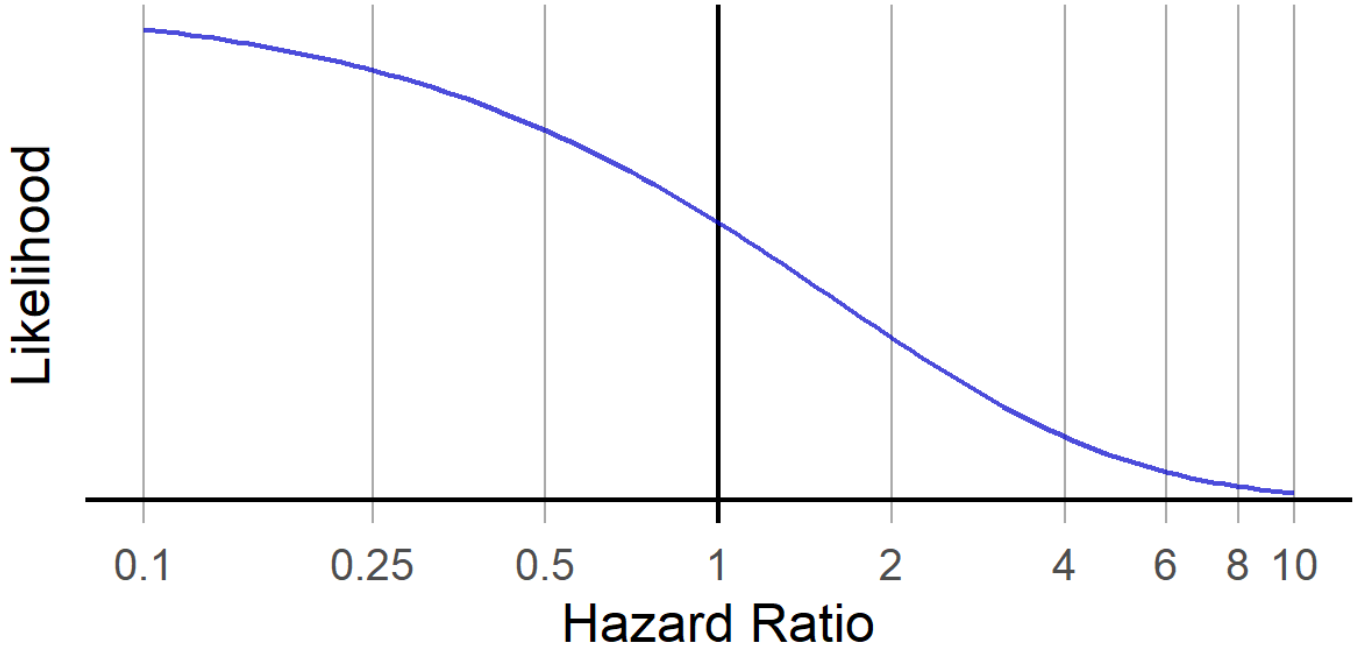
Hazard Ratio = 1.02 (0.22 – 3.31) No shape assumption



	Subjects	Outcomes
Target	22,002	3
Comparator	130,200	25

* Real data, no simulation

Even more when counts are 0



	Subjects	Outcomes
Target	2,834	0
Comparator	15,168	10

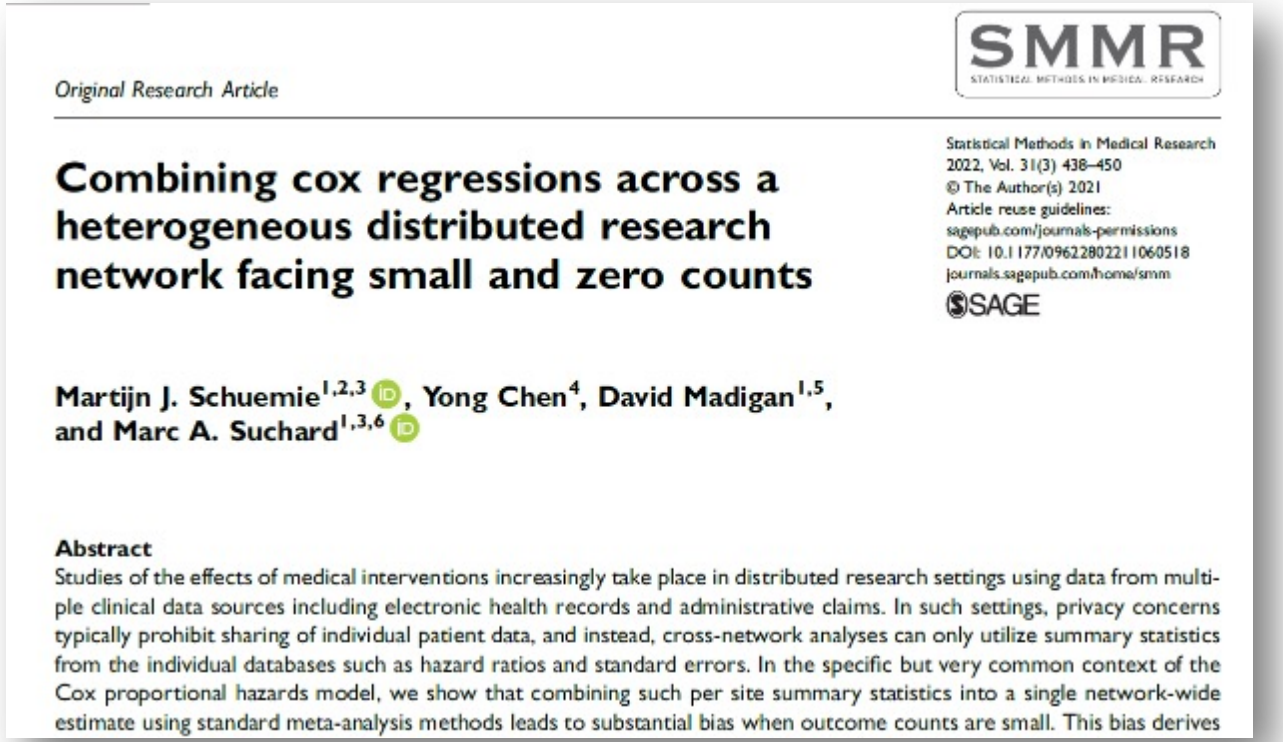
* Real data, no simulation

Solution – likelihood profiling

Data sites share actual shape of likelihood, instead of just the hazard ratio + confidence interval

Key ideas

- directly approximate the shape of the per-site log likelihood function
- communicate the parameters that summarize the shape of likelihood function from each site



Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Combining cox regressions across a heterogeneous distributed research network facing small and zero counts

Statistical Methods in Medical Research
2022, Vol. 31(3) 438–450

DOI: 10.1177/09622802211060518
journals.sagepub.com/home/smm
SAGE

Martijn J. Schuemie[1,2,3], Yong Chen[4], David Madigan[1,5], and Marc A. Suchard[1,3,6]

Abstract

Studies of the effects of medical interventions increasingly take place in distributed research settings using data from multiple clinical data sources including electronic health records and administrative claims. In such settings, privacy concerns typically prohibit sharing of individual patient data, and instead, cross-network analyses can only utilize summary statistics from the individual databases such as hazard ratios and standard errors. In the specific but very common context of the Cox proportional hazards model, we show that combining such per site summary statistics into a single network-wide estimate using standard meta-analysis methods leads to substantial bias when outcome counts are small. This bias derives

Methods

- Normal approximation:

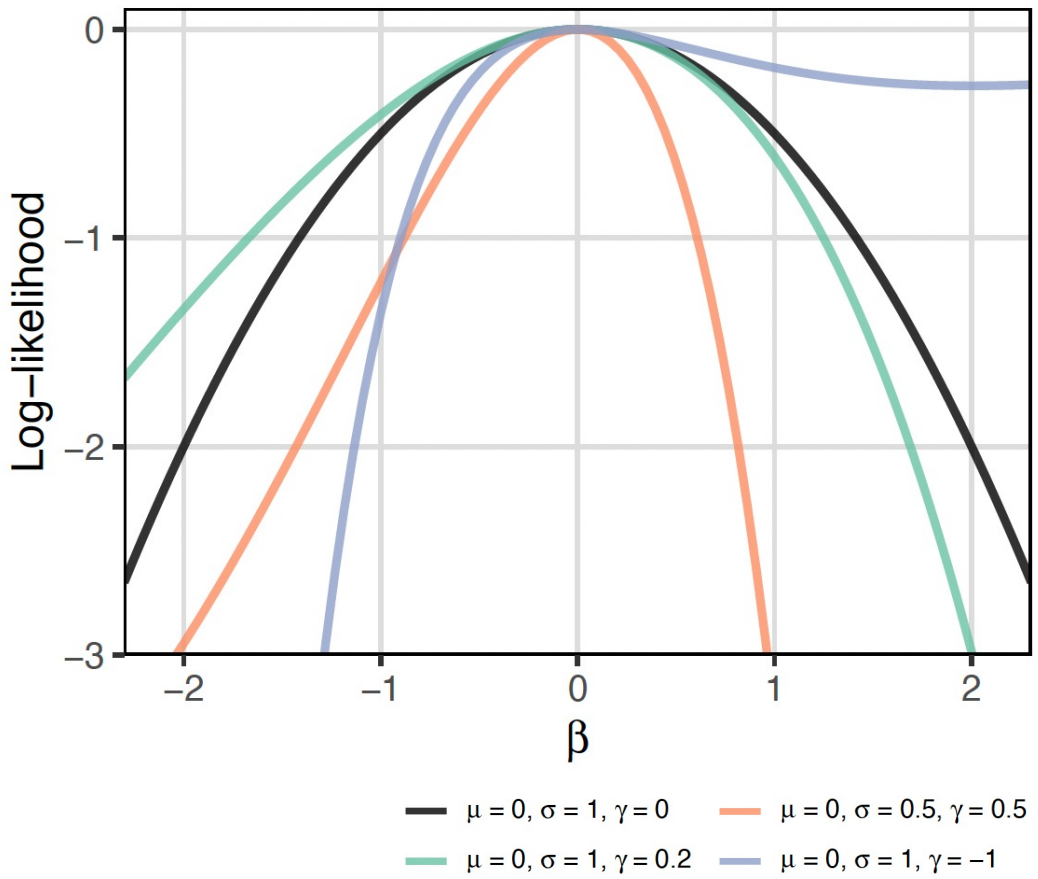
$$\phi(\beta, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\beta-\mu)^2}{2\sigma^2}}.$$

  - Estimate $\hat{\mu}$ and $\hat{\sigma}$
- Skew-normal approximation
  - Generalized normal distribution for a degree of skewness
- Custom approximation:
  - For severe skewness, a novel "custom function":

$$ln(f(\beta, \mu, \sigma, \gamma)) = -\frac{(\beta-\mu)^2}{2\sigma^2} e^{\gamma(\beta-\mu)}$$

  - Estimate $\hat{\mu}, \hat{\sigma}$, and $\hat{\gamma}$

The custom approximation function, under various parameter choices



Solution – likelihood profiling

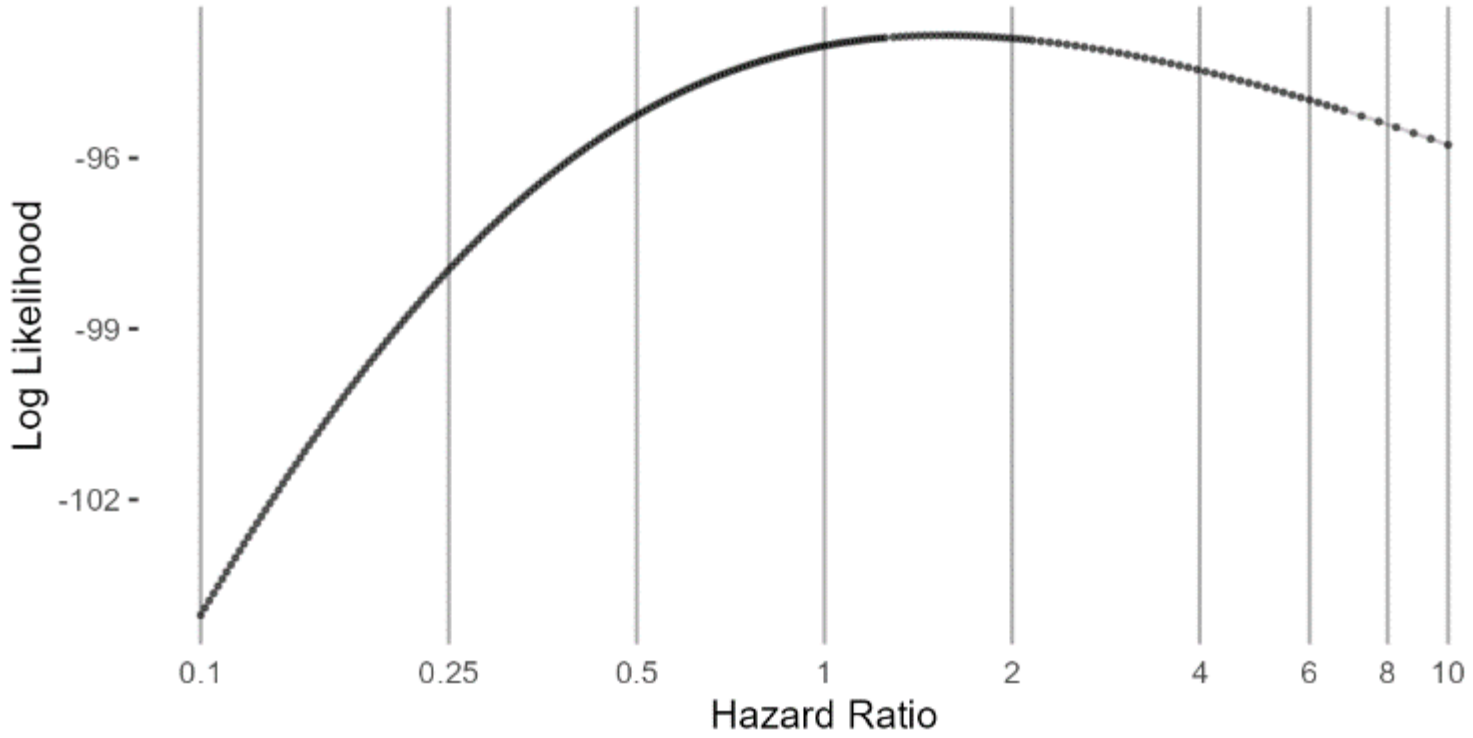
- Grid
  - communicate the (log) partial likelihood function by sampling values at predefined points in a one-dimensional grid of hazard ratios over a plausible range
    - For example, we define the grid from a log hazard ratio as 1,000 equally spaced points spanning log(0:1) to log(10).
  - zero counts do not impact this approximation and increasing the grid size can provide an arbitrarily high-quality approximation.

Solution – likelihood profiling

Data sites share actual shape of likelihood, instead of just the hazard ratio + confidence interval

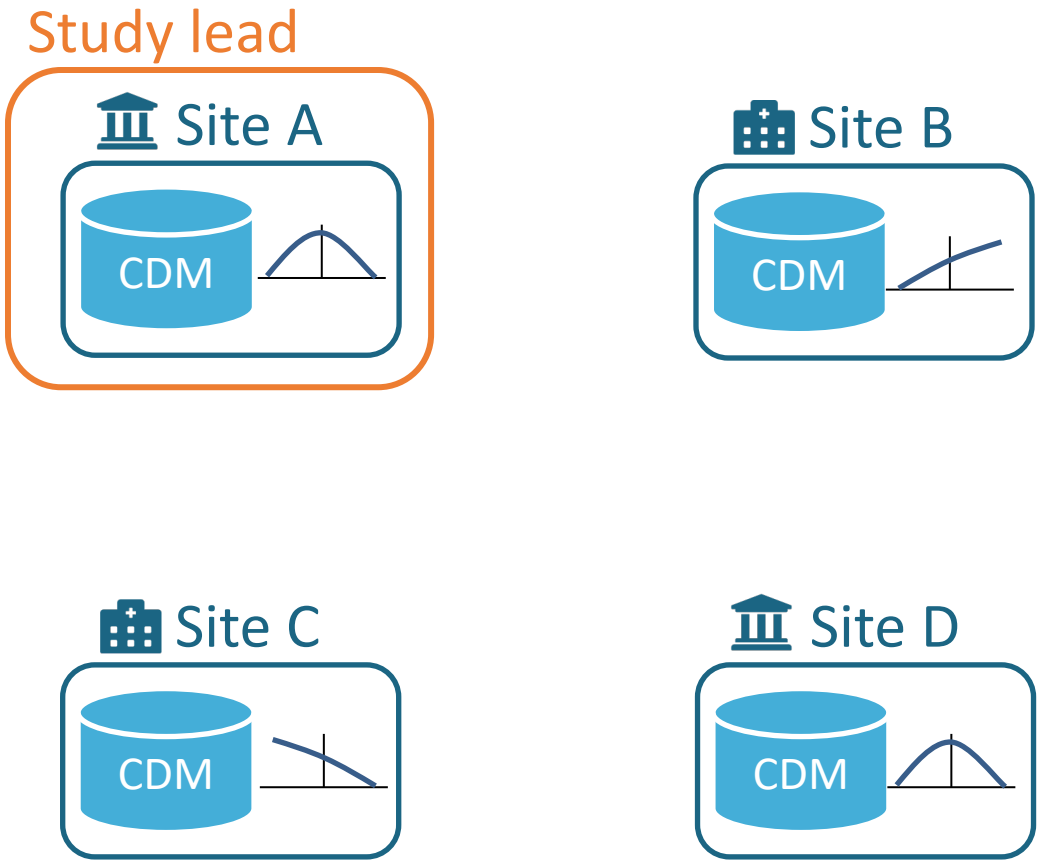
Current best-practice:

- Adaptive grid



Distributed Research Network

- Data sites share shape of likelihood



Evidence Synthesis

The **EvidenceSynthesis** package implements

- Fixed-effects model
- Random-effects model using a Bayesian approach.
  - Uses the BEAST MCMC engine.

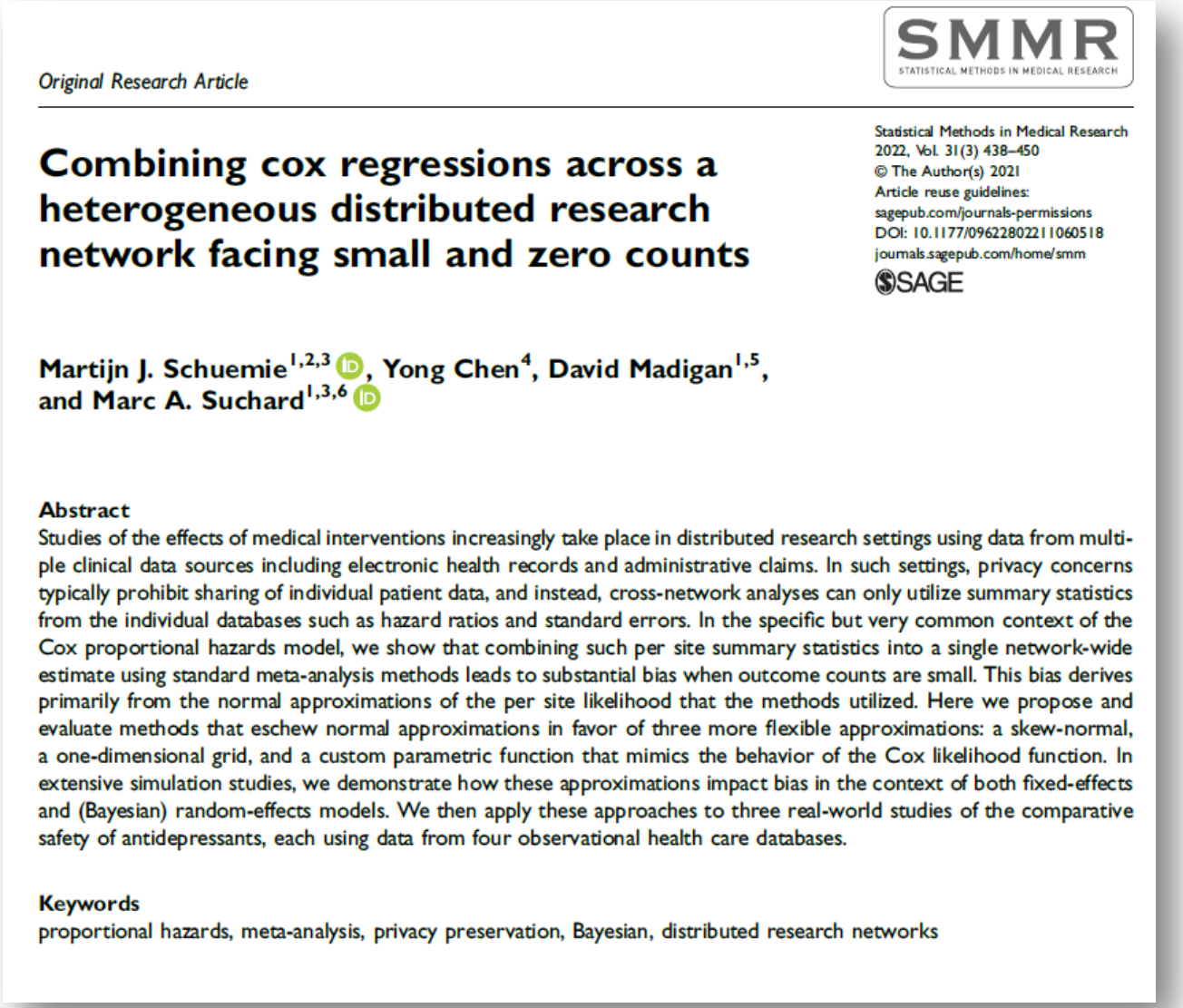
BEAST

Bayesian Evolutionary Analysis Sampling Trees

Evidence Synthesis in ASSURE / OHDSI

Two important differences with vanilla meta-analysis

- Bayesian random effects
- Likelihood profiling



Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2022, Vol. 31(3) 438–450

DOI: 10.1177/09622802211060518
journals.sagepub.com/home/smm
SAGE

Combining cox regressions across a heterogeneous distributed research network facing small and zero counts

Martijn J. Schuemie[1,2,3], Yong Chen[4], David Madigan[1,5], and Marc A. Suchard[1,3,6]

Abstract

Studies of the effects of medical interventions increasingly take place in distributed research settings using data from multiple clinical data sources including electronic health records and administrative claims. In such settings, privacy concerns typically prohibit sharing of individual patient data, and instead, cross-network analyses can only utilize summary statistics from the individual databases such as hazard ratios and standard errors. In the specific but very common context of the Cox proportional hazards model, we show that combining such per site summary statistics into a single network-wide estimate using standard meta-analysis methods leads to substantial bias when outcome counts are small. This bias derives primarily from the normal approximations of the per site likelihood that the methods utilized. Here we propose and evaluate methods that eschew normal approximations in favor of three more flexible approximations: a skew-normal, a one-dimensional grid, and a custom parametric function that mimics the behavior of the Cox likelihood function. In extensive simulation studies, we demonstrate how these approximations impact bias in the context of both fixed-effects and (Bayesian) random-effects models. We then apply these approaches to three real-world studies of the comparative safety of antidepressants, each using data from four observational health care databases.

Keywords

proportional hazards, meta-analysis, privacy preservation, Bayesian, distributed research networks

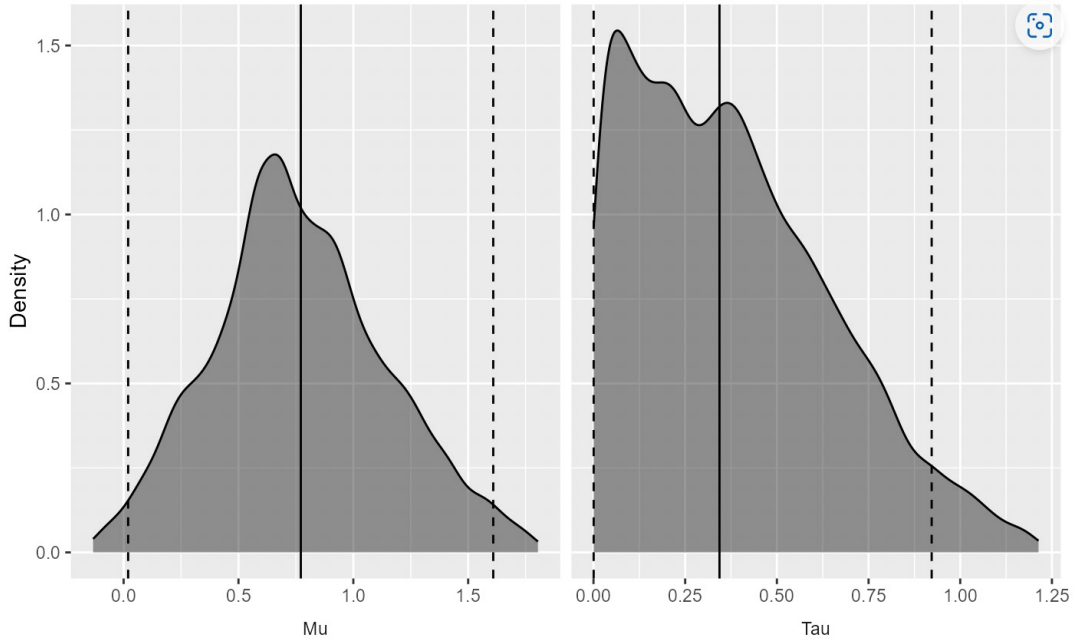
Bayesian random-effects

- Compute a Bayesian meta-analysis using the MCMC engine BEAST.

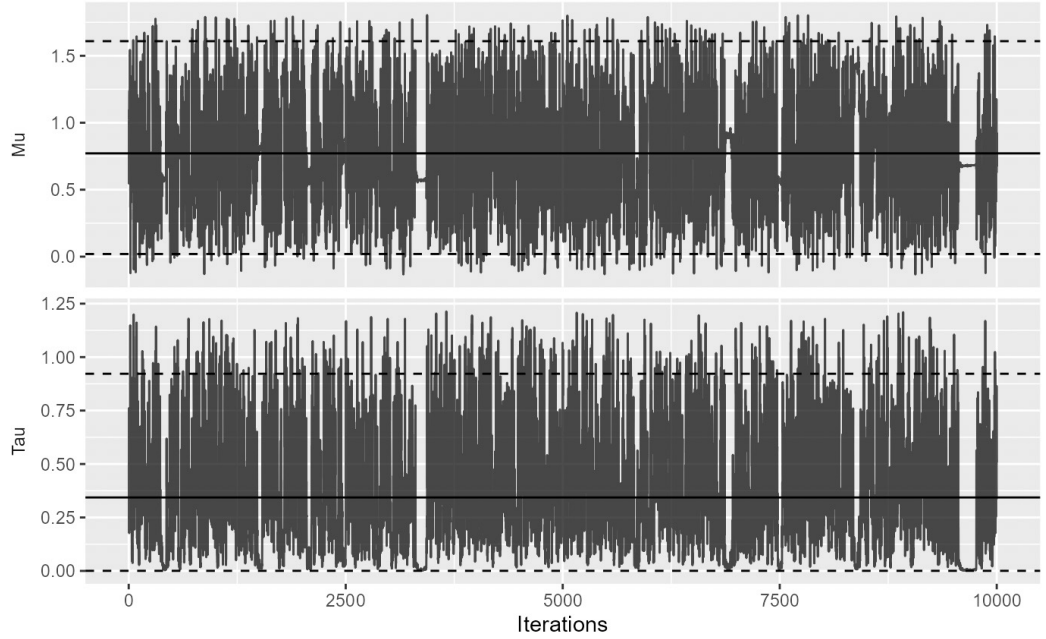
$$\beta_i \sim \text{Normal}(\mu, \tau^2)$$

- A normal and half-normal prior are used for the μ and τ parameters, respectively

Full posterior distribution for μ and τ

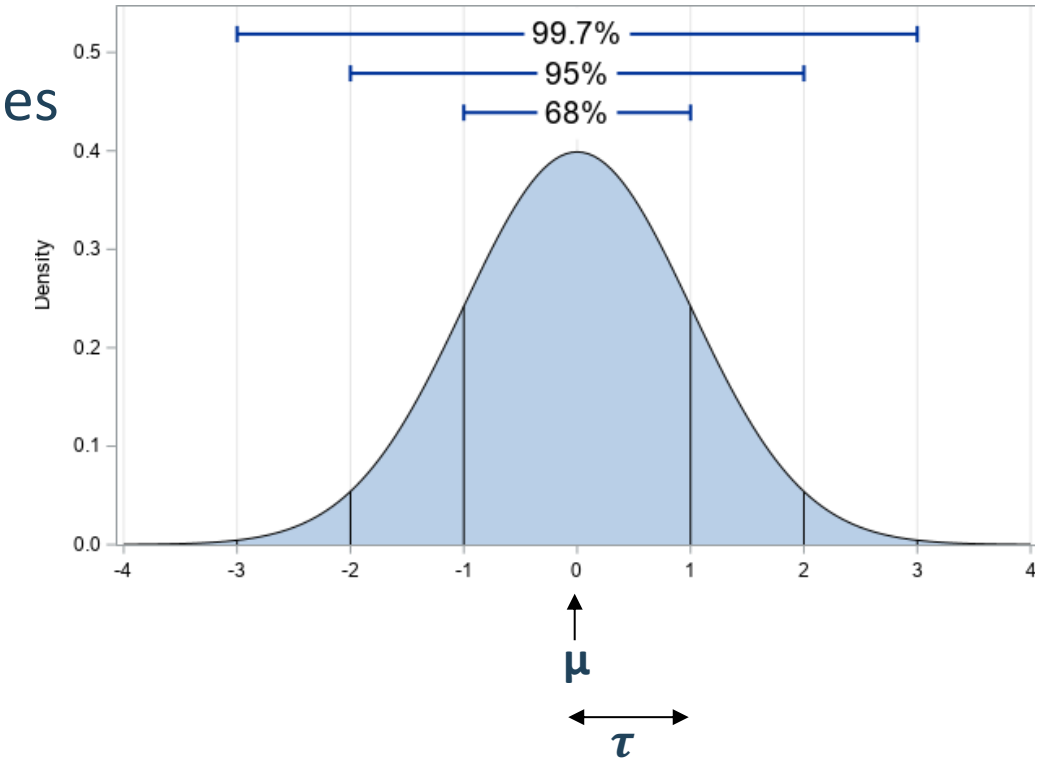


Trace of the MCMC



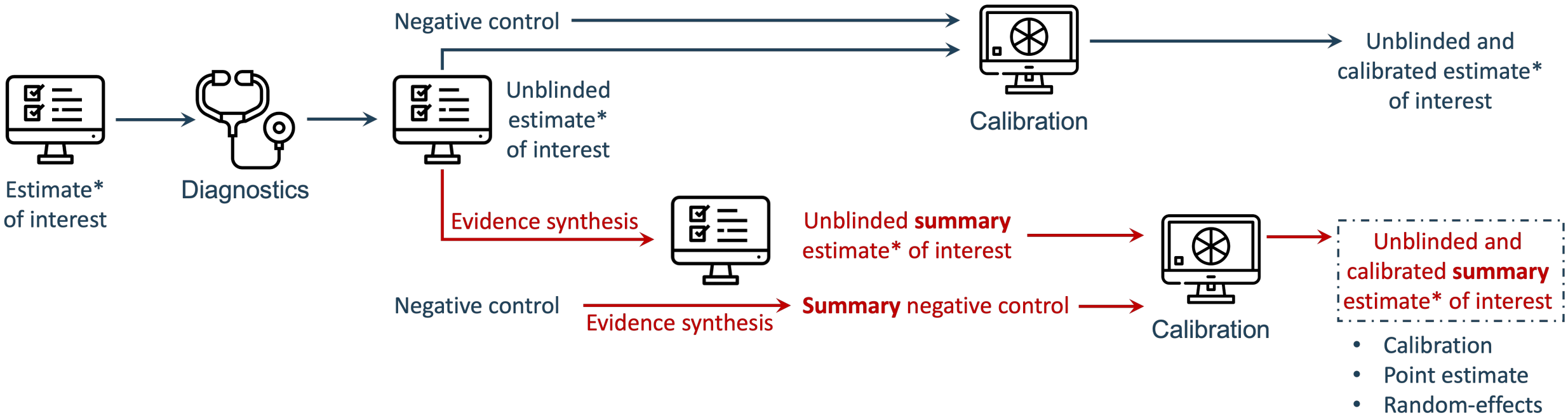
Bayesian random-effects

- Random effects assumes each site draws effect size from a normal distribution with mean μ and standard deviation τ
- τ is ill-defined when we have a few small databases
- What assumption do we want to make?
  - If we don't know, τ must be 0 (no heterogeneity)
  - If we don't know, τ might be 0, or might be >= 0
- We can model this using a Bayesian approach
  - Default: half-normal prior with scale = 0.5
  - (Use fancy MCMC engine to compute: BEAST)



BEAST
Bayesian Evolutionary Analysis Sampling Trees

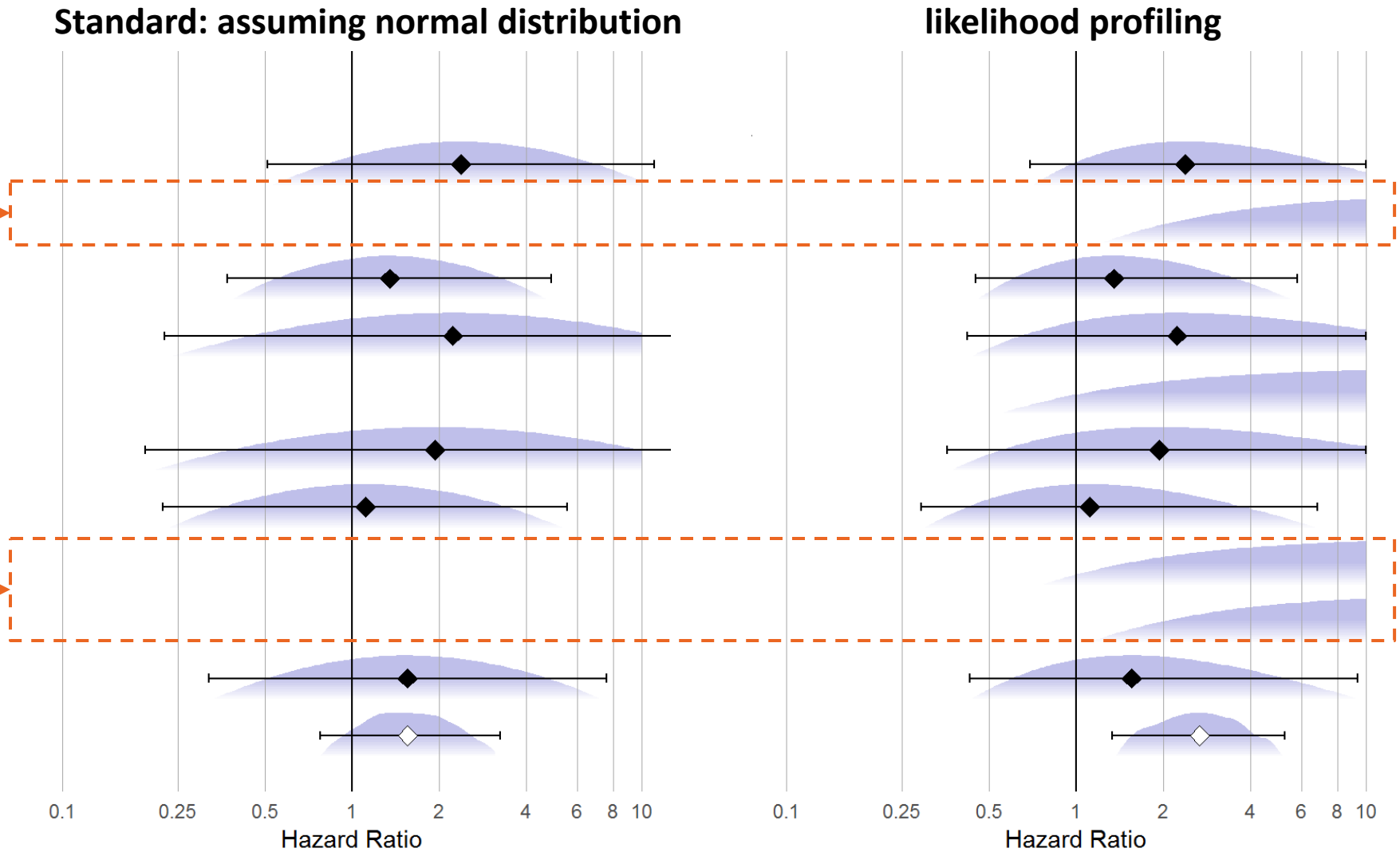
From per-database to summary estimate



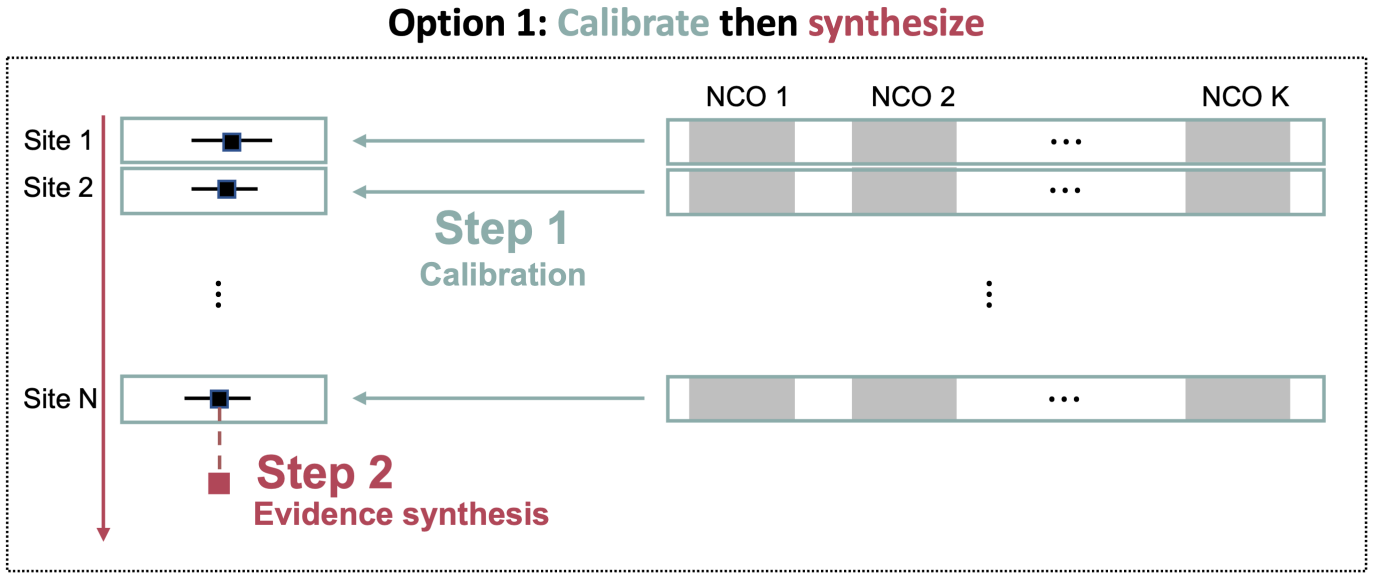
* Estimate here means likelihood profile, which may not be expressed as an estimate

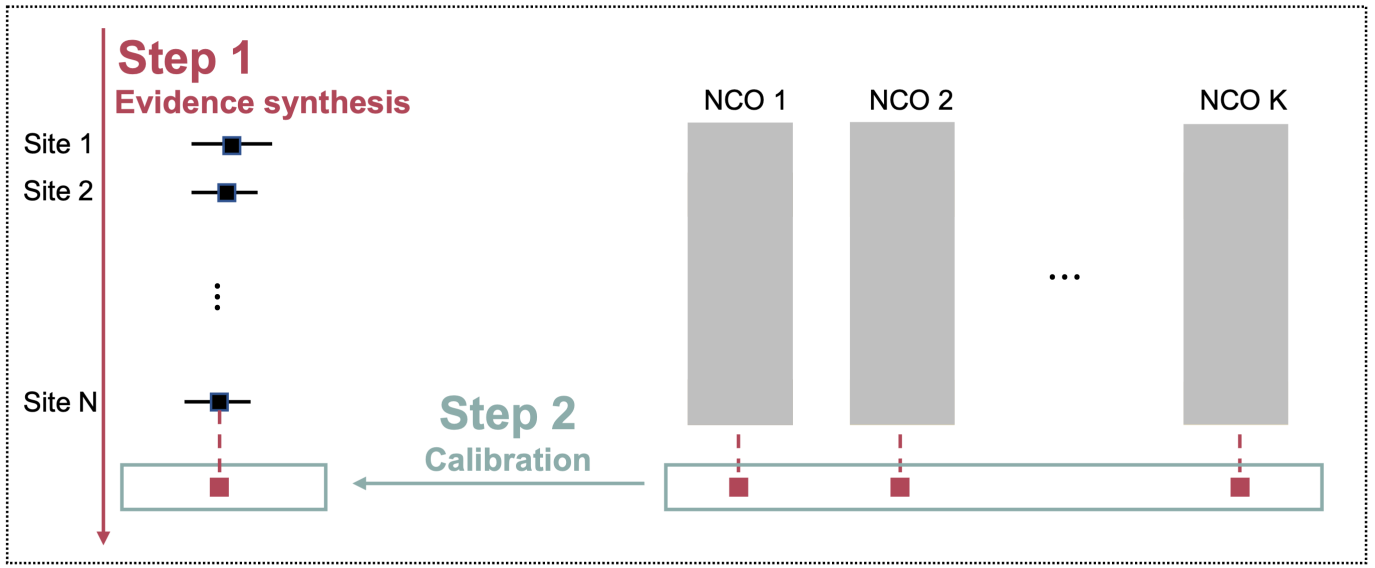
Likelihood profiling

Some sites have zero outcomes in T or C, so no estimate



From per-database estimate to summary estimate





* Estimate here means likelihood profile, which may not be expressed as an estimate