Meta-Analysis with Untrusted Data

Shiva Kaul SKKAUL@CS.CMU.EDU
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213

Geoffrey J. Gordon GGORDON@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Meta-analyses are usually conducted on small amounts of "trusted" data, ideally from randomized, controlled trials. Excluding untrusted (observational) data — such as medical records and related scientific literature — avoids potential confounding and ensures unbiased conclusions. Unfortunately, this exclusion can reduce predictive accuracy to the point of clinical irrelevance, especially when trials are heterogeneous. This paper shows how untrusted data can be safely incorporated into meta-analysis, improving predictions without sacrificing rigor or introducing unproven assumptions. Our approach, called conformal meta-analysis, consists of (1) learning a (potentially flawed) prior distribution from the untrusted data, (2) using the prior and trusted data to derive a simple, fully-conformal prediction interval for the observed trial effect, and (3) analytically extracting an interval for the true (unobserved) effect. In multiple experiments on healthcare datasets, our algorithms deliver tighter, sounder intervals than traditional ones. This paper conceptually realigns meta-analysis as a foundation for evidence-based medicine, embracing heterogeneity and untrusted data for more nuanced, precise predictions.

Keywords: meta-analysis, conformal prediction, noise-tolerant learning, ridge regression

Data and Code Availability This paper uses publicly-available datasets from the Penn Machine Learning Benchmark (Olson et al., 2017). It also collects a small, novel clinical trial dataset to facilitate a case study. This dataset, along with code replicating the experiments, is available at https://github.com/shivak/conformal-meta.

Institutional Review Board (IRB) This paper focuses on the statistical methodology of meta-analysis. As such, it does not require ethical review.

1. Introduction

Meta-analysis is the statistical bedrock of evidence-based medicine, underpinning most modern clinical practice guidelines (Higgins et al., 2019; Hoffmann et al., 2021). As depicted in Figure 1, meta-analysis can be thought of as a structured analogue to language models for answering scientific questions in a rigorous, unbiased manner. In meta-analysis, the training data are n randomized, controlled trials. In each trial, there is an observed effect $Y_i \in \mathbb{R}$, which is the average difference in outcomes between the patients assigned the treatment and those assigned the control. Due to the limited number of patients, the observed Y_i is a noisy version of the true, unobserved effect $U_i \in \mathbb{R}$. We model this as $Y_i \sim N(U_i, V_i)$, where $V_i > 0$ is an observed variance (larger trials tend to have smaller V_i). This model is reasonable because Y_i is an unbiased sample average over individual patients and is thus subject to the central limit theorem. Each trial also has features X_i which describe aspects of how the trial was conducted; for example, the mean age of the patients, the dosage of the drug, and the duration of the treatment. The goal is to learn a predictor C of causal effect: given the features of a future trial x whose true effect is u, we want the prediction interval $C(x) \subseteq \mathbb{R}$ to contain u with high probability.

It is somewhat unusual to describe meta-analysis as a prediction problem involving features. Meta-analysis is more commonly viewed as a basic inferential problem, where the task is to produce a confidence interval for the average treatment effect (ATE, the true mean of the U_i and u) by ignoring the X_i and appropriately averaging the Y_i. However, the global average is less clinically informative than the effects for specific circumstances described by the features (Simonsohn et al., 2022; Subramanian et al., 2018; Gould, 2010; Feinstein, 1995). Even when ignoring features, reporting prediction intervals is strongly encouraged to

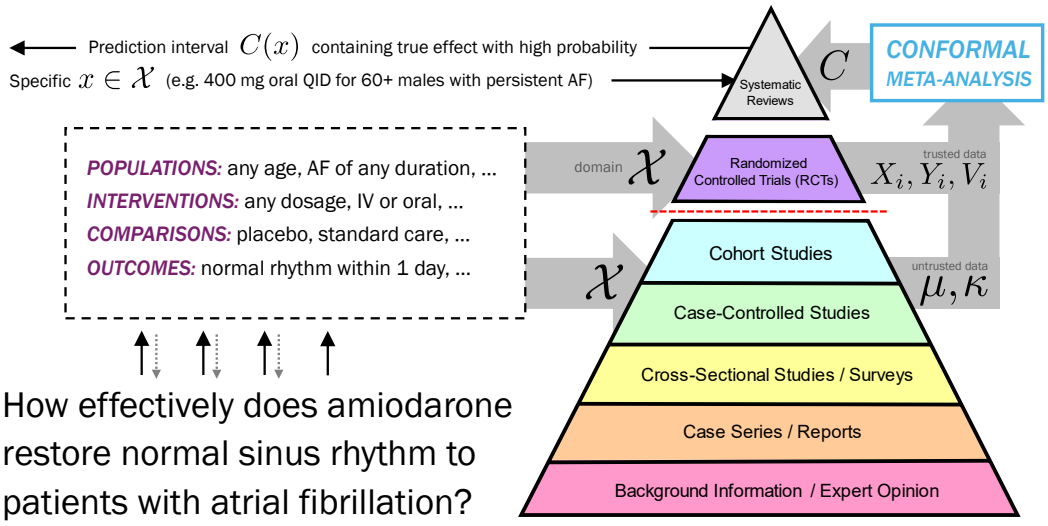


Figure 1: How conformal meta-analysis answers scientific questions. First, a relatively broad domain $\mathcal{X}$ for the meta-analysis is determined, possibly through interaction with the user. This allows more expansive questions which include more data. Next, both trusted and untrusted data relevant to $\mathcal{X}$ are retrieved. Conformal meta-analysis takes these and produces not just a single interval, but a predictive model C. Given specific trial circumstances x, the model predicts $C(x)$ which, under standard assumptions, contains the true effect with high probability.

convey the variation of u around ATE (IntHout et al., 2016; Riley et al., 2011; Borenstein, 2024).

There is an obvious, albeit unfortunate reason why prevalent meta-analysis algorithms ignore features: they involve just a tiny number of trials, far too small to train a predictor of u given x. This number n is usually around 10 or 20 (Hoffmann et al., 2021), and reaches only about 500 on the upper end (Cipriani et al., 2018). The reason for this shortfall: randomized, controlled trials are trusted to support causal conclusions, but they are expensive and rare. Untrusted, observational data — including medical records, insurance claims, related scientific literature, and personal experience — are far more voluminous, but can be badly confounded. Extracting causal conclusions from them requires unproven, possibly dubious assumptions. These are unacceptable in meta-analysis, which is supposed to authoritatively validate the conclusions of earlier research in their entirety.

Since prevalent algorithms ignore x, their (monolithic) prediction intervals $C(x)$ must be wide enough to accommodate nearly all x. Among meta-analyses whose confidence intervals exclude the null effect, about one-third have prediction intervals which include it (IntHout et al., 2016). This imprecision makes it difficult to establish good scientific evidence in fields with heterogeneous trials, i.e. when substantial variation in x leads to concomitant variation in u. For example, in exercise science, the treatment effect may strongly depend on a large number of variables such as frequency, duration, equipment, technique, age, and diet (Rippetoe, 2017; Ferreira et al., 2010). In most psychological research, the variation in u is attributable primarily to between-study heterogeneity rather than within-study sampling variance (Stanley et al., 2018). Because prediction intervals tend to be so wide, few meta-analyses bother to even report them (Seehra et al., 2021; Borg et al., 2024), despite their clinical importance.

1.1. Our Approach

This paper demonstrates that untrusted data — with all its possible confounding, biases, and even outright errors — can be incorporated into meta-analysis while remaining rigorous and unbiased. In fact, this paper offers stronger, provable guarantees while weakening the assumptions traditionally employed in meta-analysis. The solution is based upon *conformal prediction* (Vovk et al., 2005, 2009; Lei et al., 2018; Angelopoulos et al., 2023b). While conformal prediction

aptly manages the inclusion of untrusted data, there are two unresolved challenges when applying it to meta-analysis. The first challenge is noise: though we aim to predict true effects u, the observed effects $Y_i \sim N(U_i, V_i)$ are blurred by limited trial sizes. This noise is curiously challenging to manage, since small (high noise) studies can have fundamentally different designs than large (low noise) studies. This reflects difficulties in clinical practice, where large-scale trials routinely fail to confirm the results of smaller ones (Ioannidis, 2005; Komajda et al., 2010; Manson et al., 2019). The second challenge arises from the limited sample ($n \leq 500$) of included trials. This essentially mandates the use of full (rather than split) conformal prediction, which poses a computational burden, and complicates efforts to handle noise.

We resolve the aforementioned challenges of applying conformal prediction, giving rise to *conformal meta-analysis*. This approach consists of the following layers: (1) representation of the untrusted data as a Gaussian process prior (i.e. a mean function $\mu(x)$ predicting u and a kernel function $\kappa(x, x')$ quantifying similarity and uncertainty), (2) a simple implementation of full conformal prediction of y, based on residuals produced by kernel ridge regression (KRR), and (3) a strategy for predicting u, exploiting the simplicity of the conformal intervals for y. To obtain (2), we show that sufficiently high regularization makes KRR *idiocentric*: as y varies, the residual for the example (x, y) changes more than the other residuals. Under this condition, fully-conformal KRR can be simplified to computing quantiles in two lists. Its simplicity allows us to prove that its prediction intervals for y typically contain the true effects u as well, with just a slight loss in confidence — thereby achieving (3).

Our experiments have two goals: (1) to quantify how much conformal meta-analysis could improve predictions when used, as intended, with large amounts of untrusted data, and (2) to more qualitatively assess, before such data are available, how it would impact the experience of producing and consuming meta-analyses. At a high level, we find that conformal meta-analysis could improve how the medical community interacts with evidence.

2. Preliminaries

These are the predictive goals of meta-analysis, as informally described in the previous section. Readers unfamiliar with meta-analysis are encouraged to review the background material in Appendix A.

Task 1 (Predicting Effects) *Let $(X_1, U_1, V_1), \ldots, (X_n, U_n, V_n), (x, u, v)$ be exchangeable random variables, where $X_i, x \in \mathcal{X}$ are trial features, $U_i, u \in \mathbb{R}$ are (unobserved) effects, and $V_i, v > 0$ are variances. Let $Y_i = U_i + \mathcal{E}_i$, where independently $\mathcal{E}_i | V_i \sim N(0, V_i)$. Let $\mu : \mathcal{X} \to \mathbb{R}$ and $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be fixed mean and positive-definite kernel functions defining a Gaussian process. From (μ, κ), the (X_i, Y_i, V_i), and x, for a desired confidence level $\alpha \in (0, 1)$, produce an interval $C(x)$ such that $\mathbb{P}(u \in C(x)) \geq 1 - \alpha$, where the probability is over all the random variables.*

Task 2 (Predicting Trials) *Same as above, except C also takes v, and should satisfy $\mathbb{P}(y \in C(x, v)) \geq 1 - \alpha$, where $y = u + \epsilon$ for independent $\epsilon | v \sim N(0, v)$.*

The first task is more practically useful and technically involved. However, since u is not observable, but y is, the second task is more easily verifiable. It is not immediately clear which task is more challenging, in the sense of needing wider intervals. On one hand, y has inherently more variance than u. On the other, the prediction of u is made without knowing v, which might otherwise distinguish between small and large trials having characteristically different u.

Note about assumptions. The random-effects model of meta-analysis (DerSimonian and Laird, 1986; Higgins et al., 2009) underlies the majority of published meta-analyses. Both Task 1 and Task 2 make the same assumptions as this standard model, with two exceptions. The standard model assumes that the true effects $U_1, \ldots, U_n, u$ are normally distributed around the average effect ATE. We make no such normality assumption; in this sense, our approach is more robust than the usual one. Our only non-standard assumption is that the untrusted data are fixed relative to the trusted data, i.e. the randomized, controlled trial results are not copied back into the prior. This assumption can be mechanically enforced by explicitly excluding the trials from the prior. See Appendix A.4 and Appendix A.5 for further discussions about our (relatively lightweight) assumptions.

2.1. Related Work (Continued in Appendix)

Causal inference from observational data. Performing randomized, controlled trials is not the only way to estimate causal effects. After making appropriate assumptions, causal inferences can be extracted from observational data (Imbens and Rubin, 2015; Pearl, 2009; Spirtes et al., 2001). This is an extensive

research endeavor encompassing many fields; we mention some of the most relevant work here. The survey by Colnet et al. (2024) discusses various approaches to integrating RCTs with observational data. To estimate the CATE, causal forests (Wager and Athey, 2018) and metalearners (Künzel et al., 2019) combine machine learning techniques with causal reasoning. The most widespread assumption of such methods is ignorability, or unconfoundedness. It requires that, having observed the features x, the treatment assigned to a participant is independent of their potential outcomes $\rho(0)$ and $\rho(1)$. That is, there are no unmeasured variables outside of x that could bias treatment towards different participants. Another widespread assumption is positivity, or overlap: for every x, both the treatment and the comparison have a chance of being assigned.

Such strong, unproven assumptions are plausible in many circumstances, but they are not appropriate for systematic reviews. At some point, assumptions must be tested; systematic reviews, more confirmatory than exploratory in nature, often serve this crucial purpose. Nevertheless, conformal meta-analysis allows causal inference methods to be (indirectly) used in systematic reviews, without any concerns about their unproven assumptions. These methods can ideally be used to extract better μ and κ from the untrusted data. Thus, conformal meta-analysis doesn't replace these methods; rather, it expands their domain of application to more scientific settings.

Conformal prediction of latent variables. Previous works have examined how to conformally predict an underlying u while observing only noisy $Y_1, \ldots, Y_n$. It is often empirically observed that conformal prediction can be obliviously robust to label noise, in the sense that $C(x)$, without any involvement of V or v, manages to covers u without any loss in confidence. However, provable guarantees remain elusive. Feldman et al. (2023) show that if $C(x)$ always contains the median of $u \mid x$, then $C(x)$ covers u with no loss in confidence. This is a very strong assumption in meta-analysis, as it essentially posits that the relationship between x and u has been globally determined, and the main difficulty of conformal prediction is to account for the uncertainty driven by the unobserved variables ξ. Most approaches to (non-obliviously) handling noise involve some modification to split conformal prediction. In classification, the (discrete) labels may be noisy because they are the majority vote from some underlying probability distribution, which reflects uncertainty over the true class. Stutz et al. (2023) adapt split conformal prediction to account for this uncertainty by sampling multiple labels from the underlying distribution. Sesia et al. (2023) and Penso and Goldberger (2024) modify split conformal prediction to estimate the amount of over (or under) coverage of $C(x)$. Unfortunately, splitting the data is not feasible in meta-analysis, where n is small. Label noise should be distinguished from label shift, when the training $Y_1, \ldots, Y_n$ are sampled from a different distribution than the test y (Podkopaev and Ramdas, 2021).

3. Conformal Meta-Analysis

This section presents the three steps of conformal meta-analysis. Section 3.1 describes how a prior (μ, κ) can be learned from untrusted data. Our focus is not on strategies for learning such priors, but on understanding the potential benefits of incorporating them into meta-analysis. Section 3 applies conformal prediction in conjunction with KRR. This yields a set $C(x, v)$ which contains y with the required probability, regardless of the veracity of (μ, κ). Under a novel condition called idiocentricity, we show $C(x, v)$ becomes a simple interval for all linear smoothing methods. (We chose KRR primarily because it can incorporate μ and κ). Finally, Section 3.2 describes how a prediction interval $C(x)$ for u can be extracted from $C(x, v)$, with just a small drop in coverage probability. This can be proven (in Theorem 12) because, under idiocentricity, $C(x, v)$'s width as a function of v can be tightly bounded.

3.1. Untrusted Data as a Prior Distribution

As depicted in Figure 1, untrusted data form the lower levels of the evidence hierarchy: observational studies, individually-published cases, hands-on experience, and personal belief (Murad et al., 2016). It is difficult to rigorously infer causation from such untrusted (or "real-world") data, since they are observational and may have deeply-embedded biases. Nonetheless, it is often found that untrusted data agree with RCTs (Benson and Hartz, 2000; Concato et al., 2000). Untrusted data originate from different kinds of sources and experiences; for example, in Section 5, they stem partially from background knowledge, and partially from RCTs that would be excluded from the meta-analysis. A modern approach to capturing large, disparate collections of knowledge is to (pre)train foundation models. Such models are already being developed

```python
def predict_trial(Y, V_, M_, K_, alpha, eta):
  n = len(Y); tau = ceil((1-alpha)*(n+1)).astype(int32)
  if tau > n: return -inf, inf # not enough training trials for conformal prediction
  lam = amax(diag(K_)) # ensure idiocentricity, per Theorem 6
  G, H = theorem4(Y, V_, M_, K_, alpha, eta, lam) # r <= Ri iff y in interval Gi±Hi
  Ln, Lp = G-H, G+H # lower and upper endpoints of the intervals
  yn = flip(sort(Ln))[tau-1] # return quantiles of the lower and upper endpoints
  yp = sort(Lp)[tau-1]
  return yn, yp
```

Algorithm 1: Python/NumPy code for conformal prediction of an (empirical) effect y. Y is a vector of the n training trial effects; $\bar{V} = [V; v]$, $\bar{M}$, and $\bar{K}$ are the effect variances, prior means, and prior kernel matrix for all $n+1$ trials, both training and testing. Algorithm 3 in Appendix B.2 implements the computations involved in Theorem 4.

for healthcare (Moor et al., 2023; Singhal et al., 2023; Tu et al., 2024). When applied to meta-analysis, this approach would involve learning an embedding $\phi(x)$ which maps trial features x into a Euclidean space having inner product $\kappa(x, x') = \phi(x)^T \phi(x')$. On top of this embedding, a linear predictor of u could be trained as $\mu(x) = w^T \phi(x)$. We will call the Gaussian process defined by (μ, κ) a prior, since it is trained separately from the trials included in the meta-analysis. However, unlike in Bayesian meta-analysis, this prior is untrusted; our predictions must retain coverage guarantees even if the prior has severe flaws. See Appendix A.5 for further discussion of untrusted data.

3.2. (Simply) Predicting Trials

This section develops Algorithm 1 for Task 2. The parameter $\eta \geq 0$ controls the extent of noise correction.

Theorem 1 (Conformal Trial Prediction) *Let $\eta \geq 0$. In the setting of Task 2, Algorithm 1 returns $[y_-, y_+]$ satisfying $\mathbb{P}(y \in [y_-, y_+]) \geq 1 - \alpha$.*

This algorithm is derived by calculating KRR's residuals, obtaining the conformal prediction interval $C(x, v)$, and simplifying it under idiocentricity. Given a ridge parameter $\lambda \in \mathbb{R}$, prior (μ, κ), and length-$(n+1)$ dataset $([X; x], [Y; y])$, KRR learns a posterior $(\hat{\mu}, \hat{\kappa})$. Let the posterior mean on $[X; x]$ be $[\widehat{M}; \hat{m}]$. Let the diagonal of the posterior kernel matrix be $[S^2; s^2]$. Let $Z_i = \mathbb{E}_{\mathcal{E},\epsilon}(\widehat{M}_i - Y_i)^2 - (\widehat{M}_i - U_i)^2 \geq 0$ and $z = \mathbb{E}_{\mathcal{E},\epsilon}(\hat{m} - y)^2 - (\hat{m} - u)^2 \geq 0$ be the expected impact of the noise upon the squared training errors. Though other choices are possible, we define the residuals as follows, using the expected impacts $[Z; z]$ to correct for within-trial variation:

$$R_i = \left((\widehat{M}_i - Y_i)^2 - \eta Z_i\right) / S_i^2 \qquad r = \left((\hat{m} - y)^2 - \eta z\right) / s^2$$

Subtracting (an η fraction of) Z_i and z effectively reduces the importance of smaller (noisier) trials. Concretely deriving these residuals for KRR, with normal noise in Y, is basic linear algebra and probability. As Appendix B.2 shows, the residuals are squares of affine functions in y. That is, for some A_i, B_i, a, and b:

$$R_i = \frac{(A_i y + B_i)^2 - \eta Z_i}{S_i^2} \qquad r = \frac{(ay + b)^2 - \eta z}{s^2} \tag{1}$$

Residuals of this form are actually shared by any learning algorithm where $[\widehat{M}; \hat{m}]$ are linear in $[Y; y]$, albeit being nonlinear in $[X; x]$. Such algorithms, called linear smoothers, include k-nearest neighbors, Nadaraya-Watson kernel regression, and smoothing splines (Buja et al., 1989). With these residuals defined, we can apply (full) conformal prediction. Remarkably, this yields a rigorous prediction set $C(x, v)$, regardless of the veracity of the untrusted data, or the complexity of the prior trained upon it.

Proposition 2 (Conformal Prediction) *Let $(X_1, Y_1, V_1), \ldots, (X_n, Y_n, V_n), (x, y^*, v)$ be exchangeable. Let $[R; r]$ be the residuals (1) of a symmetric (i.e. unaffected by the order of its inputs) learning algorithm upon the augmented data $[X; x]$, $[Y; y]$ and $[V; v]$. Given any $\alpha \in (0, 1)$, let $\tau = \lceil (1-\alpha)(n+1) \rceil$. Define the prediction interval as $C(x, v) = \{y : r \text{ is among the } \tau \text{ smallest of } R_1, \ldots, R_n\}$. Then $\mathbb{P}(y^* \in C(x, v)) \geq 1 - \alpha$. (Vovk et al., 2005)*

Burnaev and Nazarov (2016), building upon Nouretdinov et al. (2001), derived an algorithm for computing $C(x, v)$ for KRR. Though their algorithm is computationally efficient, it returns a general prediction set (a union of disjoint intervals and singletons) which isn't amenable to analytic reasoning. We substantially simplify the algorithm under the following condition.

Definition 3 (Idiocentricity) *The residuals in (1) are idiocentric if* $\frac{|a|}{s} > \frac{|A_i|}{S_i}$ *for all* i.

This condition means that changing the test example's y changes its own residual more than it changes the residuals of other examples. (It can be generalized in terms of derivatives, but we keep it specific to linear smoothers for the presentation here). First, let us show how idiocentricity simplifies $C(x, v)$.

Theorem 4 *For* $i = 1, \dots, n$, *let* $\rho_i = \eta(Z_i s^2 - z S_i^2)$. *Define intervals* $L_i = G_i \pm H_i$, *where:*

$$G_i = \frac{A_i B_i s^2 - abS_i^2}{(aS_i)^2 - (A_i s)^2} \text{ and}$$

$$H_i = \frac{\sqrt{\max\left(0,\ s^2 S_i^2 (A_i b - a B_i)^2 - \rho_i((aS_i)^2 - (A_i s)^2)\right)}}{(aS_i)^2 - (A_i s)^2}$$

With idiocentric residuals (1), $C(x, v)$ *simplifies to* $\{y : y$ *is inside more than* $n - \tau$ *of the* $L_1, \dots, L_n\}$.

Proof Since the residuals defined in (1) are squared, we can flip the signs of b and B_i to standardize on $a, A_i \geq 0$. $r \leq R_i$ rewrites to $S_i^2(ay + b)^2 + \rho_i \leq s^2(A_i y + B_i)^2$. Under the condition $a/s > A_i/S_i \geq 0$, this is equivalent to $y \in L_i$. ■

We slightly loosen the defining condition of $C(x, v)$ to obtain an even simpler algorithm.

Lemma 5 *In the notation of Theorem 4, let* y_+ *be above* τ *of the upper endpoints of the* L_i, *and let* y_- *be below* τ *of the lower endpoints of the* L_i. *Then* $C(x, v) \subseteq [y_-, y_+]$.

Proof The upper endpoint y_+ is met when, for τ of the $i \in \{1, \dots, n\}$, we have $y_+ \leq L_i$ or $y_+ \geq L_i$. Ignore the first possibility, which becomes more unlikely as y_+ increases, for a potentially looser but nonetheless valid interval. A similar argument justifies y_-. ■

Next, we show that KRR is idiocentric when the ridge parameter λ is set sufficiently large.

Theorem 6 *The residuals in (1) are idiocentric if* $\lambda \geq \max\{\kappa(X_1, X_1), \dots, \kappa(X_n, X_n), \kappa(x, x)\}$.

To prove Theorem 1, use the λ of Theorem 6 to earn the simplified interval of Theorem 4, which is supported by the coverage guarantee of Theorem 2.

Alternative Choices of λ. Throughout the remainder of this paper, λ is chosen to satisfy Theorem 6. (The satisfying value is hardcoded in line 4 of Algorithm 1). However, the bound given by Theorem 6 can be loose. With the training and test examples held fixed, for any given value of λ, idiocentricity can be numerically verified by simply computing the terms in the definition. Since these terms don't depend on the test example's y, a grid search can be used to find the smallest λ which ensures idiocentricity, without interfering with the ensuing conformal prediction.

A smaller value of λ may potentially be statistically desirable, since it controls the balance between fitting the data and staying close to the prior. Note the equation (7) for the posterior mean $\widehat{M}$: at $\lambda = 0$, KRR interpolates the data, and as $\lambda \to \infty$, KRR ignores the data and sticks with the prior. Smaller λ could potentially be preferable when n is relatively large, justifying a posterior farther from the prior. However, the optimal setting of λ for regression may not coincide with the optimal setting for conformal prediction. For example, $\lambda = 0$ can be effective for "ridgeless" regression (Hastie et al., 2022; Liang and Rakhlin, 2020), but it is useless for full conformal prediction, since its residuals are all zero.

3.3. Predicting Effects

The culmination of this paper is Algorithm 2, for predicting causal effects, which builds upon Algorithm 1. The proof of its coverage guarantee is technical, so it is reserved for Appendix B.4. Here, let us simply understand its statement in the following theorem.

Theorem 7 (Conformal Effect Prediction) *Let* $\eta \geq 0$. *In the setting of Task 1, Algorithm 2 returns* $[u_-, u_+]$ *satisfying* $\mathbb{P}(u \in [u_-, u_+]) \geq 1 - \frac{\alpha}{(1-\alpha)\text{erfc}\sqrt{\eta/2}}$.

Setting $\eta = 0$ (i.e. disabling noise correction) obtains confidence $\frac{1-2\alpha}{1-\alpha}$, which is just a slight loss from $1 - \alpha$ when $\alpha \approx 0$. (For example, 0.95 confidence drops to 0.9473, which probably doesn't change $\tau = \lceil (1-\alpha)(n+1) \rceil$). This setting is appropriate when $V \approx 0$, i.e. the trials all have a large number of participants. By setting $\eta = 2 \cdot \text{inverfc}(\frac{1}{c(1-\alpha)})^2$, the confidence drops to $1 - c \cdot \alpha$. More noise correction is conceptually more appropriate when analyzing mixtures of small and large trials. However, the loss of confidence means larger n is needed, which may not be a worthwhile tradeoff. Conformal prediction is usable only when $\tau \leq n$; with $c = 2$, a final confidence of 95% requires $n \geq 40$, which is twice the n needed for $\eta = 0$.

While the overhead at $\eta = 0$ is not practically important, it indicates either the algorithm or its analysis are suboptimal. When meta-analysis is very close

```
def predict_effect(Y, V, M_, K_, alpha, eta):
  V_ = append(V, 0) # as if the test trial had zero effect variance
  return predict_trial(Y,V_,M_,K_,alpha,eta)
```

Algorithm 2: Python/NumPy code for conformal prediction of an effect u. It submits $v = 0$ to Algorithm 1.

to regression ($V \approx 0$), the $1-\alpha$ coverage of Theorem 2 should be smoothly recovered. Appendix B.5 present another approach which behaves correctly in this regard. It is based on fundamentally different techniques which can be extended to non-normal (i.e. merely bounded) noise. It determines a probability $1-\delta$ region $\mathcal{U}$ for the true effects U. Then, it poses an optimization problem to bound all the intervals which could have been generated by any $\widehat{U} \in \mathcal{U}$.

Theorem 8 (Conformal Effect Prediction) *Let $\delta > 0$. In the setting of Task 1, the respective solutions u_- and u_+ to (9) and (10) in Appendix B.5 satisfy $\mathbb{P}(u \in [u_-, u_+]) \geq (1-\alpha)(1-\delta)$.*

4. Experiments

This paper is the first to study meta-analysis with the involvement of untrusted priors and features. There are no existing meta-analytic datasets upon which we can quantitatively evaluate our methods, and generating such a dataset is a significant undertaking. Thus, our main experiments use partially-synthetic data. We performed four types of simulations on three biomedical datasets from the Penn Machine Learning Benchmark (Olson et al., 2017). These regression datasets define K and Y; we generated synthetic M and V according to parameters prior error ≥ 0 and effect noise ≥ 0, respectively. We use Algorithm 2 with $\eta = 0$. We compare it to the state-of-the-art HKSJ method, which is described in Appendix A.6.

Simulation 1: This investigates when conformal meta-analysis is superior to traditional meta-analysis. For different settings of prior error, we compare the widths of the intervals obtained by different meta-analysis algorithms. The only situation in which HKSJ is competitive with conformal meta-analysis is when the prior is bad and the number of trials is small/moderate. Otherwise, conformal meta-analysis is superior, sometimes achieving intervals that are dramatically thinner than those of HKSJ.

Simulation 2: This experiment checks whether the desired 95% confidence level is still achieved as effect noise increases. Conformal meta-analysis succeeds, whereas HKSJ fails badly. On the other datasets (see appendix), HKSJ sometimes drops below 80% confidence. This deficiency is present at all settings of effect noise, though it aggravates at higher values. This simulation shows that conformal meta-analysis has a rigorous coverage guarantee, and HKSJ does not. It should be noted that HKSJ was developed to improve the coverage guarantee of the more prevalent Higgins-Thompson-Spiegelhalter method.

Simulation 3: This experiment compares different instantiations of Algorithm 2: one with $\eta = 0$, and the other with $\eta = 0.4015$, with α adjusted so both ultimately seek a 90% confidence level. With the higher setting of η, over-coverage is consistently demonstrated. This suggests that the analysis of Section 3.2 can be improved, at least in some settings.

Simulation 4: Our approach assumes that, in many fields, it should be possible to develop good priors from large volumes of untrusted data. However, if these priors are indeed very accurate, it is unclear whether using KRR (upon just n trials) is worth the complexity, and possible statistical overhead, over just using the prior as a fixed predictor. (This is conceptually equivalent to using a very large ridge parameter λ). This simulation indicates there is no such overhead: our fully-conformal intervals are strictly superior to those derived from a fixed prior. Thus, unless assumptions stronger than exchangeability are used to derive prediction intervals, learning is superior to mere validation.

5. Case Study: Amiodarone

Evaluating our approach on real meta-analytic data is valuable, even if at limited scale due to the practical difficulties of collecting a large dataset. We revisit the systematic review of Letelier et al. (2003), which assessed the effectiveness of amiodarone for atrial fibrillation (AF) patients. Its outcome measure is the relative risk of normal sinus rhythm; that is, the probability of restoring normal rhythm when administered amiodarone, divided by the probability of restoration with placebo. The review involved $n = 21$ trials, which we use as training data. For test

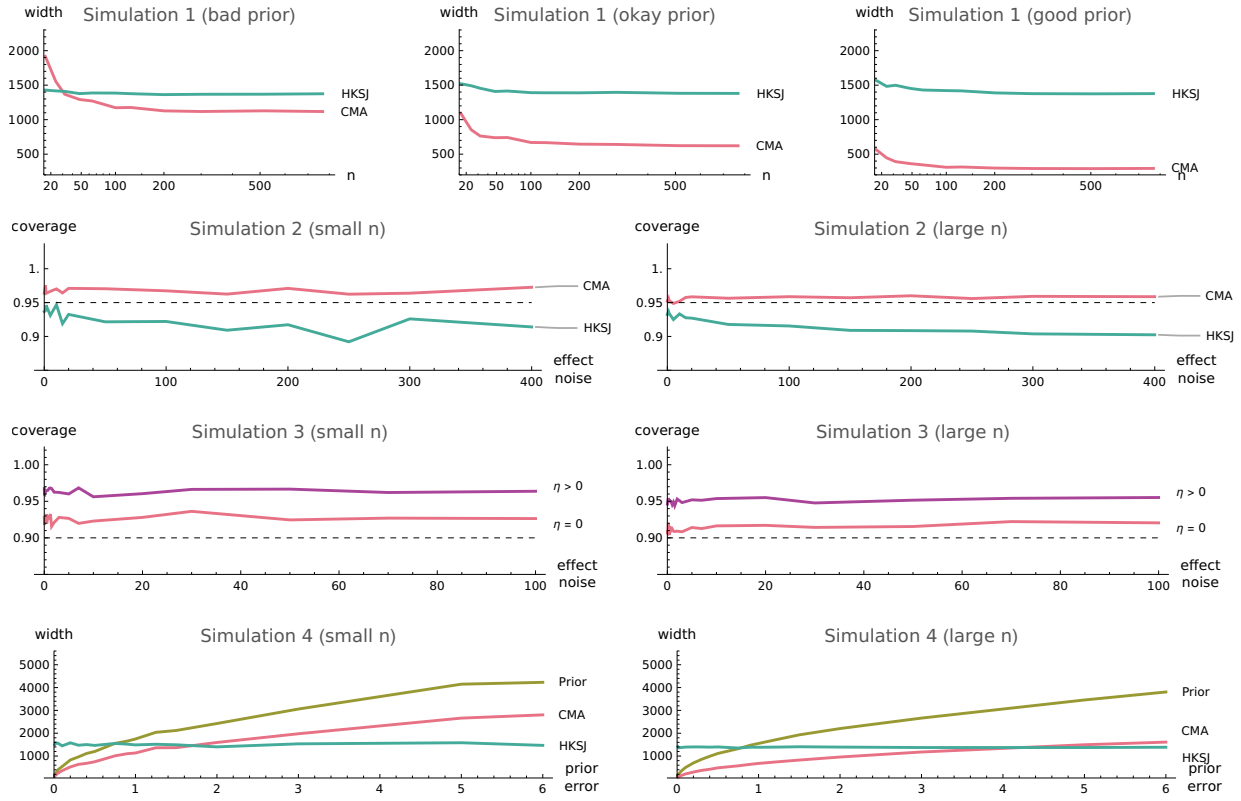


Figure 2: Results of all simulations on a single exemplar dataset. See Appendix B.6 for congruent results on the other datasets. The parameter "prior error" controls the distance between the true effects Y and the prior predictions M; zero prior error means $M = Y$. The parameter "effect noise" controls the size of the noise variances V; zero effect noise means $V = 0$. Their mathematical definitions are also presented in Appendix B.6. Overall, conformal meta-analysis can deliver much tighter intervals than traditional methods (Simulation 1), even though traditional methods have weak coverage guarantees (Simulation 2), whereas our algorithms, or their analyses, have (overly) strong guarantees (Simulation 3). Our algorithms, not just good priors, are essential to this performance (Simulation 4).

data, we identify 4 trials that were published after the review, but would have met its inclusion criteria (Thomas et al., 2004; Kochiadakis et al., 2007; Balla et al., 2011; Karaçağlar et al., 2019). Per Task 2, we compare traditional meta-analysis (the Bayesian algorithm of Theorem 11, described in Appendix A.6) with conformal meta-analysis (Algorithm 1, with $\eta = 1$).

Our goal is not to make scientific claims about amiodarone; that would require following a formal, preregistered protocol. Though we temper our quantitative findings, depicted in Figure 3, we find them qualitatively interesting. Conformal meta-analysis manages to correctly predict all 4 trials, whereas traditional meta-analysis suffers a misprediction. This is not statistically convincing, but it aligns with the fact that conformal meta-analysis has a rigorous coverage guarantee, whereas traditional algorithms do not. (See Appendix A.6 for more details). Not all of the conformal intervals overlap, but traditional intervals all inherently overlap. This suggests conformal meta-analysis can make predictions that are meaningfully responsive to the details of trials, perhaps distinguishing between effective and ineffective ones.

Appendix B.7 describes how we conducted the case study, which differed from a typical meta-analysis. The first change is training a prior on helpful data that would otherwise be ignored. We identify 8 trials that did not meet the inclusion criteria, since they were not placebo controlled. To generate pseudo-effects for these trials, we must estimate the placebo response, and then subtract it from the amiodarone response. This leads to the second major change, which is holistically including the perspectives of practitioners. The critique of Slavik and Zed (2004), written by two

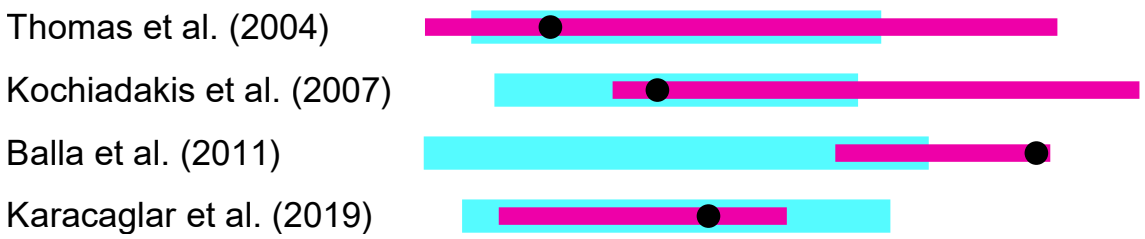


Figure 3: Prediction intervals for new observed effects y (black dots) produced by traditional meta-analysis (light blue) and conformal meta-analysis (magenta, thin). On average, they are comparable in width, but conformal meta-analysis manages to cover the discrepant trial of Balla et al. (2011). Thus, the conformal intervals are more responsive and do not sacrifice precision. The prior for conformal meta-analysis was produced **post-hoc**, having already seen the results of Letelier et al. (2003). Thus, these intervals are not quantitative evidence, they are merely qualitative illustrations of the behavior of conformal meta-analysis. Axes are intentionally omitted to deemphasize specific quantitative results.

doctors of pharmacy, gave estimates for the placebo response on sinus rhythm (i.e. spontaneous conversion) in different circumstances. We use these estimates to generate the pseudo-effects. Finally, we use LLMs (specifically, GPT-4 and Claude 3.5 Sonnet) to extract trial data, performing in-context parsing upon published trial documents. In this manner, LLMs can be used to aid meta-analysis, much as meta-analysis serves as a question-answering system. This experience, and the paper overall, reflect positively on the following dilemma: *can language models be used to rigorously answer scientific questions?*

6. Conclusion

In their seminal paper on random-effects meta-analysis, DerSimonian and Laird (1986) expressed hope for resolving heterogeneity by using features x. 35 years later, Bryan et al. (2021) declared that such a "heterogeneity revolution" had still not occurred. Conformal meta-analysis could help spark this revolution, but much further research is warranted. Both our algorithms and their analyses could be quantitatively improved. It should be possible to further relax our statistical assumptions, such as exchangeability (Barber et al., 2023; Gibbs and Candès, 2024). Collecting larger meta-analysis datasets, and training priors on larger volumes of untrusted data, would help further validate our approach.

Acknowledgements

We thank Ryan Tibshirani, Erik van Zwet, Maya Mathur, Robert Tibshirani, Lu Tian, and Michael Oberst for their helpful feedback and discussion.

References

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671): 669–674, 2023a.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4): 494–591, 2023b.

David Armstrong. Professionalism, indeterminacy and the ebm project. *BioSocieties*, 2(1):73–84, 2007.

W. L. Baker, C. Michael White, J. C. Cappelleri, J. Kluger, C. I. Coleman, and From the Health Outcomes, Policy, and Economics (HOPE) Collaborative Group . Understanding heterogeneity in meta-analysis: the role of meta-regression. *International Journal of Clinical Practice*, 63(10):1426–1434, 2009. doi: https://doi.org/10.1111/j.1742-1241.2009.02168.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-1241.2009.02168.x.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

Idriz Balla, Elizana Petrela, and Anesti Kondili. Pharmacological conversion of recent atrial fibrillation: a randomized, placebo-controlled study of three antiarrhythmic drugs/yeni baslayan atriyal fibrilasyonun ilaçla sinüs ritmine döndürülmesi: Üç antiaritmik ilaçla gerçeklestirilen randomize, plasebo-

kontrollü çalisma. *The Anatolian Journal of Cardiology*, 11(7):600, 2011.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Kjell Benson and Arthur J Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342(25): 1878–1886, 2000.

Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

Dean Billheimer. Predictive inference and scientific reproducibility. *The American Statistician*, 73(sup1): 291–295, 2019.

Michael Borenstein. Avoiding common mistakes in meta-analysis: Understanding the distinct roles of q, i-squared, tau-squared, and the prediction interval in reporting heterogeneity. *Research Synthesis Methods*, 15(2):354–368, 2024. doi: https://doi.org/10.1002/jrsm.1678. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1678.

David N Borg, Franco M Impellizzeri, Samantha J Borg, Kate P Hutchins, Ian B Stewart, Tamara Jones, Brenton J Baguley, Lucas BR Orssatto, Aaron JE Bach, John O Osborne, et al. Meta-analysis prediction intervals are under reported in sport and exercise medicine. *Scandinavian Journal of Medicine & Science in Sports*, 34(3):e14603, 2024.

Christopher J Bryan, Elizabeth Tipton, and David S Yeager. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, 5(8):980–989, 2021.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

Evgeny Burnaev and Ivan Nazarov. Conformalized kernel ridge regression. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, pages 45–52. IEEE, 2016.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Andrea Cipriani, Julian PT Higgins, John R Geddes, and Georgia Salanti. Conceptual and technical challenges in network meta-analysis. *Annals of internal medicine*, 159(2):130–137, 2013.

Andrea Cipriani, Toshi A Furukawa, Georgia Salanti, Anna Chaimani, Lauren Z Atkinson, Yusuke Ogawa, Stefan Leucht, Henricus G Ruhe, Erick H Turner, Julian PT Higgins, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet*, 391(10128):1357–1366, 2018.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39 (1):165–191, 2024.

John Concato, Nirav Shah, and Ralph I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892, 2000.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28, 2015.

Jonathan J Deeks and Julian PT Higgins. Statistical algorithms in review manager 5. *Statistical methods group of the Cochrane Collaboration*, 1(11), 2010.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=QKnWXX3aVm.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986. ISSN 0197-2456. doi:

https://doi.org/10.1016/0197-2456(86)90046-2. URL https://www.sciencedirect.com/science/article/pii/0197245686900462.

Alvan R Feinstein. Meta-analysis: statistical alchemy for the 21st century. *Journal of clinical epidemiology*, 48(1):71–79, 1995.

Shai Feldman, Bat-Sheva Einbinder, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, pages 624–626. PMLR, 2023.

Manuela L. Ferreira, Rob J.E.M. Smeets, Steven J. Kamper, Paulo H. Ferreira, and Luciana A.C. Machado. Can We Explain Heterogeneity Among Randomized Clinical Trials of Exercise for Chronic Back Pain? A Meta-Regression Analysis of Randomized Controlled Trials. *Physical Therapy*, 90 (10):1383–1403, 10 2010. ISSN 0031-9023. doi: 10.2522/ptj.20090332. URL https://doi.org/10.2522/ptj.20090332.

Christian Fiedler, Carsten W Scherer, and Sebastian Trimpe. Practical and rigorous uncertainty bounds for gaussian process regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7439–7447. AAAI, 2021.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Stephen Jay Gould. The median isn't the message. *Ceylon Medical Journal*, 49(4), 2010.

Peter D Grünwald. The e-posterior. *Philosophical Transactions of the Royal Society A*, 381(2247): 20220146, 2023.

Etash Kumar Guha, Eugene Ndiaye, and Xiaoming Huo. Conformalization of sparse generalized linear models. In *International Conference on Machine Learning*, pages 11871–11887. PMLR, 2023.

Gordon Guyatt, David Sackett, D Wayne Taylor, John Chong, Robin Roberts, and Stewart Pugsley. Determining optimal therapy–randomized trials in individual patients. *The New England journal of medicine*, 314(14):889–892, 1986.

Yuta Hamaguchi, Hisashi Noma, Kengo Nagashima, Tomohide Yamada, and Toshi A Furukawa. Frequentist performances of bayesian prediction intervals for random-effects meta-analysis. *Biometrical Journal*, 63(2):394–405, 2021.

Joachim Hartung and Guido Knapp. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in medicine*, 20(12):1771–1782, 2001.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Rachel Heyard, Leonhard Held, Sebastian Schneeweiss, and Shirley V Wang. Design differences and variation in results between randomised trials and non-randomised emulations: meta-analysis of rct-duplicate data. *BMJ Medicine*, 3(1), 2024. doi: 10.1136/bmjmed-2023-000709. URL https://bmjmedicine.bmj.com/content/3/1/e000709.

Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(1):137–159, 2009.

Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Chichester, UK, 2nd edition, 2019.

Falk Hoffmann, Katharina Allers, Tanja Rombey, Jasmin Helbach, Amrei Hoffmann, Tim Mathes, and Dawid Pieper. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. *Journal of Clinical Epidemiology*, 138:1–11, 2021.

Hilde M Huizenga, Ingmar Visser, and Conor V Dolan. Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology*, 64(1):1–19, 2011.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Joanna IntHout, John PA Ioannidis, and George F Borm. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard dersimonian-laird method. *BMC medical research methodology*, 14:1–12, 2014.

Joanna IntHout, John PA Ioannidis, Maroeska M Rovers, and Jelle J Goeman. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open*, 6(7):e010247, 2016.

John PA Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294 (2):218–228, 2005.

Dan Jackson and Ian R White. When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058, 2018.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Emir Karaçağlar, İlyas Atar, Süleyman Özbiçer, Atilla Sezgin, Salih Özçobanoğlu, Ayse Canan Yazici, Bülent Özin, and Haldun Müderrisoğlu. Amiodarone versus direct current cardioversion in treatment of atrial fibrillation after cardiac surgery. *Turkish Journal of Clinics and Laboratory*, 10(1):26–32, 2019.

George E Kochiadakis, Nikos E Igoumenidis, Michail E Hamilos, Maria E Marketou, Gregory I Chlouverakis, and Panos E Vardas. A comparative study of the efficacy and safety of procainamide versus propafenone versus amiodarone for the conversion of recent-onset atrial fibrillation. *The American journal of cardiology*, 99(12):1721–1725, 2007.

Michel Komajda, John JV McMurray, Henning Beck-Nielsen, Ramon Gomis, Markolf Hanefeld, Stuart J Pocock, Paula S Curtis, Nigel P Jones, and Philip D Home. Heart failure events with rosiglitazone in type 2 diabetes: data from the record clinical trial. *European heart journal*, 31(7):824–831, 2010.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165, 2019.

Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106 (4):749–764, 2019.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Luz M Letelier, Kamol Udol, Javier Ena, Bruce Weaver, and Gordon H Guyatt. Effectiveness of amiodarone for conversion of atrial fibrillation to sinus rhythm: a meta-analysis. *Archives of Internal Medicine*, 163(7):777–785, 2003.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329 – 1347, 2020. doi: 10.1214/19-AOS1849. URL https://doi.org/10.1214/19-AOS1849.

Tengyuan Liang and Benjamin Recht. Randomization inference when n equals one. *arXiv preprint arXiv:2310.16989*, 2023.

Jona Lilienthal, Sibylle Sturtz, Christoph Schürmann, Matthias Maiworm, Christian Röver, Tim Friede, and Ralf Bender. Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. *Research Synthesis Methods*, 15(2): 275–287, 2024.

Ziyu Liu, Fahad M Al Amer, Mengli Xiao, Chang Xu, Luis Furuya-Kanamori, Hwanhee Hong, Lianne Siegel, and Lifeng Lin. The normality assumption on between-study random effects was questionable in a considerable number of cochrane meta-analyses. *BMC medicine*, 21(1):112, 2023.

Andreas Lundh, Joel Lexchin, Barbara Mintzes, Jeppe B Schroll, and Lisa Bero. Industry sponsorship and research outcome. *Cochrane database of systematic reviews*, (2), 2017.

JoAnn E Manson, Nancy R Cook, I-Min Lee, William Christen, Shari S Bassuk, Samia Mora, Heike Gibson, David Gordon, Trisha Copeland, Denise D'Agostino, et al. Vitamin d supplements and prevention of cancer and cardiovascular disease. *New England Journal of Medicine*, 380(1):33–44, 2019.

Javier Abad Martinez, Umang Bhatt, Adrian Weller, and Giovanni Cherubin. Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6631–6639, 2023.

Luca Masserano, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, and Ann B Lee. Simulator-based inference with waldo: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. *Proceedings of Machine Learning Research*, 206, 2023.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. *Communications of the ACM*, 65(7):33–35, 2022.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Marilyn C Morris and Robert M Nelson. Randomized, controlled trials as minimal risk: an ethical analysis. *Critical care medicine*, 35(3):940–944, 2007.

Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9, 2017.

M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127, 2016.

Kengo Nagashima, Hisashi Noma, and Toshi A. Furukawa. pimeta: an r package of prediction intervals for random-effects meta-analysis. arXiv, 2021.

Eugene Ndiaye and Ichiro Takeuchi. Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32, 2019.

Eugene Ndiaye and Ichiro Takeuchi. Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1):151–176, 2023.

Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic learning theory*, pages 963–982. PMLR, 2021.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Ilia Nouretdinov, Thomas Melluish, and Volodya Vovk. Ridge regression confidence machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 385–392, 2001.

Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10: 1–13, 2017.

Christopher Partlett and Richard D Riley. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following reml estimation. *Statistics in medicine*, 36(2):301–317, 2017.

J. Pearl. *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009. ISBN 9780521895606. URL https://books.google.com/books?id=f4nuexsNVZIC.

Coby Penso and Jacob Goldberger. A conformal prediction score that is robust to label noise. *arXiv preprint arXiv:2405.02648*, 2024.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR, 2021.

Kirsty M. Rhodes, Rebecca M. Turner, Ian R. White, Dan Jackson, David J. Spiegelhalter, and Julian P. T. Higgins. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Statistics in Medicine*, 35(29):5495–5511, 2016. doi: https://doi.org/10.1002/sim.7090. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7090.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123(3):A12–A13, 1995.

Richard D Riley, Julian P T Higgins, and Jonathan J Deeks. Interpretation of random effects meta-analyses. *BMJ*, 342, 2011. ISSN 0959-8138. doi: 10.1136/bmj.d549. URL https://www.bmj.com/content/342/bmj.d549.

Mark Rippetoe. *Starting Strength: Basic Barbell Training*. Aasgaard Company, 3rd edition, 2017. ISBN 9780982522738.

Christian Röver. Bayesian random-effects meta-analysis using the bayesmeta r package. *arXiv preprint arXiv:1711.08683*, 2017.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Matthias Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

Jadbinder Seehra, Daniel Stonehouse-Smith, and Nikolaos Pandis. Prediction intervals reporting in orthodontic meta-analyses. *European journal of orthodontics*, 43(5):596–600, 2021.

Matteo Sesia, YX Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *arXiv preprint arXiv:2309.05092*, 2023.

John Shawe-Taylor and Robert C Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

Kurex Sidik and Jeffrey N Jonkman. On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics-Simulation and Computation*, 32(4):1191–1203, 2003.

Uri Simonsohn, Joseph Simmons, and Leif D Nelson. Above averaging in literature reviews. *Nature Reviews Psychology*, 1(10):551–552, 2022.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Richard S Slavik and Peter J Zed. Intravenous amiodarone for conversion of atrial fibrillation: Misled by meta-analysis? *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 24(6):792–798, 2004.

Teresa C Smith, David J Spiegelhalter, and Andrew Thomas. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in medicine*, 14(24):2685–2699, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

T. D. Stanley and Stephen B. Jarrell. Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys*, 3(2):161–170, 1989. doi: https://doi.org/10.1111/j.1467-6419.1989.tb00064.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.1989.tb00064.x.

Tom D Stanley, Evan C Carter, and Hristos Doucouliagos. What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin*, 144(12):1325, 2018.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=CAd6V2qXxc.

SV Subramanian, Rockli Kim, and Nicholas A Christakis. The "average" treatment effect: A construct ripe for retirement. a commentary on deaton and cartwright. *Social science & medicine*, 210:77–82, 2018.

Stuart P Thomas, Duncan Guy, Elisabeth Wallace, Roselyn Crampton, Pat Kijvanit, Vicki Eipper, David L Ross, and Mark J Cooper. Rapid loading of sotalol or amiodarone for management of recent onset symptomatic atrial fibrillation: a randomized, digoxin-controlled trial. *American heart journal*, 147(1):E3, 2004.

Simon G. Thompson and Julian P. T. Higgins. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21(11): 1559–1573, 2002. doi: https://doi.org/10.1002/sim.1187. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1187.

Ingrid Toews, Andrew Anglemyer, John LZ Nyirenda, Dima Alsaid, Sara Balduzzi, Kathrin Grummich, Lukas Schwingshackl, and Lisa Bero. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study. *Cochrane Database of Systematic Reviews*, (1), 2024.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138. URL https://ai.nejm.org/doi/abs/10.1056/AIoa2300138.

Areti Angeliki Veroniki. Random-effects meta-analysis methods in revman (cochrane statistical editor training 2022). YouTube video, 2022. URL https://www.youtube.com/watch?v=4gsaUl5uh70.

Areti Angeliki Veroniki, Dan Jackson, Ralf Bender, Oliver Kuss, Dean Langan, Julian P.T. Higgins, Guido Knapp, and Georgia Salanti. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods*, 10(1): 23–43, 2019. doi: https://doi.org/10.1002/jrsm.1319. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1319.

Wolfgang Viechtbauer. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, 2005.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *The Annals of Statistics*, pages 1566–1590, 2009.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Shirley V Wang, Sebastian Schneeweiss, Jessica M Franklin, Rishi J Desai, William Feldman, Elizabeth M Garry, Robert J Glynn, Kueiyu Joshua Lin, Julie Paik, Elisabetta Patorno, et al. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. *Jama*, 329(16):1376–1385, 2023.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, 119(545):122–135, 2024.

Hye Sun Yun, David Pogrebitskiy, Iain J Marshall, and Byron C Wallace. Automatically extracting numerical results from randomized controlled trials with large language models. *arXiv preprint arXiv:2405.01686*, 2024.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

Appendix A. Background

A.1. Outcomes and Effects

Let $x \in \mathcal{X}$ be features describing a trial. This consists of the prospectively-set criteria of its population, intervention, comparison, and measure of outcome, commonly abbreviated as PICO (Richardson et al., 1995). For example, x may include the duration of an exercise program and the minimum age of its participants. It may also include auxiliary information that was collected passively and retrospectively, though (as described in the next section) this may complicate the interpretation of the meta-analysis. x does not have to be numerical; it can be, for example, a published document describing a clinical trial. The number of participants in such a trial should not be intentionally encoded in x, since a treatment should be applicable to any number of people. However, avoiding implicit, unintentional correlations between trial design and trial size may be difficult or impossible. Let ξ encode factors which influence the treatment, but are neither controlled nor observed. For example, the effect of an exercise program may surreptitiously depend on the altitude of the training facility or the jobs of the participants.

In the Neyman-Rubin framework of potential outcomes (Neyman, 1923; Rubin, 1974), for a single participant denoted by ρ, $\rho(1) \in \mathbb{R}$ is the outcome when assigned the treatment, and $\rho(0) \in \mathbb{R}$ is the outcome when assigned the comparison. Each outcome may be a final measurement (such as the amount of strength gained after training), or its change from a baseline measurement, or the logarithm of the ratio of final to baseline. The difference $\rho(1) - \rho(0)$ is the individual effect of the treatment. The potential outcomes framework is challenging because we cannot observe both terms in $\rho(1) - \rho(0)$, since each participant is assigned to either the treatment or the comparison. The conditional average treatment effect (CATE), denoted by u, quantifies the expected difference between the treatment and comparison for a new participant:

$$u(x, \xi) = \mathbb{E}_{\rho} \left(\rho(1) - \rho(0) \mid x, \xi \right) \tag{2}$$

The CATE is usually defined solely in terms of the observed variables x. We include ξ to emphasize the influence of unobserved variables, which are sometimes ignored in causal inference.

A.2. Different Goals of Meta-Analysis

The CATE is the predictive target of meta-analysis. With high probability (typically 95%, with $\alpha = 0.05$), the CATE should lie within the predicted interval:

$$\mathbb{P}_{C,x,\xi} \left(u(x, \xi) \in C(x) \right) \geq 1 - \alpha \tag{3}$$

Rather than predicting relatively specific, tangible effects, meta-analysis often focuses on estimating more abstract, harder-to-verify quantities. Meta-analyses usually report a confidence interval $\mathrm{CI} \subset \mathbb{R}$ which, with high probability, should contain the average treatment effect (ATE, also known as the summary effect or grand mean):

$$\mathbb{P}_{\mathrm{CI}} \left(\mathrm{ATE} \in \mathrm{CI} \right) \geq 1 - \alpha \quad \text{where} \quad \mathrm{ATE} = \mathbb{E}_{x,\xi} \; u(x, \xi)$$

Whereas the confidence interval merely needs to capture the ATE, the prediction interval must capture most of the dispersion around it. (Formally, a prediction interval covers a random variable, and its coverage probability must also account for the randomness of that variable, whereas a confidence interval covers a fixed value). In the presence of significant heterogeneity, the confidence interval is much tighter than the prediction interval, and has little chance of capturing the effect of a future treatment. Due to this potentially unintuitive behavior, and the possibility of instilling overconfidence in evidence about the treatment, many prominent researchers strongly encourage systematic reviews to report prediction intervals (IntHout et al., 2016; Riley et al., 2011; Borenstein, 2024). According to some researchers, the relative ease of corroborating (or refuting) predictions makes them essential for scientific rigor and reproducibility (Billheimer, 2019).

These problems are exacerbated by the introduction of features (x) and larger numbers of trials (n), as proposed in this paper. Since confidence intervals are tighter than prediction intervals, it may be technically tempting to use untrusted priors to analogously tighten intervals for ATE. However, when considering many trials with substantially different features, ATE becomes a useless quantity (Simonsohn et al., 2022; Subramanian et al., 2018; Gould, 2010; Feinstein, 1995). It is arguably misleading to use features within a statistical analysis but to simultaneously obfuscate their existence in the reported statistic. This is why prediction intervals are presently the preferred solution concept.

While prediction intervals avoid some of the unintuitive pitfalls of confidence intervals, it is important to note that the predictive guarantee (3) has subtleties of

its own. It is a mixed observational-causal guarantee: coverage does not hold for all x, just marginally (on average) over x. For example, if $\alpha = 0.05$, then it is possible for coverage to be 99% for patients younger than 60 and only 80% for patients between 60 and 70, so long as the average is at least 95%. This marginal guarantee should not be confused with the following, stronger conditional guarantee, which does not average over x:

$$\mathbb{P}_{C,\xi}(u(x,\xi) \in C(x) \mid x) \geq 1-\alpha, \quad \text{for all } x \in \mathcal{X}$$

Achieving conditional coverage guarantees is not possible without further assumptions (Lei and Wasserman, 2014). Since prevalent meta-analysis algorithms do not involve x, their guarantees are of course marginal over x.

The guarantee (3) is most reliable when the distribution over x is explicitly specified by a generative model. If trial designs are actually chosen according to this distribution, and x consists solely of prospectively-set, controllable variables, then it is easy to sample future x for which the coverage guarantee holds. If x includes retrospectively-collected information, or the trials are designed according to unspecified criteria, then the guarantee becomes less meaningful.

A.3. Randomized Controlled Trials (RCTs)

An RCT enrolls m participants with potential outcomes $\rho_1, \ldots, \rho_m$. Uniformly at random, it assigns m_0 of them to group 0 (the comparison), and the remaining m_1 to group 1 (the treatment). Most RCTs do not report individual outcomes. Rather, they report the mean and (corrected) variance of the comparison outcomes as $y^{(0)}$ and $v^{(0)}$. The same statistics are reported for the treatment outcomes as $y^{(1)}$ and $v^{(1)}$. These are combined into y, the difference in means, and v, a sum of the squared standard errors (Deeks and Higgins, 2010). These statistics are defined as:

$$y^{(g)} = \frac{1}{m_g} \sum_{i \text{ in group } g} \rho_i(g) \qquad y = y^{(1)} - y^{(0)}$$

$$v^{(g)} = \frac{1}{m_g - 1} \sum_{i \text{ in group } g} (\rho_i(g) - y^{(g)})^2 \qquad v = \frac{v^{(0)}}{m_0} + \frac{v^{(1)}}{m_1}$$

Condensing the data into y and v has the following rationale. It can be shown that y is an unbiased estimate of the CATE:

$$\mathbb{E}(y \mid x, \xi) = \mathbb{E}(u \mid x, \xi)$$

Thus, as the RCT enrolls a very large number of participants, y converges to u, regardless of x and ξ. This is the primary reason why RCTs are so valuable. v is an estimate of y's variance around u, under conditions discussed in the next section.

A.4. Random-Effects Model of the Data

Meta-analysis is conducted upon n trials, each with data $X_i \in \mathcal{X}$, $Y_i \in \mathbb{R}$ and $V_i > 0$ for $i = 1, \ldots, n$. As discussed above, each trial's Y_i is centered around U_i, but varies around it due to its limited number of participants. Because Y_i is a sample average, by the central limit theorem, it is asymptotically normally distributed around U_i. The random-effects model of meta-analysis (DerSimonian and Laird, 1986; Higgins et al., 2009) asserts, as a simplifying assumption, that Y_i is exactly (not just asymptotically) normally distributed around U_i with true variance equal to the observed one. That is, $Y_i \sim N(U_i, V_i)$. This can be written in a way that highlights a key difference between the standard random-effects model and this paper's model:

$$Y_i(X_i,\xi_i) = \text{ATE} + \underbrace{U_i(X_i,\xi_i) - \text{ATE}}_{\text{between-trial heterogeneity}} + \underbrace{N(0,V_i)}_{\text{within-trial variation}} \tag{4}$$

The first and last terms are the same in both models. The random-effects model asserts that the middle term $U_i - \text{ATE} \sim N(0, \nu)$ where ν (often denoted by τ^2) is called the heterogeneity variance. By contrast, in this paper, U_i depends on the features X_i, and may also involve arbitrary (non-Gaussian) noise through ξ_i. Thus, this paper eliminates a normality assumption which is viewed as dubious in practice (Liu et al., 2023). The normality of within-trial variation, though less controversial, may be tenuous for small trials (Jackson and White, 2018).

A.5. Untrusted Data as a Probability Distribution

Independently of RCTs, practitioners and researchers often possess deep intuitions about the CATE. These intuitions arise from the lower levels of the evidence hierarchy: observational studies, individually-published cases, hands-on experience, and personal belief (Murad et al., 2016). It is difficult to rigorously infer causation from such untrusted (or "real-world" data,

since they are observational and may have deeply-embedded biases. Nonetheless, it is often found that untrusted data agree with RCTs (Benson and Hartz, 2000; Concato et al., 2000). Retrospectively, Toews et al. (2024) found the ratio of risk-ratios between RCTs and observational studies to be approximately 1.08. The prospective RCT-DUPLICATE trial found their Pearson correlation to be 0.82 (Wang et al., 2023), with much of the discrepancy attributable to readily-identified factors (Heyard et al., 2024). For example, observational claims data do not typically record whether treatment was initiated in a hospital, but this may affect the outcomes of RCTs.

Since untrusted data originate from different kinds of sources and experiences, they do not share the form of RCTs. A modern approach to capturing large, disparate collections of knowledge is to (pre)train foundation models. Such models are already being developed for healthcare (Moor et al., 2023; Singhal et al., 2023; Tu et al., 2024). When applied to meta-analysis, this approach would involve learning an embedding $\phi(x)$ which maps features x into a Euclidean space having inner product $\kappa(x, x') = \phi(x)^T \phi(x')$. On top of this embedding, a linear predictor of the CATE could be trained as $\mu(x) = w^T \phi(x)$. Practically, this representation (μ, κ) encompasses nearly every useful way of predicting the CATE. Mathematically, this representation constructs a Gaussian process, a probability distribution over functions $f : \mathcal{X} \mapsto \mathbb{R}$, with higher probability placed on f which could plausibly approximate the CATE (Kanagawa et al., 2018; Williams and Rasmussen, 2006). In this probabilistic perspective, $\mu(x) = \mathbf{E}_f f(x)$ and $\kappa(x, x') = \mathbf{E}_f(f(x) - \mu(x))(f(x') - \mu(x'))$. Gaussian processes are often used as prior probability distributions in Bayesian inference (Gelman et al., 1995). (See Section 2.1 for further comparison to Bayesian inference).

A significant restriction is that μ and κ are fixed relative to the data. In practical terms, this means the outcomes of the trials are not reincorporated into the prior. Otherwise, the trials could trivially, erroneously serve as their own reality check. Thus, although μ and κ are completely untrusted in terms of their veracity and utility, their provenance (especially the data used to generate them) must be clearly understood. Practices such as preregistration and data transparency can facilitate this understanding (Munafò et al., 2017). Importantly, this assumption is about the processes used to include data, which are under our control. It is not about the complex phenomena which generate the data itself. In this sense, it is much weaker than the assumptions of ignorability and positivity which are made in causal inference.

The assumption of fixed μ and κ is technically stronger than necessary. The following task description more precisely specifies the exchangeability requirement which is required for our techniques to apply.

Task 3 (Predicting Effects (Technical)) *Let $\mu : \mathcal{X} \to \mathbb{R}$ and $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Let $\bar{X} = [X; x] \in \mathcal{X}^{n+1}$, $\bar{U} = [U; u] \in \mathbb{R}^{n+1}$, $\bar{V} = [V; v] \in \mathbb{R}_+^{n+1}$, $\bar{M} = [\mu(\bar{X}_i)]_i$, and $\bar{K} = [\kappa(\bar{X}_i, \bar{X}_j)]_{i,j} \succ 0$ be random variables. Suppose, for any permutation σ of $\{1, \ldots, n+1\}$, the joint distribution of the $(\bar{X}_i, \bar{U}_i, \bar{V}_i, \bar{M}_i, [\bar{K}_{i,j}]_j)$ equals that of the $(\bar{X}_{\sigma(i)}, \bar{U}_{\sigma(i)}, \bar{V}_{\sigma(i)}, \bar{M}_{\sigma(i)}, [\bar{K}_{\sigma(i),\sigma(j)}]_j)$. Let $Y_i = U_i + \mathcal{E}_i$, where independently $\mathcal{E}_i | V_i \sim N(0, V_i)$. From $(\bar{X}, Y, V, \bar{M}, \bar{K})$, for a desired confidence level $\alpha \in (0, 1)$, produce an interval $C(x)$ such that $\mathbb{P}(u \in C(x)) \geq 1 - \alpha$, where the probability is over all the random variables.*

An advantage of this more technical formulation is that its underlying exchangeability assumption can be tested (Vovk, 2021). Thus, even when the prior has unknown provenance, a diagnostic hypothesis test can potentially check if its involvement in the meta-analysis is valid.

A.6. Standard Meta-Analysis Algorithms

As previously mentioned, prevalent algorithms for meta-analysis ignore the features x; in the parlance of the field, they perform mean-effect prediction rather than meta-regression. Thus, they simply return a single prediction interval $C \subset \mathbb{R}$ rather than a prediction band. Because the model (4) is not analytically solvable, there is no exact, rigorous frequentist prediction interval. Instead, there are many different formulae (Veroniki et al., 2019; Nagashima et al., 2021), each involving approximations which hold only as $n \to \infty$. Most of the prediction intervals have this form:

$$C = \widehat{\text{ATE}} \pm t \sqrt{\hat{\nu} + \widehat{\text{Var}}(\widehat{\text{ATE}})} \tag{5}$$

In this expression, the variance estimates $\hat{\nu}$ and $\widehat{\text{Var}}(\widehat{\text{ATE}})$ are usually algorithm-specific. More generally, t is the $1 - \frac{\alpha}{2}$ quantile of a Student t distribution with either $n-1$ or $n-2$ degrees of freedom. $\widehat{\text{ATE}}$ is an estimate of ATE, usually based upon inverse-variance

weighting:

$$\widehat{\text{ATE}} = \sum_i w_i Y_i \Big/ \sum_i w_i \quad \text{where} \tag{6}$$
$$w_i = \frac{1}{V_i + \hat{\nu}} \quad \text{for each } i = 1, \ldots, n$$

In practice, the most widely-used prediction interval is based on the classical heterogeneity estimator $\hat{\nu}$ of DerSimonian and Laird (1986), and an estimator $\widehat{\text{Var}}(\widehat{\text{ATE}})$ proposed by Higgins et al. (2009). When n is small, experimental evidence indicates this interval is too small to satisfy (3) with the desired probability $1 - \alpha$. To the best of our knowledge, this method does not have a proven coverage guarantee, so the following result is stated imprecisely.

Proposition 9 (Classical Prediction Interval) *Assume the model (4) with $U_i \sim N(ATE, \nu)$. Define the following quantities within (5):*

$$\hat{\nu} = \frac{Q - (n-1)}{S_1 + S_2/S_1} \qquad \widehat{Var}(\widehat{ATE}) = (\sum_i w_i)^2$$
$$\bar{Y} = \frac{\sum_{i=1}^n V_i^{-1} Y_i}{\sum_{i=1}^n V_i^{-1}} \qquad Q = \sum_{i=1}^n V_i^{-1}(Y_i - \bar{Y})^2$$
$$S_r = \sum_{i=1}^n V_i^{-r}$$

Then C, as defined in (5), approximately satisfies (3) as $n \to \infty$.

Partlett and Riley (2017) proposed an alternative prediction interval based upon restricted maximum likelihood (REML) and Hartung-Knapp-Sidik-Jonkman (HKSJ) estimators (Nagashima et al., 2021). REML obtains $\hat{\nu}$ and $\widehat{\text{ATE}}$ as the maximizers of a log-likelihood function $\ell(\hat{\nu}, \widehat{\text{ATE}})$ which is filtered to remove influences from irrelevant variables (Viechtbauer, 2005). It is not concave, so it cannot be maximized by standard algorithms. However, its stationary points $\partial\ell \;/\; d\hat{\nu} = 0$ (for fixed $\widehat{\text{ATE}}$) and $\partial\ell \;/\; d\widehat{\text{ATE}} = 0$ (for fixed $\hat{\nu}$) have closed-form expressions, so it is amenable to alternating maximization. The following estimator $\widehat{\text{Var}}(\widehat{\text{ATE}})$ was developed independently by Hartung and Knapp (2001) and Sidik and Jonkman (2003). Cochrane Statistical Methods and other groups endorse the use of HKSJ (IntHout et al., 2014; Veroniki, 2022; Veroniki et al., 2019). This method also does not have a proven coverage guarantee.

Proposition 10 (HKSJ Prediction Interval) *Assume the model (4) with $U_i \sim N(ATE, \nu)$. Initialize $\hat{\nu} = 0$. Alternate the updates to $\widehat{ATE}$ and w in (6) with the following update of $\hat{\nu}$, until a fixed point is approximately reached:*

$$\hat{\nu} \leftarrow \frac{\sum_{i=1}^n w_i^2((Y_i - \widehat{ATE})^2 - V_i)}{\sum_{i=1}^n w_i^2} + \frac{1}{\sum_{i=1}^n w_i}$$
$$\widehat{Var}(\widehat{ATE}) = \sum_{i=1}^n \frac{(Y_i - \widehat{ATE})^2 w_i}{(n-1)\sum_j w_j}$$

Then C, as defined in (5), approximately satisfies (3) as $n \to \infty$.

In addition to these frequentist intervals, Bayesian intervals for u can also be obtained (Smith et al., 1995; Gelman et al., 1995). These begin with prior distributions over ATE and ν. Improper (i.e. unnormalized) uniform priors are a default uninformative choice (Röver, 2017). Using the random-effects model as a likelihood, Bayes' theorem obtains the posterior distribution over ATE and ν, which induces a (normal) posterior distribution over u. From this posterior distribution, a prediction interval for u can be derived. Such intervals can be highly sensitive to the choice of uninformative prior, which is partially why Bayesian methods are less common in systematic reviews (Hamaguchi et al., 2021). Nonetheless, there are some circumstances where the flexibility of Bayesian methods is desirable. For example, the Bayesian approach can be extended to predicting trials. The posterior distribution for future $y \sim N(u, v)$ is just u's posterior with v more variance.

Proposition 11 (Bayesian Trial Prediction) *Let the prior distribution over ATE be improper uniform. Assume the likelihood (4) with $U_i \mid ATE, \nu \sim N(ATE, \nu)$. Then, recalling (6), the posterior predictive distribution conditioned on ν is $y \mid \nu = \hat{\nu} \sim N\left(\widehat{ATE}, (\sum_i w_i)^{-1} + \hat{\nu} + v\right)$. (Röver, 2017)*

A.7. The Ethics of Meta-Analysis

Healthcare is important, uncertain, and sometimes controversial. Evidence-based medicine was introduced to help resolve some of these issues, but it involves controversy of its own. It unavoidably privileges certain kinds of experiences and opinions over others. This paper does not introduce these problems, but it does operate in their midst. Let us examine how

these problems could be ameliorated or aggravated by our approach.

Currently, meta-analysis in evidence-based medicine is highly exclusionary. The "lower levels" of the evidence hierarchy are deprecated in favor of RCTs in an effort to preserve rigor and eliminate bias. However, this introduces some bias of its own. For example, RCTs are expensive to conduct. Any methodology that substantially prefers RCTs may be substantially influenced by funding agencies and associated institutions (Lundh et al., 2017). Furthermore, RCTs are not ethical to conduct in many situations (Morris and Nelson, 2007). Conformal meta-analysis recognizes that RCTs are especially valuable, but it holistically incorporates data of less rarified origin. Even when our methods do not lead to quantitative improvements, they are arguably more fair, inclusive, and comprehensive. They could ameliorate concerns that evidence-based medicine limits the autonomy of healthcare professionals (Armstrong, 2007).

However, conformal meta-analysis introduces additional computational and statistical complexity into the process of meta-analysis. This complexity could be exploited by bad actors, with negative societal consequences. For example, a malicious meta-analyst could sneak trial data into their prior to arrive at intentionally biased conclusions. To prevent such harms from occurring, any rigorous conclusions derived from conformal meta-analysis need to be accompanied by safeguards on the handling of data.

Appendix B. Supplemental Material

B.1. Additional Related Work

Efficient algorithms for full conformal prediction. In general, full conformal prediction is computationally intractable, requiring retraining at every possible value of y. However, there are many special cases in which the full conformal prediction set can be efficiently computed. Nouretdinov et al. (2001) and Burnaev and Nazarov (2016) describe fast algorithms for (kernel) ridge regression. Similarly exploiting piecewise linearity, Lei (2019) derives an algorithm for ℓ_1-regularized regression, including the lasso and elastic net. Homotopy and numerical continuation techniques can approximate the prediction set for regularized and sparse generalized linear models (Ndiaye and Takeuchi, 2019; Guha et al., 2023). Influence functions and root-finding techniques can be similarly employed (Martinez et al., 2023; Ndiaye and Takeuchi, 2023). Unlike these previous works, we seek a prediction interval $C(x, v)$ which isn't merely easy to compute, but is easy to mathematically analyze. In particular, we need to analyze its width as a function of v. Our simplifying idiocentricity condition is most closely related to Corollary 3.5 of Lei (2019), which describes when the elastic net's prediction set is an interval. Aside from the fact that this result applies to a different algorithm, its condition does not match ours, and does not lead to the same simplifications.

Meta-regression. A meta-regression fits the observed effects Y_i as a (typically linear) function of the features X_i (Stanley and Jarrell, 1989). Meta-regression is usually conducted to diagnose which features are responsible for heterogeneity. It can also generate useful hypotheses for future research, by identifying which features are associated with higher or lower effects. While meta-regression and conformal-meta-analysis are similar in form, there are a number of crucial differences. Most importantly, unlike conformal-meta analysis, meta-regression does not offer predictive guarantees for new x; the fit to the data is post-hoc and interpretive (Baker et al., 2009; Thompson and Higgins, 2002). The (non-predictive) statistical task in meta-regression is to determine which features have a statistically significant relationship with the effect (Huizenga et al., 2011). To limit spurious findings, meta-regression is typically performed on a small number of prespecified features. By contrast, conformal meta-analysis fits powerful, nonlinear models on a potentially large number of features. In conformal meta-analysis, the regression, as embodied by the prediction band C, is presented as the main result, not just an adjunct diagnostic.

Individual treatment effects. This paper improves predictions by tailoring them to specific patient populations described by x. However, it still averages over individuals within those populations. There are multiple approaches to accounting for this heterogeneity by predicting individual treatment effects. One approach is to perform n-of-1 trials, where a single individual serves as both the treatment and control by applying the treatment at different times (Guyatt et al., 1986; Liang and Recht, 2023). Another approach is to conduct causal inference, under stronger assumptions, on individual-level data from randomized and/or observational studies (Bica et al., 2021). As part of this approach, conformal prediction has been employed to obtain prediction intervals for potential outcomes (Lei and Candès, 2021), possibly as a function of a parameter Γ bounding the amount of

unobserved confounding (Jin et al., 2023; Yin et al., 2024). These approaches require individual-level data, different experimental designs, or stronger assumptions, which are worth pursuing primarily when individual (within-trial) variation is significant relative to between-trial variation. Whether this occurs depends on the nature of the treatment as well as the granularity of $\mathcal{X}$.

Bayesian priors. Conformal meta-analysis takes a prior probability distribution, along with trial data, and makes predictions from a posterior distribution — a process that mirrors Bayesian inference (Gelman et al., 1995). The choice of prior can substantially influence Bayesian inference, sometimes for the better: for example, informative, data-driven priors for the heterogeneity variance ν can mitigate excessive posterior uncertainty (Rhodes et al., 2016; Lilienthal et al., 2024). However, in a Bayesian meta-analysis, the prior can potentially hurt the empirical coverage of the reported intervals. As a simple example, even if all the trial data indicate a large treatment effect, a prior which heavily concentrates on zero effect would nonetheless result in tight posterior intervals around zero. Such behavior is inappropriate for systematic reviews, which are meant to resolve collective uncertainty among parties who do not necessarily share the same prior beliefs. One attempt to address this problem is to use uninformative priors. However, even such choices can seriously impact the empirical validity of a Bayesian meta-analysis (Hamaguchi et al., 2021). In conformal meta-analysis, by contrast, even strong beliefs can be safely encoded into the prior without breaking empirical coverage guarantees. In the aforementioned example of a concentrated, incorrect prior, conformal meta-analysis would merely yield loose intervals.

Uniform confidence bands. Prediction intervals also should not be confused with uniform confidence bands, which offer the following stronger guarantee, and do not involve unobserved ξ:

$$\mathbb{P}_C\left(\text{for all } x \in \mathcal{X},\ u(x) \in C(x)\right) \geq 1 - \alpha$$

Such bands have been developed for Gaussian process regression in the context of online optimization, where new points x are sequentially, adaptively chosen to minimize uncertainty about u (Srinivas et al., 2009; Chowdhury and Gopalan, 2017; Fiedler et al., 2021; Neiswanger and Ramdas, 2021). Since subsequent x are chosen adaptively using the band, it is essential for the band to hold for arbitrary x rather than just randomly-sampled x. Strictly speaking, these bands are correct for arbitrary μ and κ. However, their widths depend on the smoothness of u, as quantified by its norm in the reproducing kernel Hilbert space induced by κ. Since u is unknown, this quantity is also unknown. As a practical matter, when μ and κ can range from very good to very poor, the band is either very wide or unknown. Though conformal meta-analysis only offers prediction intervals with marginal coverage guarantees, their width and coverage do not depend on unknown quantities.

Utilizing unlabeled data. Trusted labels are generally considered a scarce resource in machine learning, especially compared to unlabeled data (i.e. x sampled from the marginal distribution of $\mathbb{P}$). Unlabeled data are commonly used to pretrain large foundation models (Dahl et al., 2011; Dai and Le, 2015). Semi-supervised learning studies how to rigorously use unlabeled data to improve predictions (Balcan and Blum, 2010). Angelopoulos et al. (2023a) recently proposed prediction-powered inference as an approach to safely tighten confidence intervals by using unlabeled data along with a prior derived from separate, untrusted data. In this approach, (1) the unlabeled data and prior (which is temporarily treated as correct) are used to estimate the parameter, (2) concentration inequalities are applied to bound the estimation error arising from limited unlabeled data, and (3) the labeled data are used to correct the estimation error due to inaccuracy of the prior. Subsequently, Zrnic and Candès (2024) proposed cross-prediction-powered inference, which has similar goals but does not utilize untrusted data. Instead, it splits the data (as in cross-validation) to train a prior. Such methods have been used to improve out-of-distribution causal inference (Demirel et al., 2024). However, these methods are not directly applicable to predictive meta-analysis, in which there are no available unlabeled data. Furthermore, these methods are designed to produce confidence intervals rather than prediction intervals.

Safely using untrusted data. Various endeavors in statistics and machine learning involve making predictions that are rigorously guaranteed, even though they use untrusted data. To some extent, all these techniques manage to circumvent the "garbage-in, garbage-out" principle. PAC-Bayesian generalization theory formalizes inductive bias as an (untrusted) prior probability distribution (Shawe-Taylor and Williamson, 1997; McAllester, 1998; Seeger, 2002). Its generalization bounds are tight when the prior and data align, so that a learning algorithm (producing a posterior distribution) can fit the data without di-

verging far from the prior. While PAC-Bayes is a very useful theoretical tool, conformal prediction bounds are quantitatively tighter, especially when n is small. In statistics, an untrusted prior distribution can be used to define an e-value, a nonnegative statistic whose mean is at most one (Neiswanger and Ramdas, 2021). Using its reciprocal as an unnormalized density leads to e-posteriors, which can be used as the basis for valid inferences and decisions (Grünwald, 2023). To derive confidence intervals with conditional coverage guarantees, likelihood-free inference methods can exploit untrusted prior information (Masserano et al., 2023). In computer science, algorithms can be infused with untrusted predictions, also called side information, advice, or hints (Mitzenmacher and Vassilvitskii, 2022). When the predictions are good, the algorithms run faster; when the predictions are bad, the algorithms retain acceptable worst-case performance. A prototypical example is binary search, which can be modified to run in $O(1)$ time given a good prediction of the target's index, and in $O(\log n)$ time no matter how bad the prediction was.

B.2. Computations for KRR

Let M and K be the mean and kernel function applied to the training features:

$$M = [\mu(X_1), \ldots, \mu(X_n)]^T \in \mathbb{R}^n$$
$$K = [\kappa(X_i, X_j)]_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$$

Given a parameter $\lambda \in \mathbb{R}$ and observations $U \in \mathbb{R}^n$, KRR learns the following posterior on the training features:

$$\widehat{M} = (\widehat{K}/\lambda)U + (K/\lambda + I)^{-1}M \qquad (7)$$
$$\widehat{K} = \lambda(K + \lambda I)^{-1}K$$

In full conformal prediction, KRR is applied to the training set (X, U) augmented by (x, u). We will use bars to denote this augmentation, so $\bar{X} = [X; x]$, $\bar{U} = [U; u]$. Let $m = \mu(x)$, $k = [\kappa(X_1, x), \ldots, \kappa(X_n, x)]^T$, $k_0 = \kappa(x, x)$, and:

$$\bar{I} = \begin{bmatrix} I & 0 \\ 0 & 1 \end{bmatrix} \quad \bar{K} = \begin{bmatrix} K & k \\ k^T & k_0 \end{bmatrix}$$
$$\bar{Q} := (\bar{K} + \lambda\bar{I})^{-1}\bar{K} = \begin{bmatrix} Q & q \\ q^T & q_0 \end{bmatrix}$$

Then, the augmented posterior mean is:

$$\begin{bmatrix} \widehat{M} \\ \hat{m} \end{bmatrix} = \bar{Q}\begin{bmatrix} U \\ u \end{bmatrix} + \overbrace{(\bar{K}/\lambda + \bar{I})^{-1}\begin{bmatrix} M \\ m \end{bmatrix}}^{\bar{t}}$$

So the differences between the observations and posterior means are:

$$\begin{aligned} \begin{bmatrix} U - \widehat{M} \\ u - \hat{m} \end{bmatrix} &= (\bar{I} - \bar{Q})\begin{bmatrix} U \\ u \end{bmatrix} - \bar{t} \\ &= \begin{bmatrix} (I - Q)U - qu \\ -q^T U + (1 - q_0)u \end{bmatrix} - \bar{t} \\ &= \begin{bmatrix} Au + B \\ au + b \end{bmatrix} \end{aligned}$$

with the abbreviations:

$$\begin{bmatrix} A \\ a \end{bmatrix} = \begin{bmatrix} -q \\ 1 - q_0 \end{bmatrix} \quad \begin{bmatrix} B \\ b \end{bmatrix} = \begin{bmatrix} I - Q \\ -q^T \end{bmatrix} U - \bar{t}$$

The augmented posterior kernel matrix is $\lambda\bar{Q}$. Thus, $S_i = \sqrt{\lambda Q_{ii}}$ and $s = \sqrt{\lambda q_0}$. To determine Z_i and z, decompose the differences between the observations and the posterior means. As before, denote augmentation with overlines, as in $\bar{\mathcal{E}} = [\mathcal{E}; \epsilon]$.

$$\begin{aligned} \begin{bmatrix} Y - \widehat{M} \\ y - \hat{m} \end{bmatrix} &= (\bar{I} - \bar{Q})(\bar{U} + \bar{\mathcal{E}} - \bar{M}) - \bar{z} \\ &= \begin{bmatrix} U - \widehat{M} \\ u - \hat{m} \end{bmatrix} - (\bar{I} + \bar{Q})\bar{\mathcal{E}} \\ &= \begin{bmatrix} U - \widehat{M} \\ u - \hat{m} \end{bmatrix} + \begin{bmatrix} (I - Q)\mathcal{E} - q\epsilon \\ -q^T\mathcal{E} + (1 - q_0)\epsilon \end{bmatrix} \end{aligned}$$

Now, calculate the mean squared error with respect to $\mathcal{E}_i \sim N(0, V_i)$ and $\epsilon \sim N(0, v)$:

$$\begin{aligned} &\mathbb{E}\,(Y_i - \widehat{M}_i)^2 \\ &= \mathbb{E}\,(U_i - \widehat{M}_i + (e_i - Q_i)^T\mathcal{E} - q_i\epsilon)^2 \\ &= (U_i - \widehat{M}_i)^2 + \mathbb{E}\left((1 - Q_{ii})\mathcal{E}_i - \sum_{j \ne i} Q_{i,j}\mathcal{E}_j - q_i\epsilon\right)^2 \\ &= (U_i - \widehat{M}_i)^2 + \underbrace{\overbrace{(1 - Q_{ii})^2 V_i + \sum_{j \ne i} Q_{i,j}^2 V_j}^{Z_i} + \underbrace{q_i^2}_{A_i^2} v}_{D_i} \end{aligned}$$

```python
def theorem4(Y, V_, M_, K_, alpha, eta, lam):
  I_ = eye(len(M_))
  t_ = solve(K_/lam + I_, M_)
  Q_ = solve(K_+lam*I_, K_)
  Q = Q_[:-1,:-1]
  q = Q_[-1,:-1]
  qo = Q_[-1,-1]

  V, v = V_[:-1], V_[-1]
  A = -q
  a = 1-qo
  B = Y - Q@Y - t_[:-1]
  b = -q@Y - t_[-1]
  # a is already positive; flip signs (wlog) so that a,Ai >= 0
  B *= sign(A) + (A == 0)
  A *= sign(A)
  S2 = lam*diag(Q)
  s2 = lam*qo
  D = square(I-Q) @ V
  d = square(q) @ V

  a2A2 = a**2*S2 - A**2*s2
  rho = eta*(D*s2 - d*S2 - a2A2*v)
  G = (A*B*s2 -a*b*S2) / a2A2
  H = sqrt(maximum(0, s2*S2*(A*b - a*B)**2 - rho*a2A2)) / a2A2

  return G, H
```

Algorithm 3: Python / NumPy code for common linear-algebraic computations described in Section 3. In this code, and the code throughout the paper, some elisions and deoptimizations are made for readability. In particular, import statements are omitted.

Similarly:

$$\begin{aligned}\mathbb{E}\,(y-\hat{m})^2 &= \mathbb{E}\,(u-\hat{m}-q^T\mathcal{E}+(1-q_0)\epsilon)^2 \\ &= (u-\hat{m})^2 + \overbrace{\underbrace{\sum_j q_j^2 V_j}_{d} + \underbrace{(1-q_0)^2}_{a^2} v}^{z}\end{aligned}$$

B.3. Proof of Theorem 6

Recalling Theorem 3 and the computations in Appendix B.2, we seek to prove:

$$\frac{|q_i|}{\sqrt{Q_{ii}}} < \frac{|1-q_0|}{\sqrt{q_0}} \iff \frac{|q_i|}{\sqrt{Q_{ii}\cdot q_0}} < \frac{|1-q_0|}{q_0}$$

Since $\bar{Q}$ is positive definite, its entries are the inner products among some vectors $f_0,\ldots,f_n$. In particular, $q_i = \langle f_i, f_0\rangle$. Thus, by the Cauchy-Schwartz inequality:

$$\begin{aligned}|q_i| &= |\langle f_i, f_0\rangle| \\ &\le ||f_i||\cdot||f_0|| = \sqrt{||f_i||^2\cdot||f_0||^2} = \sqrt{Q_{ii}\cdot q_0}\end{aligned}$$

Thus, it suffices to show that $1 < \frac{1-q_0}{q_0}$, that is, $0 < q_0 < \frac{1}{2}$. Since $\bar{Q}$ is positive definite, $q_0 > 0$ is obvious. To establish $q_0 < \frac{1}{2}$, let us examine the constraints on the last row of $\bar{Q}$. By the original definition of $\bar{Q}$, taking just the last column of $\bar{K}$:

$$\begin{bmatrix} q \\ q_0 \end{bmatrix} = (\bar{K}+\lambda\bar{I})^{-1}\begin{bmatrix} k \\ w \end{bmatrix}$$

Expanding and multiplying by both sides:

$$\left(\begin{bmatrix} K & k \\ k^T & k_0 \end{bmatrix} + \lambda\bar{I}\right)\begin{bmatrix} q \\ q_0 \end{bmatrix} = \begin{bmatrix} k \\ k_0 \end{bmatrix}$$

Expanding again:

$$\begin{bmatrix} K \\ k^T \end{bmatrix} q + \begin{bmatrix} k \\ k_0 \end{bmatrix} q_0 + \lambda \begin{bmatrix} q \\ q_0 \end{bmatrix} = \begin{bmatrix} k \\ k_0 \end{bmatrix}$$

This finally leads to the constraints:

$$(K + \lambda I)q = (1 - q_0)k$$
$$k^T q + \lambda q_0 = (1 - q_0)k_0$$

Inverting the first equation to solve for $q = (1 - q_0)(K + \lambda I)^{-1}k$ and plugging into the second yields:

$$(1 - q_0)k^T(K + \lambda I)^{-1}k + \lambda q_0 = (1 - q_0)k_0$$

If we take $\lambda = k_0$ then:

$$(1 - q_0)k^T(K + k_0 I)^{-1}k = (1 - 2q_0)k_0$$
$$\sum_{i=1}^{n} \frac{\tilde{k}_i^2}{\lambda_i + k_0} = \frac{1 - 2q_0}{1 - q_0}k_0$$

The left hand side is positive, so in order for the right hand to be positive, it is necessary that $q_0 < \frac{1}{2}$, as originally desired. To ensure λ (and KRR overall) remain symmetric, this analysis must be applied to any permutation of the data. Thus, λ should be larger than any diagonal entry of $\bar{K}$, not just k_0.

B.4. Proof of Theorem 7

Theorem 1 guarantees that $C(x, v)$ usually covers $y \sim N(u, v)$. We will use this guarantee to derive intervals $C(x)$ that usually cover u. We don't have a v to plug into $C(x, v)$, so we have to dig into how $C(x, v)$ works. The claim of Theorem 7 is that $C(x, 0)$ covers u just slightly less often than it covers y, so long as the level of noise correction η is not too high. This holds because of two counterbalancing properties of $C(x, v)$ that hold for all $v \geq 0$.

The first property is that most of the spread of $|N(0, v)|$ can be shaved from the edges of $C(x, v)$ without losing too many u. This is possible because, in meta-analysis, we care only about small α, ideally around 0.05. Since $C(x, v)$ covers y with high probability, there are only a few u closer than $|N(0, v)|$ to the ends of $C(x, v)$ — otherwise, bad flips of the noise could push too many y out of the interval, which would violate the coverage guarantee of $C(x, v)$. While this logic indicates shaving is a conceptually feasible strategy, it remains an abstract possibility, since we don't know v, and don't know how much to shave. (It should intuitively be $O(\sqrt{v})$, but constants matter).

The second property is that making η smaller limits the growth of $C(x, v)$. We mean this in a completely formulaic sense — we have reasonably concrete expressions for the endpoints of $C(x, v)$, and the following Theorem 12 shows they widen by $\sqrt{\eta v}$. When $\eta = 0$, $C(x, v)$ doesn't depend on v at all. In other words, when noise correction is disabled, $C(x, 0)$ must completely internalize the impact of noise, yielding a relatively wide interval. Larger settings of η allow $C(x, v)$ to grow more with v, allowing (relatively) thin intervals at small v. To concretely realize the shaving strategy, we just need to set η small enough so that, as a function of v, *the shaveable region within $C(x, v)$ grows as fast as $C(x, v)$ itself.* This allows us to obliviously use the baseline $C(x, 0)$. The conditional distribution $v \mid x$ is arbitrary and unknown, but any probability mass on $v > 0$ simply pushes more u within $C(x, 0)$.

The fact that $C(x, v)$ grows proportionally to $\sqrt{v}$ to capture the noise is not only intuitive, it is necessary. Most well-behaved learning algorithms should yield conformal intervals which grow (on average) at roughly this rate. Our ability to prove an exact growth rate, in the next lemma, relies on the simplicity of full conformal prediction for idiocentric linear smoothers.

Lemma 12 (Normal Interval Growth) *Let $C(x, v)$ be the interval from Theorem 4. For all $\eta \geq 0$ and $v > 0$, $C(x, v) \subseteq C(x, 0) \pm \sqrt{\eta v}$.*

Proof The interval for y depends on v only through ρ_i:

$$\frac{1}{\eta}\rho_i = Z_i s^2 - z S_i^2 = D_i s^2 - d S_i^2 - \overbrace{\left((a S_i)^2 - (A_i s)^2\right)}\, v$$

Under idiocentricity, $a/s > A_i/S_i$. Thus, the bracketed term above is positive, ρ_i decreases with v, the square-root radius in L_i (which subtracts ρ_i) increases with v, and the denominator in L_i is positive. Dividing by the denominator, the radius H_i is of the form $\sqrt{\ldots + \eta v} \leq \sqrt{\ldots} + \sqrt{\eta v}$. Neither the center G_i of L_i nor the other elided terms in the radius depend on v; the $\sqrt{\eta v}$ term is the only one which involves v. ■

The rest of the proof of Theorem 7 doesn't depend on either idiocentricity or linear smoothers. Theorem 13 formalizes the first property described above: most u are contained within $C(x, v)$ by a margin that grows with v. Finally, Theorem 14 shows that $C(x, v)$ can be shaved down to $C(x, 0)$, with η determining the loss in coverage of u.

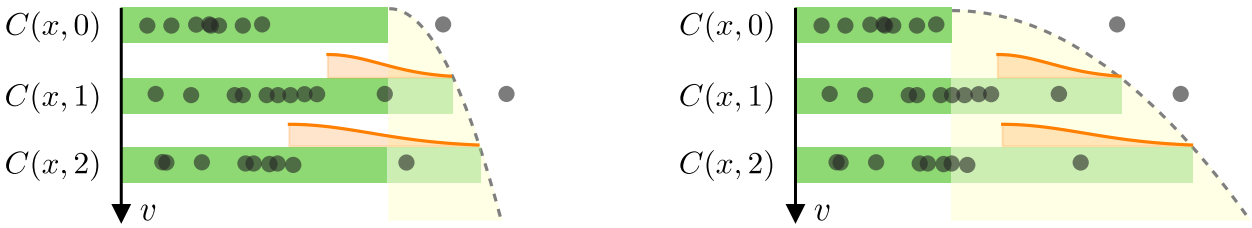


Figure 4: A high-level sketch of $C(x,0)$'s coverage of u, when η is small enough (left) versus too large (right). The gray dots are u, and its distributions conditioned on various v are shown. $C(x,0)$ is the dark green bar; as v increases, $C(x,v)$ increases by $\sqrt{\eta v}$, and that growth (in yellow) is shaved. The orange curves convey the spread of $|N(0,v)|$. With good η (left), $C(x,v)$ grows slowly compared to $|N(0,v)|$, which naturally pushes in the u (on average) as v increases. Thus, $C(x,0)$ is wide enough to contain most of the u, no matter what v is. On the right, when η is large, $C(x,v)$ adapts more dynamically to v, so $C(x,0)$ is smaller. Too many u in the yellow region are shaved.

Lemma 13 (Pay For Room) *Recall $y = u + \epsilon$ for $\epsilon \sim N(0,v)$. Let $w = [u-\epsilon, u+\epsilon]$, with possibly unsorted endpoints. If $\mathbb{P}(y \in C(x,v)) \geq 1-\alpha$, then $\mathbb{P}(w \subseteq C(x,v)) \geq (1-2\alpha)/(1-\alpha)$.*

Proof Abbreviate $C(x,v) = C$. The key property we repeatedly use is that y is one of the endpoints of w chosen uniformly at random, conditionally independent of the other data. If $w \not\subseteq C$, then either both of its endpoints are not in C, or exactly one of them isn't. In the former case, y clearly isn't in C; in the latter, it isn't with probability $\frac{1}{2}$. Let gray be the event that exactly one of w's endpoints is outside of C. First, we prove that:

$$\mathbb{P}(\mathsf{gray}) \leq 2\alpha \tag{8}$$

Let near denote both of w's endpoints are in C, and far that neither are in C, so that $\mathsf{near}, \mathsf{gray}, \mathsf{far}$ partition the probability space. By total probability, and the aforementioned reasoning about y:

$$\begin{aligned}
&\mathbb{P}(y \in C) \\
&= (1 - \mathbb{P}(\mathsf{gray}) - \mathbb{P}(\mathsf{far}))\mathbb{P}(y \in C \mid \mathsf{near}) \\
&\quad + \mathbb{P}(\mathsf{far})\mathbb{P}(y \in C \mid \mathsf{far}) + \mathbb{P}(\mathsf{gray})\mathbb{P}(y \in C \mid \mathsf{gray}) \\
&= (1 - \mathbb{P}(\mathsf{gray}) - \mathbb{P}(\mathsf{far}))(1) + \mathbb{P}(\mathsf{far})(0) + \mathbb{P}(\mathsf{gray})\frac{1}{2} \\
&\leq 1 - \mathbb{P}(\mathsf{gray}) + \mathbb{P}(\mathsf{gray})\frac{1}{2}
\end{aligned}$$

Combining this with the assumption yields (8). Next:

$$\begin{aligned}
\mathbb{P}(y \in C \mid w \not\subseteq C) &= \mathbb{P}(\mathsf{gray})\mathbb{P}(y \in C \mid \mathsf{gray}) \\
&= \mathbb{P}(\mathsf{gray})\frac{1}{2} \\
&\leq \alpha
\end{aligned}$$

With this inequality, the original claim follows from:

$$\begin{aligned}
&1 - \alpha \\
&\leq \mathbb{P}(y \in C) \\
&= \mathbb{P}(w \subseteq C, y \in C) + (1 - \mathbb{P}(w \subseteq C))\mathbb{P}(y \in C \mid w \not\subseteq C) \\
&\leq \mathbb{P}(w \subseteq C, y \in C) + (1 - \mathbb{P}(w \subseteq C))\alpha \\
&= \mathbb{P}(w \subseteq C) + (1 - \mathbb{P}(w \subseteq C))\alpha
\end{aligned}$$

Note this proof required ϵ to be symmetric, zero mean, and conditionally independent given its variance v, but not necessarily normally distributed. ■

Lemma 14 (Shaving) *If $\mathbb{P}(w \subseteq C(x,v)) \geq \frac{1-2\alpha}{1-\alpha}$, then $\mathbb{P}(u \in C(x,0)) \geq 1 - \frac{\alpha}{(1-\alpha)\mathrm{erfc}\sqrt{\eta/2}}$.*

Proof Abbreviate $C = C(x,v)$ and $\widetilde{C} = C(x,0)$. For the first inequality of the following block, the worst case is obtained when u is exactly one of the endpoints of $\widetilde{C}$ (say, the upper endpoint $\tilde{c}_+$), since that maximizes the distance from the endpoint of C (say, c_+), and therefore maximizes probability that w will still remain within C.

$$\begin{aligned}
\mathbb{P}(w \subseteq C \mid u \notin \tilde{C}) &\leq \mathbb{P}(\tilde{c}_+ + |\epsilon| \leq c_+) \\
&= \mathbb{P}(|\epsilon| \leq \sqrt{\eta v}) \\
&= \mathrm{erf}\sqrt{\frac{\eta}{2}}
\end{aligned}$$

By total probability:

$$\begin{aligned}
&\frac{1-2\alpha}{1-\alpha} \\
&\leq \mathbb{P}(w \subseteq C) \\
&= \mathbb{P}(u \in \tilde{C})\mathbb{P}(w \subseteq C \mid u \in \tilde{C}) + \mathbb{P}(u \notin \tilde{C})\mathbb{P}(w \subseteq C \mid u \notin \tilde{C}) \\
&\leq \mathbb{P}(u \in \tilde{C}) + (1 - \mathbb{P}(u \in \widetilde{C}))\mathrm{erf}\sqrt{\eta/2}
\end{aligned}$$

The desired claim follows from rearranging. ■

B.5. Predicting Clean Effects

If we had observed true effects U rather than noisy Y, then straightforward conformal prediction would yield a satisfactory interval for the true effect u.

Proposition 15 (Conformal Prediction) *Let $(X_1, U_1), \dots, (X_n, U_n), (x, u^*)$ be exchangeable. Let $[R; r]$ be the residuals of a symmetric learning algorithm upon the augmented data $[X; x]$ and $[U; u]$. Given any $\alpha \in (0, 1)$, let $\tau = \lceil (1-\alpha)(n+1) \rceil$. Define the prediction interval as $C(x) = \{u : r \text{ is among the } \tau \text{ smallest of } R_1, \dots, R_n\}$. Then $\mathbb{P}(u^* \in C(x)) \geq 1 - \alpha$. (Vovk et al., 2005)*

Let $C(x; \widehat{U})$ denote the prediction interval when $\widehat{U}$ is given as training data. Suppose we know a set $\mathcal{U}$ which contains the true U. If the outer interval $\widehat{C}(x)$ contains all $C(x; \widehat{U})$ over $\mathcal{U}$, then of course $\widehat{C}(x)$ contains $C(x; U)$ and inherits its coverage. Theorem 16 shows the uncertainty over U falling in that plausible set separates from fully-conformal KRR's uncertainty over u, given U. This is because $\mathcal{E}$ is independent from all else, given V.

Lemma 16 (Cover All Possibilities) *Let $\widehat{C}(x)$ contain all intervals induced by the ellipsoid $\mathbf{E}$:*

$$\mathbf{E} = \left\{ E : \sum_{i=1}^{n} \frac{E_i^2}{V_i} \leq \rho \right\} \quad \widehat{C}(x) = \bigcup_{E \in \mathbf{E}} C(x; \underbrace{U + \mathcal{E}}_{Y} - E)$$

Let $\rho > 0$ be chosen so that $\mathbb{P}_{\mathcal{E}}(\mathcal{E} \in \mathbf{E} \mid V) \geq 1 - \delta$. Then $\mathbb{P}(u \in \widehat{C}(x)) \geq (1-\alpha)(1-\delta)$.

Proof Let $C(x; U) = C(x)$ be the interval from Theorem 15 when computed on the true U. In the following, let rest denote X, U, x, u.

$$\begin{aligned}
&\mathbb{P}(u \in \widehat{C}(x)) \\
&\geq \mathbb{P}(u \in C(x), C(x) \subseteq \widehat{C}(x)) \\
&= \mathbb{E}_V \mathbb{E}_{\mathsf{rest}} \left(\mathbf{1}(u \in C(x)) \cdot \mathbb{P}_{\mathcal{E}}(C(x) \subseteq \widehat{C}(x) \mid V, \mathsf{rest}) \right) \\
&\geq \mathbb{E}_V \mathbb{E}_{\mathsf{rest}} \left(\mathbf{1}(u \in C(x)) \cdot (1-\delta) \right) \\
&= (1-\delta) \mathbb{P}_{\mathsf{rest}, V}(u \in C(x)) \\
&\geq (1-\delta)(1-\alpha)
\end{aligned}$$

A sufficient condition for $C(x; U) \subseteq \widehat{C}(x)$ is that $\mathcal{E} = E$ for some $E \in \mathbf{E}$, i.e. that $\mathcal{E}$ belongs to the ellipsoid. Note that $\widehat{C}(x)$ depends on U but this condition does not. Thus:

$$\begin{aligned}
\mathbb{P}_{\mathcal{E}}(C(x) \subseteq \widehat{C}(x) \mid V, \mathsf{rest}) &\geq \mathbb{P}_{\mathcal{E}}(\mathcal{E} \in \mathbf{E} \mid V, \mathsf{rest}) \\
&= \mathbb{P}_{\mathcal{E}}(\mathcal{E} \in \mathbf{E} \mid V) \\
&\geq 1 - \delta
\end{aligned}$$

■

This lemma doesn't make any smoothness assumptions on how $C(x; \hat{U})$ changes as $\widehat{U}$ varies away from U; it relies on the coverage of exactly $C(x; U)$, but not of any slight perturbation $C(x; \widehat{U})$. Furthermore, the lemma does not depend specifically on the normal distribution of $\mathcal{E}$, just that we know a set $\mathbf{E}$ which captures it with probability $1 - \delta$. For Gaussian noise, this is an ellipsoid of appropriate scale. This proof does not depend on the geometry of $\mathbf{E}$, just the fact that it contains $\mathcal{E}$ with high probability, and can be computed from Y and V. Thus, this overall strategy can be extended to handle non-Gaussian noise.

The previous lemma converts the statistical problem of covering u into the purely computational problem of determining the endpoints of $\widehat{C}(x)$. When $\eta = 0$, and if U is provided in lieue of Y, Algorithm 2 computes the interval $C(x)$ specified in Theorem 15. This allows us to concretely bound the endpoints of $\widehat{C}(x)$ as the following two optimization problems:

$$\min_{E \in \mathbf{E}} \max\{\text{bottom } n - \tau + 1 \text{ lower ends of } L_1, \dots, L_n\} \tag{9}$$

$$\max_{E \in \mathbf{E}} \min\{\text{top } n - \tau + 1 \text{ upper ends of } L_1, \dots, L_n\} \tag{10}$$

Though this a nonconvex optimization problem, it has useful structure. From Theorem 4, recall that the endpoints equal $G_i \pm H_i$, where $H_i = |\Delta_i|$. We are using $\eta = 0$, which implies $\rho_i = 0$, and in turn simplifies the equations for these variables. Recalling the equations from Appendix B.2 and Theorem 4, the conformal idiocentric KRR equations are a set of constraints in the variables $B, b, \widehat{U}$ and E, involving constants $a, A, s, S, Q, q, \bar{t}, Y$ and V:

$$G_i = \frac{A_i B_i s^2 - ab S_i^2}{(aS_i)^2 - (A_i s)^2} \quad \Delta_i = s S_i \frac{A_i b - a B_i}{(aS_i)^2 - (A_i s)^2}$$

$$\begin{bmatrix} B \\ b \end{bmatrix} = \begin{bmatrix} I - Q \\ -q^T \end{bmatrix} \widehat{U} - \bar{t} \quad \widehat{U} = Y - E \quad \sum_{i=1}^{n} \frac{E_i^2}{V_i} \leq \rho$$

The four constraints on the left are linear. The quadratic constraint is convex, since $V_i > 0$. Thus,

the above constraints are convex. Thus, despite the nonconvexity of the objective, the problems (9) and (10) may be amenable to semidefinite programming relaxations, robust optimization, and/or nonconvex optimization. The full version of this paper investigates this approach.

B.6. Simulation Details and Full Results

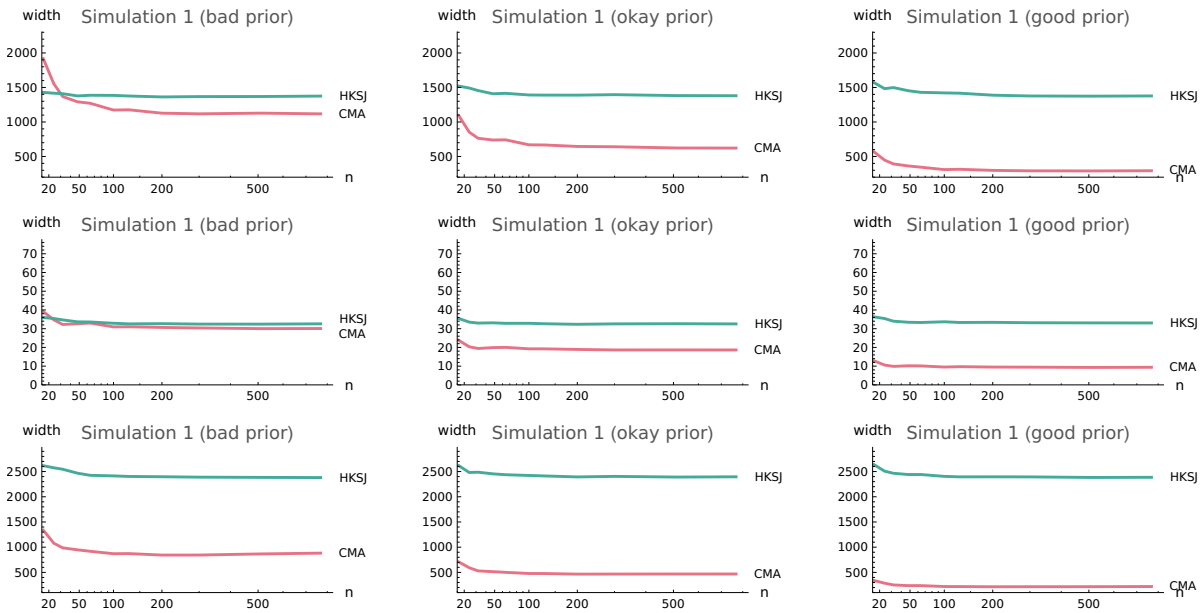
The simulations were performed using three partially-synthetic biomedical datasets from the Penn Machine Learning Benchmark (Olson et al., 2017): 1196_BNG_pharynx, 1201_BNG_breastTumor, and 1193_BNG_lowbwt. We randomly subsample training data (X, U) as well as test data (x, u). The kernel matrix K is generated using either the Gaussian or Laplace kernel as κ. For consistency across datasets having different scales, a parameter $\mathsf{effect\ noise} > 0$ is introduced, and the distribution of V is constructed to satisfy $\mathsf{effect\ noise} = \mathbb{E}(V_i)^2/\mathbb{E}|U_i|$. Specifically $V_i \sim \mathrm{Exp}(1) \cdot \sqrt{\mathsf{effect\ noise} \cdot \mathbb{E}|U_i|}$. Similarly, to produce prior means M of varying quality, a parameter $\mathsf{prior\ error} > 0$ is introduced, and the distribution of M satisfies $\mathrm{MSE}(M, U) = \mathsf{prior\ error} \cdot \mathbb{V}(U)$. Furthermore, the difference between M and U should not be purely random — otherwise, using KRR to explain this difference would be hopeless. Instead, we generate a random offset function $\tilde{f}(x) = \sum_i g_i \kappa(\tilde{x}_i, x)$ for random held-out data $\tilde{x}_i$ and $g_i \sim N(0, 1)$. Since $\tilde{f}$ is an RKHS element generated from random data, there is some hope in approximating it using the training data. Letting $\tilde{F}$ be $\tilde{f}$ applied to the training features, we generate $M = p\tilde{F} + (1-p)U$ where $p = \sqrt{\mathsf{prior\ error} \cdot \mathbb{V}(U)/\mathrm{MSE}(U, \tilde{F})}$.

All simulations are averaged over 32 random splits. Intervals are computed for between 256 and 768 test data in each run. Due to the efficiency of our proposed algorithms, all experiments are capable of running on a free Google Colab instance.

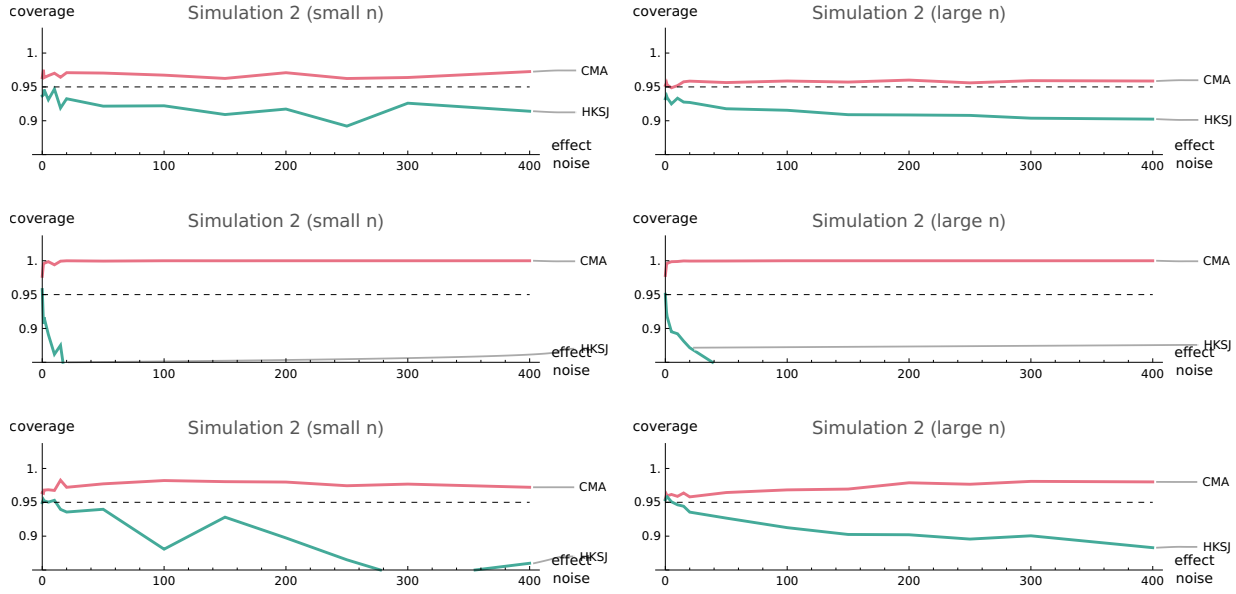
B.7. Case Study Details

We follow the meta-analysis process illustrated in Figure 1. First, we determine the domain $\mathcal{X}$ of x. Helpfully, Letelier et al. (2003) identified 10 potentially-relevant features, such as mean age, mean AF duration, and amiodarone therapy protocol (e.g. "IV, 5 mg/kg in 30 min + 10 mg/kg in 20 h" or "Oral, 600 mg/d for 3 wk"). In order to extract these features from the trial, we give their published PDFs to a publicly-available language model, along with a prompt including example output. This extraction is fairly reliable, echoing the experience of Yun et al. (2024). Next, parsing code (also written by the language model) converts the extracted textual features to numerical vectors x. As exemplified in Figure 7, this parsing can be tedious and error-prone, even with a state-of-the-art LLM. Our final predictions involve three additional extracted features: total amiodarone dosage in the first 24 hours, whether mean AF duration was above or below 48 hours, and the number of patients (which is a sensible feature when predicting trials rather than effects).

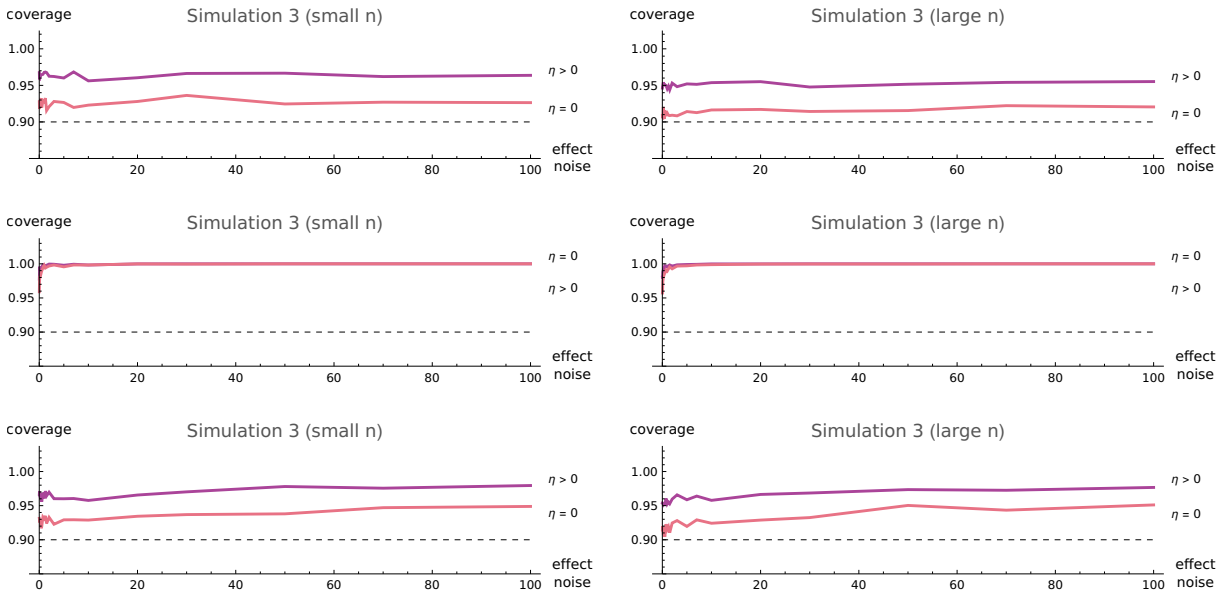
In lieue of a powerful pretrained foundation model, we base μ and κ on the critique of Slavik and Zed (2004). They describe how multiple sources of heterogeneity, such as dosage, could impact the effect of amiodarone. Most importantly, amiodarone has a relatively slow course of action, whereas patients with recent-onset AF (usually defined as an AF duration of less than 48 hours) have a high chance of spontaneously reverting to normal sinus rhythm. (Letelier et al. (2003) also noted this pattern). With recent-onset AF, median spontaneous conversion rates are "11% at 2 hours after admission, 18% at 3 hours, 25% at 4 hours, 31% at 6 hours, 39% at 8 hours, 38% at 12 hours, 58% at 24 hours, and 67% at 48 hours.". This compares to only 0–8% within the first 72 hours for patients with persistent AF. We identify 8 further trials which compared amiodarone to an active comparison. We compute pseudo-effects (as relative risk) by taking the ratio of the observed probability of conversion under amiodarone, over the aforementioned estimated probability of spontaneous conversion over time. Such indirect comparison is reminiscent of how network meta-analysis works (Cipriani et al., 2013). We trained a ReLU deep network upon the relevant features in these synthetically-labeled data.



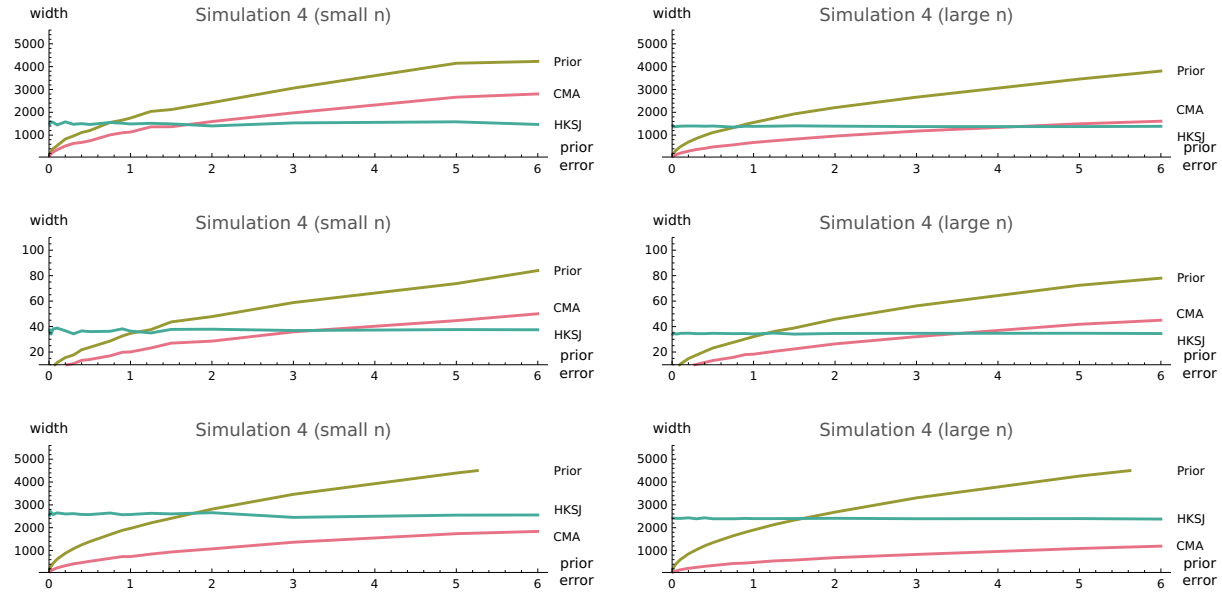
Simulation 1: Rows are different datasets; the different columns, from left to right, set prior error equal to 3.0, 0.9, and 0.2, respectively. $\alpha = 0.1$ and effect noise $= 0.5$ were used.



Simulation 2: Rows are different datasets. $n = 50$ and $n = 200$ are used in the left and right columns, respectively. prior error is set low to 0.2.



Simulation 3: Rows are different datasets; $n = 50$ and $n = 200$ are used in the left and right columns, respectively. $\alpha = 0.1$ and prior error $= 0.1$ were used.



Simulation 4: Rows are different datasets; $n = 16$ and $n = 200$ are used in the left and right columns, respectively. A low effect noise $= 0.02$ was set, along with $\alpha = 0.1$.

```
Can you extract the following features
 from the attached PDF paper? I gave
example values, from another paper,
which should be replaced with the
actual values in this paper. The only
relevant outcome is conversion to
normal sinus rhythm. Also, create a
new key like "Results": [a, b, c, d]
where a is the number of amiodarone
patients converted to sinus rhythm, b
is the total number of amiodarone
patients, c is the number of
comparison patients converted to sinus
 rhythm, and d is the total number of
comparison patients. Answer as JSON.

{"Name": "Villani et al.11 (Italy)
2000", "Features": { "Amiodarone
Therapy Protocol": "Oral, 400 mg/d for
 1 mo", "Comparison Treatment": "Oral
digoxin, 0.25 mg/d or oral diltiazem
hydrochloride 180- 360 mg/d for 1 mo",
 "Time to Outcome Measure": "1 mo", "
Number of Amiodarone Patients": "44",
"Number of Control Patients": "30", "
Fraction with CV Disease": "47", "Mean
 Left Atrium Size, mm": "50", "Mean AF
 Duration": "17 wk", "Mean Age": "58",
 "Fraction Male": "67", "Adequate
Concealment of Treatment": "No", "
Follow-up Fraction": "100", "Masked
Patients": "Yes", "Masked Caregiver":
"no", "Masked Assessor": "no" }}
```

Figure 5: Prompt used to extract relevant data from trial PDFs.

```
In the attached JSON list, each
element represents a study described
by the "Features" attribute. Convert
these features to real numbers so they
 can be provided to a learning
algorithm.

* amiodarone treatment should be the
total dosage, in milligrams, which is
given over the first 24 hours. If the
dosage is specified per kg bodyweight,
 then take into account the average
bodyweight of the patients.

* comparison treatment should be
converted to [0,1], where 0 denotes
placebo and 1 an intensive, high dose
comparison regimen.

* if the fraction of male patients is
unknown, just assume it is 0.5.

* fraction with CV disease and
followup fraction were reported as
integers, so for example 78 should be
converted to 0.78.

* number of control and amiodarone
patients should be just copied over as
 integers

* mean AF duration and time to outcome
 measure should be converted to -1 for
 <= 48 hours and 1 for > 48 hours

* mean left atrium size and mean age
should be rescaled to \[-1,1\] where 0
 is the average of the feature, -1 is
the minimum, and 1 is the maximum

* the boolean features should be
rescaled to \[-1, 1\], where -1 means
false, 1 means true, and 0 means not
present or not confident.

* include the same keys for all the
studies, using the original key names.

Answer as JSON; no further explanation
 is necessary.
```

Figure 6: Prompt used to convert extracted data to numerical features.

```python
def parse_dosing_protocol(protocol):
    if protocol is None or protocol.lower() == 'not specified':
        return 0

    weight = 70  # Average body weight in kg
    total_mg = 0  # Initialize total milligrams

    # Normalize and break down the protocol into components
    protocol = protocol.lower().replace('over', 'in').replace('plus', ',')
    phases = protocol.split('+')

    for phase in phases:
        parts = phase.split(',')
        for part in parts:
            part = part.strip()
            tokens = part.split()
            dose = 0
            rate_based = False
            duration = 24  # Default duration is 24 hours unless specified

            # Parse the dose and units
            for i, token in enumerate(tokens):
                try:
                    # Attempt to convert token to float to find numeric values
                    potential_dose = float(token)

                    # Check for units immediately following the numeric value
                    if i + 1 < len(tokens):
                        unit = tokens[i + 1]
                        if 'g' in unit and 'mg' not in unit:
                            potential_dose *= 1000  # Convert grams to milligrams
                        elif 'mg/kg' in unit:
                            potential_dose *= weight  # Convert to total mg for given weight

                        # Determine if the dose is time-bound
                        if 'hour' in unit or 'h' in unit or 'min' in unit:
                            rate_based = True  # The dose is a rate per time
                            duration = extract_duration(part)
                            if 'min' in unit:
                                duration /= 60  # Convert minutes to hours
                    dose = potential_dose
                    break
                except ValueError:
                    continue  # Not a number, move to next token

            # Apply the dose calculation based on the duration and whether it's rate-based
            if rate_based:
                total_mg += min(duration, 24) * dose  # Apply the rate up to 24 hours
            elif 'day' in part:
                if 'first' in part or '1 day' in part or '1 week' in part:
                    total_mg += dose  # Apply if it specifies the first day or week
            else:
                total_mg += dose  # Single dose or calculated for the duration

    return total_mg
```

Figure 7: Python code generated by GPT-4 to parse and convert amiodarone therapy protocols. Generating this code required multiple rounds of interaction with the language model. This code still has mild bugs, which are left untouched to accurately convey contemporary expectations about in-context parsing.