arXiv:2605.03819v1 [stat.ME] 5 May 2026

Meta-Analysis of High-Dimensional Surrogate Markers

Arthur Hughes[1,2], Rodolphe Thiébaut[1,2,3], Layla Parast[4], and Boris P. Hejblum[1,2,*]

[1]INSERM Bordeaux Population Health Research Center, INRIA SISTM, University of Bordeaux, F-33000 Bordeaux, France

[2]Vaccine Research Institute, F-94000 Créteil, France

[3]Centre Hospitalier Universitaire de Bordeaux, Service d'Information Médicale, INSERM, Bordeaux F-33000, France

[4]Department of Statistics and Data Science, University of Texas at Austin, Austin, TX 78712, United States of America

[*]**Corresponding author:** Boris P. Hejblum, Bordeaux Population Health, Université de Bordeaux, 146 Rue Léo Saignat 33076 Bordeaux, France. Email: `boris.hejblum@u-bordeaux.fr`

Abstract

When direct measurement of a clinically relevant primary endpoint in a clinical trial is infeasible, a surrogate endpoint may be used instead to infer treatment effects. Trial-level surrogates predict the average treatment effect on the primary endpoint and may be evaluated within the meta-analytic framework. However, traditional methods are ill-suited to the complex high-dimensional data now increasingly collected in modern trials, such as omics data. Although methods for high-dimensional surrogate evaluation exist, they have largely been developed for

single-trial settings and therefore cannot assess surrogate generalisability. Here, we propose RISE-Meta, an approach for evaluating trial-level surrogate markers in the multi-trial, high-dimensional setting. In the first stage, an existing nonparametric method is applied to individual participant data to derive study-level surrogacy metrics for each candidate marker. Next, random-effects meta-analysis combines these metrics across studies, and equivalence testing provides operational criteria for surrogate validity. Finally, a subset of candidates is combined into a composite signature through a weighting scheme to improve surrogacy relative to any individual candidate. We evaluate RISE-Meta in both simulation studies and real data applications. In an application to high-dimensional data, we analyse gene expression as trial-level surrogate markers for the antibody response to seasonal influenza vaccination, while in a low-dimensional application we compare RISE-Meta to a reference meta-analytic approach and observe strong agreement between the two.

Keywords: high-dimension, meta-analysis, nonparametric, surrogate marker, trial-level surrogate

1 Introduction

The main objective of a clinical trial is to evaluate the effect of a treatment on a clinically relevant primary endpoint. However, direct measurement of the primary endpoint may not be feasible due to cost, time, or ethics. In such cases, one may instead rely on a surrogate marker, a variable used to infer treatment effects on the primary endpoint. For example, in trials evaluating the efficacy of yellow fever vaccination in European populations, relying on infection rates is problematic since yellow fever virus is not endemic to Europe. Treatment efficacy may be assessed instead using neutralising antibody titres, which serve as a surrogate marker for infection rates (Jonker, Visser, and Roukens, 2013).

However, establishing the validity of a candidate surrogate has been a longstanding problem in biostatistics, and has been approached using multiple paradigms. Among these is the meta-analytic framework, which evaluates surrogate candidates using data from multiple clinical trials (Alonso et al., 2004; Buyse et al., 2000; Daniels and Hughes, 1997; Molenberghs et al., 2002). Where data from multiple trials are available, the meta-analytic framework is attractive as it exploits between-trial heterogeneity to assess whether surrogates predict treatment effects across clinical and demographic settings.

The historical development of the meta-analytic framework has clarified the conceptualisation of surrogate markers. Buyse et al. (2000) introduced the distinction between *individual-level surrogates* (ILS), associated with the primary endpoint at the participant level, and *trial-level surrogates* (TLS), for which the average treatment effect on the surrogate predicts the average treatment effect on the primary endpoint across trials. Two metrics, R^2_{indiv} and R^2_{trial}, were proposed to assess surrogacy at the individual- and trial-level, respectively. These distinctions are important as ILS and TLS differ both conceptually (Korn, Sachs, and McShane, 2016) and in their use cases (Alonso et al., 2005). In this work, we focus on meta-analytic methods for evaluating trial-level surrogate markers.

Advances in the cost and accuracy of high-throughput technologies have enabled more routine generation of *high-dimensional* biomarkers in modern biomedical studies. In such settings, the number of candidate markers can exceed the number of samples by

several orders of magnitude. A typical example is gene expression profiling, where tens of thousands of genes are measured simultaneously in treated and control individuals. These data can characterise the molecular mechanisms underlying treatment effects and identify biomarkers predictive of the host response (Prachi, Biagini, and Bagnoli, 2012). Thus, high-dimensional markers are both practical and biologically relevant candidates for surrogate markers, but their scale and complexity pose challenges for evaluation.

Public data-sharing initiatives have generated an abundance of high-dimensional datasets for the same treatment, accessible through repositories such as the *Cancer Genome Atlas*, *Gene Expression Omnibus*, and *ImmPort* (Bhattacharya et al., 2018; Edgar, 2002; Weinstein et al., 2013). Although individual studies are often limited in sample size, their aggregation can produce large datasets. This naturally motivates the use of meta-analytic approaches, which provide a principled framework for combining information across studies. Such approaches have already been applied in high-dimensional settings outside of the context of surrogate evaluation (Choi et al., 2003; Liyanage et al., 2024; Tseng, Ghosh, and Feingold, 2012).

However, although traditional meta-analytic approaches for surrogate evaluation are well established, they are not well suited to high-dimensional settings. A key limitation is their reliance on strong distributional assumptions for individual participant data (IPD). For example, the methods described by Buyse et al. (2000) and Molenberghs et al. (2002) assume a bivariate normal distribution for the surrogate and the primary endpoint. These assumptions become increasingly hard to verify and restrictive as the dimensionality and complexity of the data grow, limiting the applicability of such methods in many modern contexts.

The evaluation of high-dimensional surrogate markers is an emerging area of research, with only a few methods currently addressing this setting. Agniel et al. (2022) developed methodology to estimate the proportion of treatment effect explained (PTE) by a group of markers but did not provide quantitative rules for selecting individual markers. Zhou, Zhao, and Parast (2022) used high-dimensional linear mediation models to estimate the PTE, but their method is not robust to model misspecification. Hughes et al. (2025) introduced

RISE, a two-stage, nonparametric approach which is robust to the distribution of the IPD and suitable for small sample sizes. A shared limitation of all these approaches is that they operate within the single-trial framework, precluding the evaluation of generalisability and effective use of the wealth of available public data.

In this work, we extend the work of Hughes et al. (2025) by adapting RISE to the meta-analytic setting, enabling the evaluation of high-dimensional surrogate markers across multiple trials. The paper is structured as follows. Section 2 summarises the existing nonparametric approaches for surrogacy evaluation and introduces RISE-Meta, our meta-analytic extension. Section 3 presents a simulation study assessing its statistical properties. In Section 4, we apply RISE-Meta to real datasets. First, in a typical high-dimensional setting, we evaluate gene expression as a trial-level surrogate for the antibody response to seasonal influenza vaccination. Then, we apply RISE-Meta to classical low-dimensional datasets to compare the results with a reference meta-analytic approach introduced in Buyse et al. (2000). Section 5 concludes with a discussion of strengths, limitations, and future research directions.

2 Methods

2.1 Notation and data structure

Active versus control design

In this setting, we observe data from $m = 1, \ldots, M$ trials, each with $i = 1, \ldots, n_m$ individuals receiving the same binary treatment $A_{i,m} \in \{0, 1\}$, where 1 denotes active treatment and 0 denotes control. Let $Y^a_{i,m}$ and $\boldsymbol{S}^a_{i,m} = (S^a_{i,1,m}, \ldots, S^a_{i,J,m})^T$ denote the primary endpoint and vector of J common surrogate candidates, respectively, for individual i in trial m receiving treatment a. Without loss of generality, we assume that larger values of Y and S_j correspond to better outcomes. In addition, we assume the trials are independent and that the data within each trial are independent and identically distributed.

Paired design

Trials with high-dimensional data do not always follow a randomised treatment/control regime, but can instead consist of the longitudinal measurements of biomarkers in the same individuals before and after receipt of treatment. In this case, for each individual and a given post-treatment timepoint, we have paired measurements $\{Y^0_{i,m}, Y^1_{i,m}, \boldsymbol{S}^0_{i,m}, \boldsymbol{S}^1_{i,m}\}$, where the superscripts $a = 0$ and $a = 1$ refer to pre- and post-treatment measurements, respectively.

2.2 Existing nonparametric approaches for surrogate evaluation

In this subsection, we summarise existing nonparametric methods for evaluating surrogate markers. This framework was first proposed by Parast, Cai, and Tian (2024) and later extended in Hughes et al. (2025) to the high-dimensional setting. Since the discussion here focuses on the single-marker, single-trial context, we drop the subscripts j and m, and denote an arbitrary surrogate candidate by S.

2.2.1 Within-study surrogacy parameter

Parast et al. (2024) first define average treatment effects on the primary endpoint and candidate S in terms of U-statistics as follows:

$$U_Y = \mathbb{P}\big(Y^1_i > Y^0_l\big) + \tfrac{1}{2}\mathbb{P}\big(Y^1_i = Y^0_l\big), \qquad U_S = \mathbb{P}\big(S^1_i > S^0_l\big) + \tfrac{1}{2}\mathbb{P}\big(S^1_i = S^0_l\big),$$

where i and l index independent observations from the treated and control arms. These measures, defined on the interval $[0, 1]$, represent the average probabilities that a randomly sampled treated individual has a greater value than a randomly sampled control for the variables Y and S, respectively. We then define within-trial surrogacy as the difference between these effects:

$$\delta = U_Y - U_S,$$

which may be interpreted as the bias in estimating the average treatment effect on Y through observation of S alone. The closer this bias is to 0, the better S is as a

trial-level surrogate in this single-trial framework. These quantities may be estimated nonparametrically using rank-based methods, with associated confidence intervals (CIs) estimated using U-statistic theory.

Hughes et al. (2025) proposed that a valid surrogate in this framework may be determined by testing the hypotheses

$$H_0 : \delta \notin \big(-\varepsilon, \varepsilon\big), \qquad H_1 : \delta \in \big(-\varepsilon, \varepsilon\big),$$

where ε is some small, pre-determined constant, called the *equivalence bound*, defining the *equivalence margin* $\big(-\varepsilon, \varepsilon\big)$. This bound can either be chosen *a priori* or through the data-driven approach described in Parast et al. (2024), whereby the investigator specifies a target power to detect a treatment effect on Y via S.

Equivalence is checked with a two one-sided test (TOST) procedure (Schuirmann, 1987), which must reject both that $\delta \leq -\varepsilon$ and $\delta \geq \varepsilon$. Further detail on estimation, as well as a set of assumptions needed to avoid a *surrogate paradox* situation, are provided in the original publications (Hughes et al., 2025; Parast et al., 2024) and in the Supplementary Material of this article.

2.3 Overview of RISE-Meta

First, we provide a broad overview of our proposed method, RISE-Meta, which is an extension of the nonparametric approach to the meta-analytic setting. Let $\widehat{\delta}_{j,m}$ refer to the estimate of δ for marker j in study m. RISE-Meta consists of three sequential steps:

1. **Within-trial screening across candidates:** for each trial, estimate marker-specific surrogacy metrics $\widehat{\delta}_{j,m}$ and their sampling error, as described in Section 2.2.

2. **Meta-analysis of within-trial effects for each candidate:** pool $\widehat{\delta}_{j,m}$ across studies using random-effects meta-analysis. Perform equivalence testing on the pooled effects and adjust all p-values to control for the overall false discovery rate. Use adjusted p-values to define a set of promising markers.

3. **Construction and evaluation of a surrogate signature**: combine the set of identified promising markers into a single composite signature using a weighted sum, where markers with desirable surrogacy properties are given larger weights. Evaluate the trial-level surrogacy of the signature using hypothesis testing, prediction intervals, and concordance-based metrics.

A graphical depiction of these three steps is given in the Supplementary Material. In order to avoid problems associated with post-selection inference and overfitting, stages 2 and 3 should be applied to different data (Ball et al., 2020). This could be done either through within-study sample splitting or, when enough studies are available, retaining a small number of studies for the third stage only.

2.4 Random-effects meta-analysis model

In this section, we present the random-effects meta-analysis model used to pool effects $\widehat{\delta}_{j,m}$ across studies, as well as our recommended estimation procedure. We note that there is considerable debate about the merits of the random and fixed-effect approaches to meta-analysis in the literature (McKenzie and Veroniki, 2024). While throughout this section we describe the random-effects approach, this is easily reduced to a fixed-effect approach, and we include this as an option in the `R` package implementing these methods, `SurrogateRank`.

2.4.1 Model specification

We assume a hierarchical model for the estimates of the within-study effects for the jth marker. In the first level, we assume that the true effects follow

$$\delta_{m,j} \sim N(\mu_j, \tau_j^2),$$

with fixed mean μ_j and variance τ_j^2, which we refer to as the *between-study variance*. In the second level, we assume that the estimates of these values $\widehat{\delta}_{m,j}$ are normally distributed

around the true parameter values

$$\widehat{\delta}_{m,j} \sim N(\delta_{m,j}, \sigma^2_{m,j}),$$

where the variance $\sigma^2_{m,j}$ is referred to as the *within-study variance*. It is worth emphasising here that these normality assumptions are imposed only on study-level quantities. In contrast to classical meta-analytic approaches, we do not assume any parametric form (such as bivariate normality of Y and S) for the IPD within studies.

2.4.2 Estimation

The within-study variances $\sigma^2_{m,j}$ are estimated within each study, and the between-study variance τ^2_j is estimated through a numerical algorithm implementing restricted maximum likelihood. The pooled effect μ_j is classically estimated using a weighted sum

$$\widehat{\mu}_j = \frac{\sum_{m=1}^{M} \widehat{w}_{m,j}\widehat{\delta}_{m,j}}{\sum_{m=1}^{M} \widehat{w}_{m,j}}$$

where the weights are the inverse of the total variance within each study, i.e. $\widehat{w}_{m,j} = (\widehat{\sigma}^2_{m,j} + \widehat{\tau}^2_j)^{-1}$. The standard error, s_j, of this pooled estimate may be estimated through the Hartung-Knapp-Sidik-Jonkman (HKSJ) approach, which is more robust than the conventional estimator, especially when the number of studies M is small (Hartung and Knapp, 2001; IntHout, Ioannidis, and Borm, 2014). Then, $(1-\alpha)\times 100\%$ CIs can be constructed as $\widehat{\mu}_j \pm q_{M-1,1-\alpha/2} \times \widehat{s}_j$, where $q_{M-1,1-\alpha/2}$ is the $(1-\alpha/2)\times 100\%$ quantile of the t-distribution with $M-1$ degrees of freedom. Further details for estimation are given in the Supplementary Material.

2.5 Hypothesis testing

Next, we perform hypothesis tests on the pooled effect sizes $\widehat{\mu}_j$ to make operational conclusions about surrogate validity. Again, we adopt the framework for equivalence

testing using the two one-sided test (TOST) procedure (Schuirmann, 1987) briefly described in Section 2.2. Specifically, this requires testing two null hypothesis, which we term upper (U) and lower (L):

$$H_{0,j}^L : \mu_j \leq -\varepsilon, \quad H_{1,j}^L : \mu_j > -\varepsilon,$$

$$H_{0,j}^U : \mu_j \geq \varepsilon, \quad H_{1,j}^U : \mu_j < \varepsilon,$$

and concluding $\mu_j \in (-\varepsilon, \varepsilon)$ only if both tests are rejected. Using the HKSJ method, p-values for the lower and upper hypotheses can be estimated as $\widehat{p}_j^L = 1 - F_{M-1}\left(\frac{\widehat{\mu}_j + \varepsilon}{\widehat{s}_j}\right)$ and $\widehat{p}_j^U = F_{M-1}\left(\frac{\widehat{\mu}_j - \varepsilon}{\widehat{s}_j}\right)$, respectively, where $F_{M-1}(.)$ denotes the cumulative distribution function of the t-distribution with $M-1$ degrees of freedom. The p-value for the TOST is defined as $\widehat{p}_j = \max(\widehat{p}_j^L, \widehat{p}_j^U)$, where a significant p-value at the α-level implies that the $(1-2\alpha) \times 100\%$ CI of $\widehat{\mu}_j$ lies within $(-\varepsilon, \varepsilon)$.

Finally, due to the large number of hypothesis tests performed, we control the false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), and refer to the resulting p-values as the adjusted p-values, denoted $\widehat{p}_j^*$.

Additional evaluation metrics. Although CIs and p-values provide useful metrics for operational decision making in the context of high-dimensional surrogate candidates, it may be unwise to uniquely rely on these criteria. In particular, prior work has shown that CIs may be misleading in the presence of substantial between-study variability (IntHout et al., 2016). This has led to the advocacy of prediction intervals (PIs), which represent the likely range of effect estimates in new, similar trials. We argue that PIs are highly relevant in the context of surrogate markers, since the goal is to use the surrogate to predict the treatment effect on the primary endpoint in a new trial. Using the HKSJ approach, a $(1-\alpha) \times 100\%$ PI can be estimated as $\widehat{\mu}_j \pm q_{M-1,1-\alpha/2} \times \sqrt{\widehat{s}_j^2 + \widehat{\tau}_j^2}$.

In addition, we use the concordance correlation coefficient (CCC) (Lin, 1989) between the treatment effects on Y, $\widehat{\boldsymbol{U}}_{\boldsymbol{Y}} = (\widehat{U}_{Y,1}, \ldots, \widehat{U}_{Y,M})^T$, and the treatment effects on $\boldsymbol{S}$, $\widehat{\boldsymbol{U}}_{\boldsymbol{S}} = (\widehat{U}_{S,1}, \ldots, \widehat{U}_{S,M})^T$, as a further validation metric for trial-level surrogacy. Like other correlation coefficients, this sits within the interval $[-1, 1]$, with similar interpretation.

However, this metric also penalises differences from the identity line, allowing for a more nuanced evaluation in the context of our treatment effects, which are on a standardised scale. Further details are given in the Supplementary Material.

2.6 Construction of a composite signature

The final stage of the method involves combining the promising surrogate candidates together to form a composite signature. First, we define a set of indices of the significant markers from the previous tests $\Gamma = \{j : \widehat{p}_j^* < \alpha\}$. We combine the markers defined by this set together using a weighted sum

$$\boldsymbol{\gamma}_\Gamma = \sum_{j \in \Gamma} \widehat{\lambda}_j \, \bar{\boldsymbol{S}}_j,$$

where $\bar{\boldsymbol{S}}_j$ is the vector of the jth surrogate standardised to have 0 mean and 1 variance, and $\widehat{\lambda}_j$ is the estimated *signature weight* for the jth marker, which we describe below.

The signature weights should be chosen in order to increase the prominence of markers with better surrogacy properties. We consider two such properties, which combine together to make the final signature weight: the strength of the surrogate, and its consistency across trials. The first component depends on the pooled mean of the jth candidate, and is formed as $\widehat{a}_j = \frac{\varepsilon - |\widehat{\mu}_j|}{\varepsilon}$, which gives maximum weight when $\widehat{\mu}_j = 0$ and minimum weight when $\widehat{\mu}_j$ is on the boundary ε. This component adapts the idea behind the weighting scheme presented in Hughes et al. (2025) to be more stable for small values of $\widehat{\mu}_j$. The second component concerns the precision of estimation within and between-trials, determined by the sum of the total variances for each trial $\widehat{b}_j = \sum_{m=1}^{M} (\widehat{\sigma}_{m,j}^2 + \widehat{\tau}_j^2)^{-1}$.

We then construct the final signature weights as a standardised combination of the two components:

$$\widehat{\lambda}_j = \frac{\widehat{a}_j \widehat{b}_j}{\max_{j \in \Gamma} \{\widehat{a}_j \widehat{b}_j\}}.$$

3 Simulation study

3.1 Simulation setup

We first use a parametric data generating process to evaluate the meta-analysis equivalence testing procedure proposed in Section 2. Our primary objectives are to assess the calibration of the test, via the empirical false positive rate (FPR), and its empirical power. The calibration assessment uses only invalid surrogates, while the power assessment uses only valid surrogates.

Valid surrogates in our framework are defined as those for which $\mu_j \in (-\varepsilon, \varepsilon)$, and invalid surrogates otherwise. As the null hypothesis is composite, its definition corresponds to a region of the parameter space. To assess calibration, it is therefore natural to consider the *least favourable configuration* (LFC), in which invalid surrogates lie on the boundary of the null, i.e. $\mu_j \in \{-\varepsilon, \varepsilon\}$. If the empirical FPR is controlled at level α in this setting, it will also be controlled for more favourable configurations.

We consider two main scenarios corresponding to the evaluation of calibration and power. Throughout, we generate results across $J = 100,000$ independent marker draws, with nominal FPR $\alpha = 0.05$ and equivalence bound $\varepsilon = 0.1$. In Scenario 1, we generate only invalid surrogates under the LFC to assess the empirical FPR. Within this scenario, we examine results at various values of M, the number of studies, ($M = 3, 10, 25$), and n_m, the sample size within each study ($n_m = 10, 50, 250$). In addition, we examine results at varying levels of between-study and within-study variability, with fixed $M = n_m = 20$. To balance effects on either side of the null boundary, we set $\mu_j = -\varepsilon$ and $\mu_j = \varepsilon$ with equal probability. In Scenario 2, we generate only valid surrogates to assess power, with μ_j allowed to vary across the valid region as $\mu_j \sim \mathcal{U}(-\varepsilon, \varepsilon)$. Within this scenario, we examine results at various values of M ($M = 3, 10, 25$), and varying levels of between-study and within-study variability, with fixed $n_m = 50$.

For each marker $j = 1, \ldots, J$, we first draw the true mean μ_j based on the scenarios below, which depend on if we want a valid or invalid surrogate. We then simulate marker-specific between-study heterogeneity as $\tau_j^2 \sim \mathcal{U}(0, u_{\tau^2,max})$. Conditional on these

parameters, the true study-specific effects are generated as

$$\delta_{m,j} \sim \mathcal{N}(\mu_j, \tau_j^2), \qquad m = 1, \ldots, M.$$

At the second stage, we define the within-study variance as $\sigma_{m,j}^2 = \nu_j / n_m$, where $\nu_j \sim \mathcal{U}(0, u_{\nu,\max})$ introduces marker-specific variability. The observed effect estimates are then generated as

$$\widehat{\delta}_{m,j} \sim \mathcal{N}(\delta_{m,j}, \sigma_{m,j}^2).$$

3.2 Results

3.2.1 Scenario 1: calibration

In scenario 1, where only invalid surrogates are generated, we investigated the calibration of the statistical test. The results are presented as calibration plots showing empirical against nominal FPR.

Figure 1A shows calibration plots for number of studies $M = 3$, 10, 25, within-study sample size $n_m = 10$, 50, 250, and fixed maximum heterogeneity parameters $u_{\tau^2,max} = \varepsilon/10$, $u_{\nu,\max} = 100\varepsilon$. Figure 1B shows calibration plots for maximum between-study variability $u_{\tau^2,max} = \varepsilon/10$, ε, 10ε and max within-study variability $u_{\nu,max}/n_m = \varepsilon/10$, ε, 10ε at fixed sample parameters $M = n_m = 10$.

In all cases, the empirical FPR was bounded by the claimed FPR, providing confidence about the test's statistical validity even when estimation is based off a small number of samples or studies, or in presence of extreme heterogeneity. As expected, calibration improves with increasing sample sizes, number of studies, and with decreasing heterogeneity, with near perfect calibration being achieved for the most favourable combinations of these values. Overall, these results indicate a conservative and valid test procedure, which is desirable in the context of surrogate marker evaluation, where an inflated false positive rate could have serious consequences.

3.2.2 Scenario 2: power

In scenario 2, where only valid surrogates are generated, we investigated the power of the statistical test across settings. Figure 2 shows the empirical power as a function of maximum between- and within-study variability $u_{\tau^2,max} = u_{\nu,max} = \{\varepsilon/100, \varepsilon/10, \varepsilon, 10\varepsilon, 100\varepsilon\}$ for number of studies $M = 3, 10, 25$ at a fixed sample size of $n_m = 50$. As expected, the power is greatest when the number of studies is high and the heterogeneity is low. The heterogeneity parameters have a greater impact on the empirical power than the number of studies, and in the cases of extremely high heterogeneity, the power to detect true positives reduces to 0.

3.3 Supplementary simulations

Although the parametric simulation allows fine control of parameters, it is also important to examine validity when the model may not hold. This can be done with nonparametric simulation, generating null datasets by random permutation of a real dataset while preserving its realistic properties. The results of this nonparametric simulation study, as well as further results from the parametric simulation study, are presented in the Supplementary Material. These results further support the validity of the test procedure and provide justification for our recommended analysis choices.

4 Data applications

4.1 High-dimensional data example

We applied RISE-Meta to public multi-trial data on trivalent inactivated influenza vaccine (TIV) to evaluate high-dimensional trial-level surrogates of the downstream immune response. These data were accessed through the ImmuneSpace platform (`immunespace.org`) as part of the Immune Signatures 2 (IS2) project (Diray-Arce et al., 2022). TIV is an inactivated vaccine designed to induce antibodies against three circulating influenza strains. The primary endpoint is the antibody response to TIV, defined as the mean of

the strain-specific antibody responses across the three strains targeted by each vaccine formulation four weeks after vaccination. Strain-specific antibody titres were measured using either neutralising antibody (nAb) or haemagglutination inhibition (HAI) assay, with nAb prioritised.

Initially, the surrogate candidates consisted of the expression of 10,086 genes sequenced from either whole blood or peripheral blood mononuclear cell samples measured the day after vaccination. Rather than working at the gene-level, we aggregated these measurements to the geneset-level, which are predefined groups of functionally related genes. This transformation provides an initial dimension reduction, thereby reducing the multiple testing burden, and improves interpretability of the results. We used the Blood Transcriptional Modules (BTMs) (Li et al., 2013) as predefined genesets and summarised each module by the mean expression of its constituent genes. After filtering for modules with descriptive titles, we obtained $J = 258$ input features. Further details, as well as an evaluation of the impact of these choices, are provided in the Supplementary Material.

Each study involved paired data, with the primary endpoint and candidate surrogates measured before and after vaccination for each subject. Antibody responses were measured at days 0 and 28, where we allowed ± 7 days for the post-vaccination measurement to account for minor differences in study design. Gene expression was measured at days 0 and 1. After filtering studies with fewer than 5 participants and individuals missing pre- or post-vaccination measurements, the data contained paired measurements from 334 individuals across 4 studies. We used 66% of the data within each study for the marker-level meta-analysis, and reserved the remaining data for evaluation of the composite signature. The unique study identifiers in ImmPort (`immport.org`) (Bhattacharya et al., 2018) and IS2, and their corresponding numbers of individuals, are SDY56 ($n = 58$), SDY80 ($n = 56$), SDY180 ($n = 12$), and SDY1276 ($n = 208$). A full description of the IS2 TIV datasets and the data preprocessing steps for this analysis are given in the Supplementary Material. The significance level was set to $\alpha = 0.05$, and the Benjamini-Hochberg (BH) procedure was applied to adjust the p-values for multiple testing. The value of ε was chosen using the power approach described in Parast et al. (2024), taking

the mean of ε across studies such that within each study there is 80% power to detect a treatment effect on the antibodies based on observation of the surrogate candidates. This gave $\varepsilon \approx 0.205$ in the screening stage and $\varepsilon \approx 0.148$ for the evaluation of the composite signature.

The results of the marker-level meta-analysis screening stage are shown in Figure 4: seven markers remained significant after multiplicity correction. Their point estimates are positive and close to 0, suggesting a small underestimation of the treatment effect. Many of the top pathways are related to interferon/antiviral signalling, dendritic cell/antigen presentation, and monocyte signatures, giving these results a coherent immunological interpretation. We return to this point in the Discussion. The results of the evaluation of the composite 7-marker signature on independent data are shown in Figure 5. The estimates of δ_m are close to 0 across trials, with minor differences in point estimates and standard errors. The pooled effect is close to 0, and its 90% CI lies within the equivalence margin, with $p = 0.004$ from the TOST. The PI is also small, reflecting minimal between-trial variability, with $\tau^2 \approx 6 \times 10^{-11}$. In addition, the CCC is 0.94, indicating strong concordance between treatment effects on the primary endpoint and on this 7-marker signature. Together, these results illustrate that RISE-Meta is capable of deriving strong and interpretable trial-level surrogates in high-dimensional experiments.

4.2 Low-dimensional data example

In addition, we applied RISE-Meta to two low-dimensional datasets to compare it with a reference method for meta-analytic surrogate evaluation. This is the only setting in which a direct comparison with classical approaches is possible, as these methods cannot be extended to high-dimensional data without substantial modification, which RISE-Meta is designed to address. Agreement in this setting would support the use of RISE-Meta in classical data settings where investigators desire to avoid strong distributional assumptions on the IPD.

The reference approach, introduced in Buyse et al. (2000) and described in full across multiple sources (Burzykowski, Molenberghs, and Buyse, 2005; Molenberghs et al., 2002;

Tibaldi et al., 2003), is referred to as the *bivariate joint modelling* approach. These reference approaches and the public datasets we use as examples in this section are accessible from the `Surrogate` package in `R` (Van Der Elst et al., 2014). Throughout, due to numerical convergence issues, we were unable to fit full or reduced bivariate mixed-effect models to the data, and opted instead to fit full bivariate fixed-effect models through the `BifixedContCont()` function. In the Supplementary Material, we give a more detailed description of this procedure. Since the low-dimensional datasets consisted of a single surrogate candidate each, the application of RISE-Meta only involves steps 1 and 2 described in Section 2.3. For TOST equivalence testing, we used equivalence bound $\varepsilon = 0.1$ and $\alpha = 0.05$.

The first dataset, `ARMD`, comes from a clinical trial of patients with age-related macular degeneration treated with interferon-α or placebo. The primary endpoint is the change in visual acuity (CVA) at 12 months, and the candidate surrogate is CVA at 6 months. Although these data come from a single trial, we follow previous applications in the literature and use the treatment center as the unit of analysis (Molenberghs et al., 2002). Further discussion on this choice is provided in the Supplementary Material and in Abrahantes et al. (2004). After filtering to centers with at least 5 total patients and at least 2 patients in each treatment group, we retained 109 patients across 15 centers. Filtering the smallest units was necessary to ensure non-degenerate estimation of the within-trial variance σ_m^2 when using RISE-Meta.

The second dataset, `Ovarian`, comes from 4 trials in patients with advanced ovarian cancer comparing cyclophosphamide plus adriamycin plus cisplatin with cyclophosphamide plus cisplatin. The primary endpoint is the log of the overall survival time (OS), and the candidate surrogate is the log of the progression-free survival time (PFS). Again, we take the treatment center as the unit of analysis. After applying the same filtering criteria, we retained 1,168 patients across 42 centers.

To compare the methods fairly, we report both the R^2_{trial}, the CCC, and their 95% CIs (estimation detail in Supplementary Material). Figure 3 shows a graphical summary of these results, indicating strong agreement between the two methods for both datasets.

In particular, both methods conclude a moderately strong surrogate in the `ARMD` data and a strong surrogate in the `Ovarian` data in the context of their respective treatment comparisons. In addition, RISE-Meta gave significant p-values (0.0067 for `ARMD` and 7×10^{-24} for `Ovarian`), indicating evidence for trial-level surrogacy. Overall, this example suggests RISE-Meta agrees with the reference method in low-dimensional settings where the normality assumption is more likely to hold.

5 Discussion

In this article, we introduced RISE-Meta, a multi-stage framework for evaluating high-dimensional trial-level surrogate candidates across studies. The method offers several advantages over classical meta-analytic approaches in complex data settings. First, the nonparametric first stage avoids restrictive distributional assumptions on the individual participant data, enhancing robustness and applicability, particularly in small samples. The resulting within-study surrogacy measures are defined on a standardised scale and are invariant to monotone transformations, depending only on the relative ordering of observations within each study. This property is especially advantageous when integrating heterogeneous data sources measured on different scales (e.g., psychological assessments, multiplex assays, gene expression, or ELISPOT data). These standardised effect measures are directly comparable across markers and studies, providing interpretable and stable inputs for downstream meta-analysis. The evaluation stage further extends this framework by combining multiple markers into a composite signature, thereby strengthening surrogacy and reducing reliance on any single candidate in isolation. Finally, the use of an equivalence testing framework enables formal assessment of surrogacy against a clinically meaningful margin. While such approaches are well established in regulatory contexts (e.g., bioequivalence studies), they remain underutilised in surrogate endpoint evaluation (Saraf, Mathew, and Roy, 2014).

We illustrated RISE-Meta on public vaccination data to evaluate gene expression biomarkers as trial-level surrogates for the clinically relevant downstream antibody re-

sponse. Here, we identified a set of 7 genesets which acted as strong trial-level surrogates, both individually and in combination on unseen data. Ranking and visualising the top candidates from the marker-level meta-analysis (as in Figure 5) allows further biological interpretation. The top of this list is enriched for interferon/antiviral (M127, M68, M75, M111.1), dendritic cell/antigen presentation (M165, M67, M95.1, S11, S5), and monocyte (M4.15, M11.0) modules, which refer to different components of the innate immune response which are expected to be observed in the first hours and days following vaccination (Hagan et al., 2022). In addition, the interferon/antiviral interpretation reinforces the conclusions found in Hughes et al. (2025), where the RISE method (i.e. the first stage of RISE-Meta) was applied to a subset of a single trial (SDY1276). This application thus illustrates the proposed value of RISE-Meta for deriving strong, interpretable TLS in complex settings, where classical approaches cannot be applied without significant modification.

RISE-Meta also has notable limitations. Firstly, it is known that the TOST procedure is conservative and underpowered (Meyners, 2012), as we confirm in our simulation studies. Despite this, alternative procedures to the TOST are more conceptually complex and do not offer major power gains when applied to well-designed studies (Meyners, 2012). Next, the choice of the equivalence margin ε has long been discussed in equivalence testing (Walker and Nowacki, 2010). A first strategy is simply to decide, based on the contextual costs and benefits of replacing the primary endpoint with a surrogate marker, a maximum bias one would tolerate when estimating the treatment effect on the former. Another, as exemplified in the application, is to use a data-driven approach based on the power within each study to detect a treatment effect on the primary endpoint based on the surrogate (Parast et al., 2024). Alternatively, one may remove the dependence on p-values and equivalence margins entirely and consider the *least equivalence allowable difference* (LEAD), which gives the smallest equivalence margin such that equivalence would be declared at a significance level α (Meyners, 2007).

Several avenues remain open for the future development of RISE-Meta. First, Bayesian meta-analysis could be used to address the power limitation, especially when the number

of studies is small (Röver, 2020). Next, RISE-Meta uses a univariate screening procedure, but markers may belong to complex networks that only demonstrate surrogacy when considered simultaneously. A natural multivariate analogue of rank-based estimation and testing uses optimal transport, which could be adopted to extend the first stage of RISE-Meta (Huang and Sen, 2025). Finally, although RISE-Meta is broadly applicable across continuous data, its utility could be widened by extending its first stage to other common data types, such as binary and survival data.

Acknowledgements

This work is part of A.H.'s PhD thesis at the University of Bordeaux, co-supervised by B.P.H. and R.T., and supported by the University of Bordeaux's Digital Public Health Graduate School, funded by France's PIA 3 scheme (Investments for the Future – Project reference: 17-EURE-0019) through the Agence Nationale de la Recherche and University of Bordeaux's France 2030 program. In addition, this work was supported by the Programme et Equipement Prioritaire de Recherche Santé Numérique (PEPR SN) project SMATCH (ref: 22-PESN-0003), the MUSICC (Mucosal Immunity in Human Coronavirus Challenge) project funded by the Coalition for Epidemic Preparedness Innovations (CEPI) and co-funded by the European Union's Horizon Europe Programme, and in part by NIDDK grant R01DK118354 (Parast). We thank the Human Immunology Project Consortium for providing the ImmPort and ImmuneSpace platforms, as well as the original investigators and participants of the studies SDY56, SDY80, SDY180 and SDY1276. We disclose that generative AI tools were used to refine the phrasing of certain sections of this manuscript, and aid with the coding of the methods, but not to generate content.

References

Abrahantes, José Cortiñas, Geert Molenberghs, Tomasz Burzykowski, Ziv Shkedy, Ariel Alonso Abad, and Didier Renard (Oct. 2004). "Choice of units of analysis and modeling

strategies in multilevel hierarchical models". In: *Computational Statistics and Data Analysis* 47.3, pp. 537–563. ISSN: 0167-9473. DOI: 10.1016/j.csda.2003.12.003.

Agniel, Denis, Boris P Hejblum, Rodolphe Thiébaut, and Layla Parast (July 2022). "Doubly robust evaluation of high-dimensional surrogate markers". In: *Biostatistics* 24.4, pp. 985–999. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxac020.

Alonso, Ariel, Helena Geys, Geert Molenberghs, Michael G. Kenward, and Tony Vangeneugden (Dec. 2004). "Validation of Surrogate Markers in Multiple Randomized Clinical Trials with Repeated Measurements: Canonical Correlation Approach". In: *Biometrics* 60.4, pp. 845–853. ISSN: 1541-0420. DOI: 10.1111/j.0006-341x.2004.00239.x.

Alonso, Ariel, Geert Molenberghs, Helena Geys, Marc Buyse, and Tony Vangeneugden (2005). "A unifying approach for surrogate marker validation based on Prentice's criteria". In: *Statistics in Medicine* 25.2, pp. 205–221. ISSN: 1097-0258. DOI: 10.1002/sim.2315.

Ball, Tali M., Lindsay M. Squeglia, Susan F. Tapert, and Martin P. Paulus (Mar. 2020). "Double Dipping in Machine Learning: Problems and Solutions". In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 5.3, pp. 261–263. ISSN: 2451-9022. DOI: 10.1016/j.bpsc.2019.09.003.

Benjamini, Yoav and Yosef Hochberg (Jan. 1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1, pp. 289–300. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bhattacharya, Sanchita, Patrick Dunn, Cristel G. Thomas, Barry Smith, Henry Schaefer, Jieming Chen, et al. (Feb. 2018). "ImmPort, toward repurposing of open access immunological assay data for translational and clinical research". In: *Scientific Data* 5.1. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.15.

Burzykowski, Tomasz, Geert Molenberghs, and Marc Buyse (2005). *The Evaluation of Surrogate Endpoints*. Ed. by Tomasz Burzykowski, Geert Molenberghs, and Marc Buyse. Springer New York. ISBN: 9780387270807. DOI: 10.1007/b138566.

Buyse, Marc, Geert Molenberghs, Tomasz Burzykowski, Didier Renard, and Helena Geys (Apr. 2000). "Statistical Validation of Surrogate Endpoints: Problems and Proposals". In: *Drug Information Journal* 34.2, pp. 447–454. ISSN: 2164-9200. DOI: 10.1177/009286150003400213.

Choi, Jung Kyoon, Ungsik Yu, Sangsoo Kim, and Ook Joon Yoo (July 2003). "Combining multiple microarray studies and modeling interstudy variation". In: *Bioinformatics* 19.suppl, pp. i84–i90. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg1010.

Daniels, Michael J. and Michael D. Hughes (Sept. 1997). "Meta-analysis for the evaluation of potential surrogate markers". In: *Statistics in Medicine* 16.17, pp. 1965–1982. ISSN: 1097-0258. DOI: 10.1002/(sici)1097-0258(19970915)16:17<1965::aid-sim630>3.0.co;2-m.

Diray-Arce, Joann, Helen E. R. Miller, Evan Henrich, Bram Gerritsen, Matthew P. Mulè, Slim Fourati, et al. (Oct. 2022). "The Immune Signatures data resource, a compendium of systems vaccinology datasets". In: *Scientific Data* 9.1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01714-7.

Edgar, R. (Jan. 2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Research* 30.1, pp. 207–210. ISSN: 1362-4962. DOI: 10.1093/nar/30.1.207.

Hagan, Thomas, Bram Gerritsen, Lewis E. Tomalin, Slim Fourati, Matthew P. Mulè, Daniel G. Chawla, et al. (Oct. 2022). "Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses". In: *Nature Immunology* 23.12, pp. 1788–1798. ISSN: 1529-2916. DOI: 10.1038/s41590-022-01328-6.

Hartung, Joachim and Guido Knapp (Dec. 2001). "A refined method for the meta-analysis of controlled clinical trials with binary outcome". In: *Statistics in Medicine* 20.24, pp. 3875–3889. ISSN: 1097-0258. DOI: 10.1002/sim.1009.

Huang, Zhen and Bodhisattva Sen (2025). *Multivariate Distribution-Free Nonparametric Testing: Generalizing Wilcoxon's Tests via Optimal Transport*. DOI: 10.48550/ARXIV.2503.12236.

Hughes, Arthur, Layla Parast, Rodolphe Thiébaut, and Boris P. Hejblum (Sept. 2025). "RISE: Two-Stage Rank-Based Identification of High-Dimensional Surrogate Markers Applied to Vaccinology". In: *Statistics in Medicine* 44.20–22. ISSN: 1097-0258. DOI: 10.1002/sim.70241.

IntHout, Joanna, John P A Ioannidis, Maroeska M Rovers, and Jelle J Goeman (July 2016). "Plea for routinely presenting prediction intervals in meta-analysis". In: *BMJ Open* 6.7, e010247. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2015-010247.

IntHout, Joanna, John PA Ioannidis, and George F Borm (Feb. 2014). "The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method". In: *BMC Medical Research Methodology* 14.1. ISSN: 1471-2288. DOI: 10.1186/1471-2288-14-25.

Jonker, Emile F. F., Leonardus G. Visser, and Anna H. Roukens (Aug. 2013). "Advances and controversies in yellow fever vaccination". In: *Therapeutic Advances in Vaccines* 1.4, pp. 144–152. ISSN: 2051-0136. DOI: 10.1177/2051013613498954.

Korn, E.L., M.C. Sachs, and L.M. McShane (Jan. 2016). "Statistical controversies in clinical research: assessing pathologic complete response as a trial-level surrogate end point for early-stage breast cancer". In: *Annals of Oncology* 27.1, pp. 10–15. ISSN: 0923-7534. DOI: 10.1093/annonc/mdv507.

Li, Shuzhao, Nadine Rouphael, Sai Duraisingham, Sandra Romero-Steiner, Scott Presnell, Carl Davis, et al. (Dec. 2013). "Molecular signatures of antibody responses derived from a systems biology study of five human vaccines". In: *Nature Immunology* 15.2, pp. 195–204. ISSN: 1529-2916. DOI: 10.1038/ni.2789.

Lin, Lawrence I-Kuei (Mar. 1989). "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* 45.1, p. 255. ISSN: 0006-341X. DOI: 10.2307/2532051.

Liyanage, Jayamini C, Luke Prendergast, Robert Staudte, and Alysha M De Livera (July 2024). "MetaHD: a multivariate meta-analysis model for metabolomics data". In: *Bioinformatics* 40.7. Ed. by Jonathan Wren. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae470.

McKenzie, Joanne E. and Areti Angeliki Veroniki (Oct. 2024). "A brief note on the random-effects meta-analysis model and its relationship to other models". In: *Journal of Clinical Epidemiology* 174, p. 111492. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2024.111492.

Meyners, Michael (Apr. 2007). "Least equivalent allowable differences in equivalence testing". In: *Food Quality and Preference* 18.3, pp. 541–547. ISSN: 0950-3293. DOI: 10.1016/j.foodqual.2006.07.005.

— (Dec. 2012). "Equivalence tests – A review". In: *Food Quality and Preference* 26.2, pp. 231–245. ISSN: 0950-3293. DOI: 10.1016/j.foodqual.2012.05.003.

Molenberghs, Geert, Marc Buyse, Helena Geys, Didier Renard, Tomasz Burzykowski, and Ariel Alonso (Dec. 2002). "Statistical challenges in the evaluation of surrogate endpoints in randomized trials". In: *Controlled Clinical Trials* 23.6, pp. 607–625. ISSN: 0197-2456. DOI: 10.1016/s0197-2456(02)00236-2.

Parast, Layla, Tianxi Cai, and Lu Tian (Jan. 2024). "A rank-based approach to evaluate a surrogate marker in a small sample setting". In: *Biometrics* 80.1. ISSN: 1541-0420. DOI: 10.1093/biomtc/ujad035.

Prachi, Prachi, Massimiliano Biagini, and Fabio Bagnoli (Oct. 2012). "Vaccinology Is Turning into an Omics-Based Science". In: *Drug Development Research* 73.8, pp. 547–558. ISSN: 1098-2299. DOI: 10.1002/ddr.21048.

Röver, Christian (2020). "Bayesian Random-Effects Meta-Analysis Using the bayesmeta R Package". In: *Journal of Statistical Software* 93.6. ISSN: 1548-7660. DOI: 10.18637/jss.v093.i06.

Saraf, Sanatan, Thomas Mathew, and Anindya Roy (Nov. 2014). "Statistical Validation of Surrogate Endpoints: Another Look at the Prentice Criterion and Other Criteria". In: *Journal of Biopharmaceutical Statistics* 25.6, pp. 1234–1246. ISSN: 1520-5711. DOI: 10.1080/10543406.2014.971174.

Schuirmann, Donald J. (Dec. 1987). "A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability". In:

Journal of Pharmacokinetics and Biopharmaceutics 15.6, pp. 657–680. ISSN: 0090-466X. DOI: 10.1007/bf01068419.

Tibaldi, Fabián, José Cortiñas Abrahantes, Geert Molenberghs, Didier Renard, Tomasz Burzykowski, Marc Buyse, et al. (Sept. 2003). "Simplified hierarchical linear models for the evaluation of surrogate endpoints". In: *Journal of Statistical Computation and Simulation* 73.9, pp. 643–658. ISSN: 1563-5163. DOI: 10.1080/0094965031000062177.

Tseng, George C., Debashis Ghosh, and Eleanor Feingold (Jan. 2012). "Comprehensive literature review and statistical considerations for microarray meta-analysis". In: *Nucleic Acids Research* 40.9, pp. 3785–3799. ISSN: 0305-1048. DOI: 10.1093/nar/gkr1265.

Van Der Elst, Wim, Florian Stijven, Fenny Ong, Dries De Witte, Gokce Deliorman, Paul Meyvisch, et al. (Mar. 2014). *Surrogate: Evaluation of Surrogate Endpoints in Clinical Trials*. DOI: 10.32614/cran.package.surrogate.

Walker, Esteban and Amy S. Nowacki (Sept. 2010). "Understanding Equivalence and Noninferiority Testing". In: *Journal of General Internal Medicine* 26.2, pp. 192–196. ISSN: 1525-1497. DOI: 10.1007/s11606-010-1513-8.

Weinstein, John N, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, et al. (Sept. 2013). "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nature Genetics* 45.10, pp. 1113–1120. ISSN: 1546-1718. DOI: 10.1038/ng.2764.

Zhou, Ruixuan Rachel, Sihai Dave Zhao, and Layla Parast (Feb. 2022). "Estimation of the proportion of treatment effect explained by a high-dimensional surrogate". In: *Statistics in Medicine* 41.12, pp. 2227–2246. ISSN: 1097-0258. DOI: 10.1002/sim.9352.

Supplementary Material

Web appendices and figures are provided in the Supplementary Material available on the publisher's website. All methods presented in this article are available from the `SurrogateRank` `R` package, available from `github.com/laylaparast/SurrogateRank`. A

tutorial for RISE-Meta based on simulated data can be found at `www.laylaparast.com/surrogaterank`. The data that support the findings of this study are openly available from ImmPort and ImmuneSpace, with study reference numbers SDY56, SDY80, SDY180 and SDY1276. Instructions to download these data and code to reproduce all results in this manuscript are available at `github.com/arthurhughes27/RISE-meta` or on Zenodo at `10.5281/zenodo.20039496`.

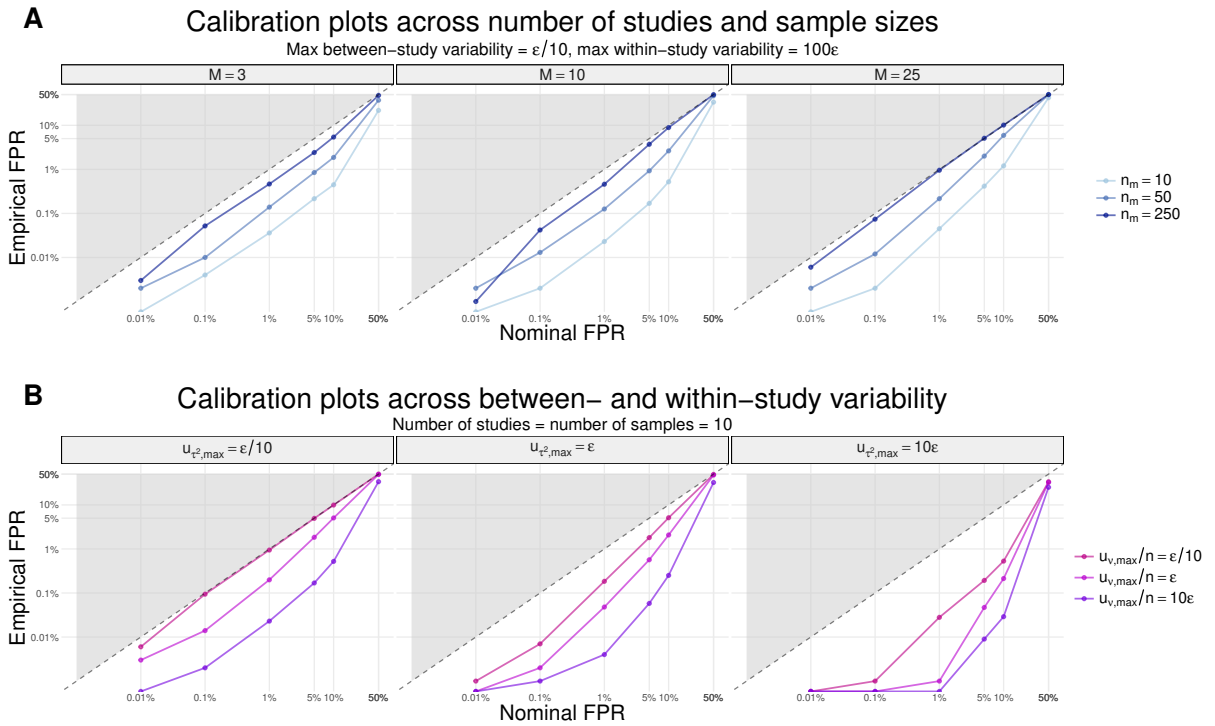


Figure 1: **Calibration plots comparing nominal and empirical false positive rates across A) samples and B) variability settings**. Each panel shows the nominal FPR α versus the empirical FPR, computed as the mean over 100,000 independent simulations for each configuration. The dashed diagonal represents perfect calibration. The shaded upper triangle delimits anti-conservative behaviour and the lower triangle conservative behaviour. The equivalence bound was set to $\varepsilon = 0.1$. In **A**), the number of studies M (facets) and within-study sample size n_m (lines) were varied, and the heterogeneity parameters were fixed at $u_{\tau^2,max} = \varepsilon/10$ and $u_{\nu,\max} = 100\varepsilon$, whereas in **B**) the between-study variability $u_{\tau^2,max}$ (facets) and within-study variability $u_{\nu,\max}/n$ (lines) were varied, and the sample parameters were fixed at $M = n_m = 10$. **Alt text:** Plots showing the claimed false positive rate against the observed false positive rate in simulations across various settings.

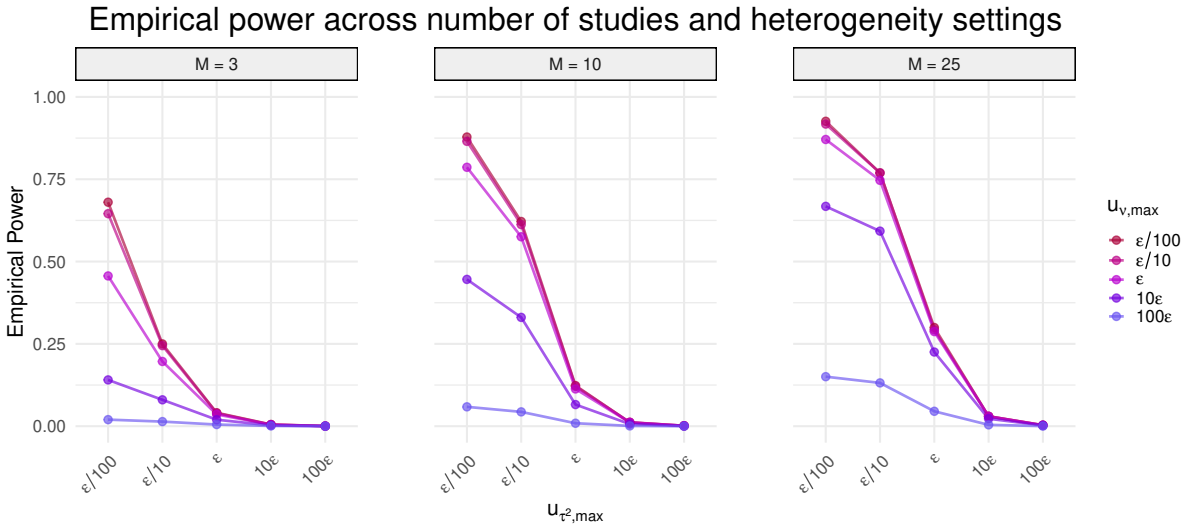


Figure 2: **Empirical power across different heterogeneity settings and number of studies**. The parameters $u_{\tau^2,max}$ and $u_{\nu,\max}$ control the maximum between-study and within-study heterogeneity, respectively. Both are expressed relative to the equivalence bound $\varepsilon = 0.1$. The within-study sample size was fixed at 50. **Alt text:** Graphs showing the observed power against the between-study heterogeneity.

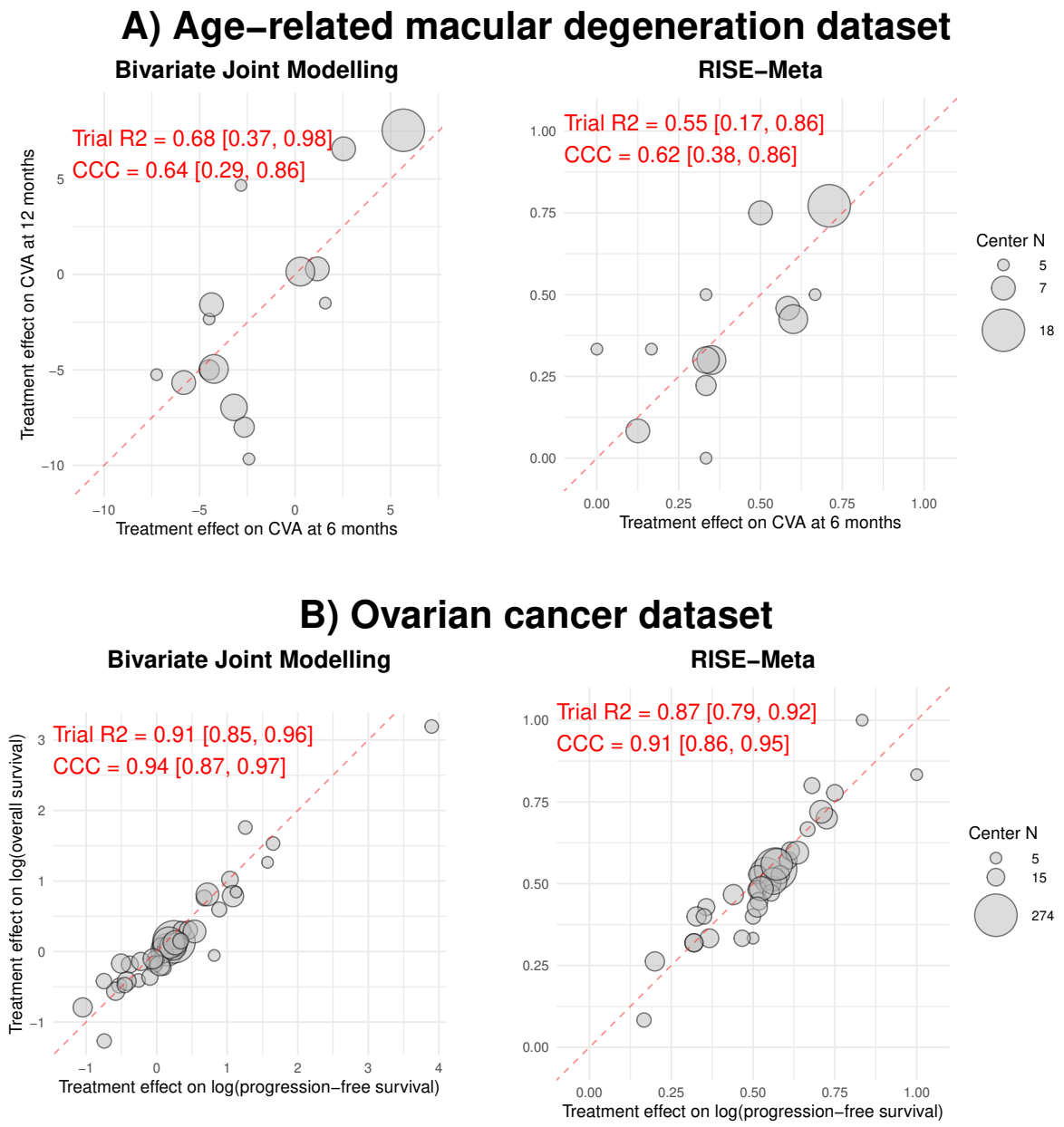


Figure 3: **Comparison of reference bivariate joint modelling approach to RISE-Meta applied to two low-dimensional datasets.** Panel **A)** gives the method comparison when applied to the `ARMD` dataset, whereas **B)** refers to the `Ovarian` dataset. Evaluation metrics and their 95% CIs for each method and dataset are given in the top left corner. Overall, this demonstrates qualitative agreement between the two methods when applied to low-dimensional data. **Alt text:** Graphical comparison of the reference meta-analysis approach against RISE-Meta in two low-dimensional examples.

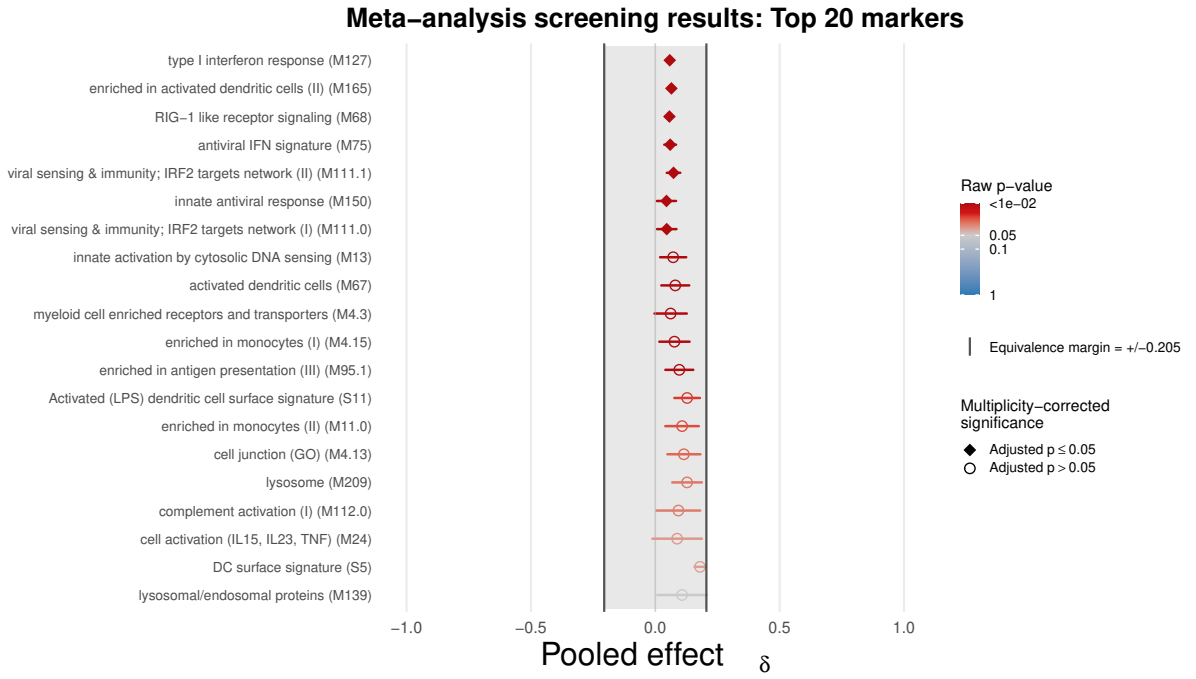


Figure 4: **Results from the marker-level meta-analysis.** This shows the pooled effect estimates for the top 20 markers by raw p-value. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.205$. The colour corresponds to the raw p-value, where shades of red indicate points whose 90% CIs are contained entirely within these bounds. The shape of the point denotes the significance decision after adjusting for test multiplicity, which includes the top 7 markers. **Alt text:** Marker-level results of meta-analysis.

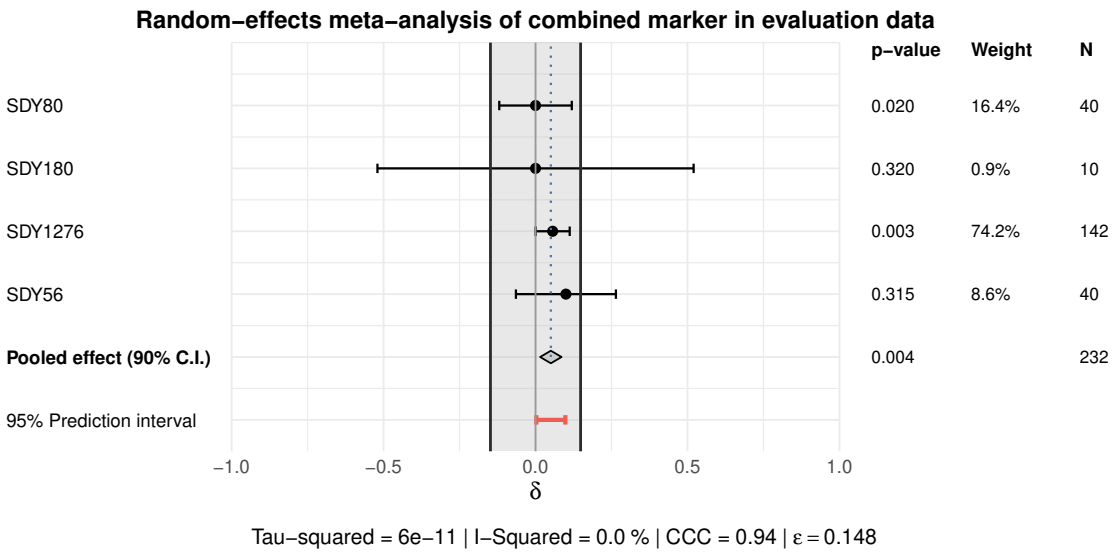


Figure 5: **Results from the evaluation of the 7-marker composite signature.** This shows the effect estimates δ for each of the 4 studies, and the pooled effect estimate. The within-study p-values, meta-analysis weights, and number of observations N are given on the right hand side. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.148$. The pooled effect and its 90% CI are contained with this region, corresponding to a TOST p-value of ≈ 0.004. The bottom row gives the 95% PI, which is also narrow due to the low between-study heterogeneity $\tau^2 \approx 6e-11$. The CCC is 0.94, indicating strong concordance between the study-level effects on the primary endpoint and on the composite signature. **Alt text:** Meta-analysis for the combined 7-marker signature.

Supplementary Material: Meta-Analysis of High-Dimensional Surrogate Markers

S1 Web Appendix A: Methods

S1.1 Notation

For a reminder of the notation used throughout this Supplementary Material: we observe data from $m = 1, \ldots, M$ trials, each with $i = 1, \ldots, n_m$ individuals receiving the same binary treatment $A_{i,m} \in \{0, 1\}$, where $a = 1$ denotes active treatment and $a = 0$ denotes control. Let $Y^a_{i,m}$ and $\boldsymbol{S}^a_{i,m} = (S^a_{i,1,m}, \ldots, S^a_{i,J,m})^T$ denote the primary endpoint and vector of J common surrogate candidates, respectively, for individual i in trial m receiving treatment a. In the paired setting, for each individual and a given post-treatment timepoint, we have paired measurements $\{Y^0_{i,m}, Y^1_{i,m}, \boldsymbol{S}^0_{i,m}, \boldsymbol{S}^1_{i,m}\}$, where the superscripts $a = 0$ and $a = 1$ refer to pre- and post-treatment measurements, respectively.

S1.2 Within-study surrogacy estimation

Here, we discuss estimation of the within-study surrogacy parameter δ, where again we omit marker and trial subscripts j, m for notational clarity. We may obtain point estimates for these parameters using rank-based procedures. First, define a function

$$G(x, y) = \begin{cases} 1, & \text{if } x > y, \\ \frac{1}{2}, & \text{if } x = y, \\ 0, & \text{if } x < y. \end{cases}$$

In the active treatment versus control context, we have N^1, N^0 treated and control individuals, respectively. Letting i index treated and j index control, we may estimate

$$\widehat{U}_Y = \frac{1}{N^1 N^0} \sum_{i=1}^{N^1} \sum_{l=1}^{N^0} G(Y^1_i, Y^0_l), \qquad \widehat{U}_S = \frac{1}{N^1 N^0} \sum_{i=1}^{N^1} \sum_{l=1}^{N^0} G(S^1_i, S^0_l),$$

In the paired data setting, we instead simply compare individuals post- and pre-treatment measurements, denoted with superscripts $1, 0$, respectively. If we have N individuals, we may estimate

$$\widehat{U}_Y = \frac{1}{N}\sum_{i=1}^{N} G(Y_i^1, Y_i^0), \qquad \widehat{U}_S = \frac{1}{N}\sum_{i=1}^{N} G(S_i^1, S_i^0).$$

Detail of the estimation of the standard error in the active versus control setting is described in Parast, Cai, and Tian (2024), whereas its derivation in the paired data setting is found in the Supplementary Material of Hughes et al. (2025).

S1.3 Between-study variability estimation

Here, we elaborate on the estimation of the between-study variability parameter τ^2, where we again omit the marker subscript j for clarity. The pooled mean effect depends on the between-study variance τ^2 through the inverse-variance weights

$$w_m(\tau^2) = \frac{1}{\widehat{\sigma}_m^2 + \tau^2}, \qquad \mu(\tau^2) = \frac{\sum_{m=1}^{M} w_m(\tau^2)\widehat{\delta}_m}{\sum_{m=1}^{M} w_m(\tau^2)}.$$

We estimate τ^2 by restricted maximum likelihood (REML), as the nonnegative solution of the restricted score equation. Numerically, the score is evaluated by recomputing the generalised least-squares estimate $\mu(\tau^2)$ at each candidate value of τ^2, and the root is sought on a bounded interval using a numerical root-finding algorithm with a small tolerance. If the score does not change sign on the search interval, or if root finding fails to converge, τ^2 is instead obtained by direct bounded maximisation of the restricted log-likelihood over $\tau^2 \geq 0$. The resulting estimate is truncated at zero to enforce the parameter constraint.

S1.4 Conditions to avoid the surrogate paradox

The surrogate paradox is a situation where the treatment effect on the surrogate is positive, the surrogate has positive association with the primary outcome, but the treatment effect

on the primary outcome is negative. It is shown in Parast et al. (2024), that under the following conditions, we may avoid this:

$$
\begin{aligned}
&(C1) \quad P(Y^a > y \mid S^a = s) \text{ increases in } s \\
&(C2) \quad P(S^1 > s) \geq P(S^0 > s) \quad \forall s \\
&(C3) \quad P(Y^1 > y \mid S^1 = s) \geq P(Y^0 > y \mid S^0 = s) \quad \forall s
\end{aligned}
$$

(C1) states that the surrogate has a non-negative association with the outcome, (C2) that treatment has a non-negative effect on the surrogate, and (C3) that the residual treatment effect on the outcome after conditioning on the surrogate is non-negative. In particular, these conditions imply that $U_S > \frac{1}{2} \implies U_Y > \frac{1}{2}$, that is, that we can conclude a positive treatment effect on Y based on a positive treatment effect observed on S.

S1.5 Pooled effect and variance estimators

The random-effects pooled effect is estimated as the inverse-variance weighted mean

$$
\widehat{\mu} = \frac{\sum_{m=1}^{M} \widehat{w}_m \widehat{\delta}_m}{\sum_{m=1}^{M} \widehat{w}_m}, \qquad \widehat{w}_m = \frac{1}{\widehat{\sigma}_m^2 + \widehat{\tau}^2},
$$

where $\widehat{\sigma}_m^2$ is the estimated within-study variance and $\widehat{\tau}^2$ is the estimated between-study variance. Under the conventional Wald-type approach, the variance of $\widehat{\mu}$ is estimated by

$$
\widehat{\mathrm{Var}}_{\mathrm{conv}}(\widehat{\mu}) = \frac{1}{\sum_{m=1}^{M} \widehat{w}_m},
$$

leading to the normal-approximation confidence interval

$$
\widehat{\mu} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}_{\mathrm{conv}}(\widehat{\mu})}.
$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. This approach does not account for uncertainty in the estimation of τ^2 and can therefore

yield confidence intervals that are too narrow when the number of studies is small or heterogeneity is large. For this reason, we prefer the Hartung-Knapp-Sidik-Jonkman adjustment (Hartung and Knapp, 2001), which rescales the variance estimator by

$$q = \frac{1}{M-1}\sum_{m=1}^{M} \widehat{w}_m(\widehat{\delta}_m - \widehat{\mu})^2, \qquad \widehat{\mathrm{Var}}_{\mathrm{HKSJ}}(\widehat{\mu}) = q \cdot \frac{1}{\sum_{m=1}^{M} \widehat{w}_m}.$$

Confidence intervals are then constructed using the t-distribution:

$$\widehat{\mu} \pm t_{M-1,\,1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}_{\mathrm{HKSJ}}(\widehat{\mu})}.$$

where $t_{M-1,\,1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the Student's t-distribution with $M-1$ degrees of freedom.

S1.6 Concordance correlation coefficient

Here we give further detail on the concordance correlation coefficient (CCC) (Lin, 1989) and its appropriateness in our setting. Here, the idea is to have a correlation-based evaluation metric which could be used as a complement to hypothesis testing and prediction intervals to summarise trial-level surrogacy. The classical methods of Buyse et al. (2000) and Molenberghs et al. (2002) use the coefficient of determination from a weighted linear regression of the trial-level effects on the primary endpoint and on the surrogate candidate, called R^2_{trial}, for this purpose, where $R^2_{trial} = 1$ corresponds to perfect surrogacy.

In our framework, these treatment effects are defined on a common scale. This allows for a stricter notion of perfect surrogacy, namely that treatment effects are not only perfectly correlated across trials but also equal, i.e. $U_{Y,m} = U_{S,m}$ for all $m = 1, \ldots, M$. Therefore, an appropriate metric of trial-level surrogacy should reflect both the correlation between trial-level effects and their agreement relative to the identity line. The CCC satisfies these criteria by jointly quantifying correlation and accuracy. Like standard correlation coefficients, it ranges in $[-1, 1]$, with values near 1 indicating strong concordance and

values near -1 indicating strong discordance. The CCC is defined as

$$\text{CCC} = \frac{2\,\text{Cov}(\boldsymbol{U_Y}, \boldsymbol{U_S})}{\text{Var}(\boldsymbol{U_Y}) + \text{Var}(\boldsymbol{U_S}) + \left(\mathbb{E}[\boldsymbol{U_Y}] - \mathbb{E}[\boldsymbol{U_S}]\right)^2},$$

where $\boldsymbol{U_Y} = (U_{Y,1}, \ldots, U_{Y,M})^T$ and $\boldsymbol{U_S} = (U_{S,1}, \ldots, U_{S,M})^T$ denote the vectors of trial-level treatment effects on the primary endpoint and generic surrogate marker S, respectively. In practice, CCC can be estimated using standard sample estimators of the moments across trials.

S1.7 Intraclass correlation coefficient

Here, we briefly introduce the intraclass correlation coefficient (ICC), a reliability index derived from linear mixed-effects models, as an alternative validation metric to the CCC presented above. There are multiple forms of the ICC, defined by the model used to derive the index, the type of measurements, and the definition of agreement (Koo and Li, 2016). The estimation and interpretation of each differs, so it is important to choose an appropriate type based on the context.

In our case, we have paired measurements $(U_{Y,m}, U_{S,m}) \quad \forall m \in \{1, \ldots, M\}$. The appropriate form is therefore the two-way random-effects model, assessing absolute agreement between paired trial-level estimates at the individual level, which corresponds to the ICC(2,1) (Koo et al., 2016). This may be estimated as follows.

$$\bar{U_S} = \frac{1}{M}\sum_{m=1}^{M} U_{S,m}, \quad \bar{U_Y} = \frac{1}{M}\sum_{m=1}^{M} U_{Y,m}, \quad \bar{U} = \frac{1}{2M}\sum_{m=1}^{M}\left(U_{S,m} + U_{Y,m}\right),$$

$$\bar{U_m} = \frac{U_{S,m} + U_{Y,m}}{2}.$$

$$MS_R = \frac{2}{M-1}\sum_{m=1}^{M}\left(\bar{U_m} - \bar{U}\right)^2,$$

$$MS_C = M\left[\left(\bar{U_S} - \bar{U}\right)^2 + \left(\bar{U_Y} - \bar{U}\right)^2\right],$$

$$MS_E = \frac{1}{M-1}\sum_{m=1}^{M}\Big[\left(U_{S,m} - \bar{U_m} - \bar{U_S} + \bar{U}\right)^2 + \left(U_{Y,m} - \bar{U_m} - \bar{U_Y} + \bar{U}\right)^2\Big],$$

$$\text{ICC}(2,1) = \frac{MS_R - MS_E}{MS_R + MS_E + \frac{2}{M}(MS_C - MS_E)}.$$

The ICC(2,1) may be interpreted as the proportion of total variability in the paired trial-level estimates that is attributable to between-trial heterogeneity, rather than disagreement between the surrogate and primary treatment effect estimates. In other words, it quantifies the extent to which the surrogate and primary effects are consistent in capturing differences across trials, relative to their lack of agreement within trials.

We have a slight preference for the CCC as it provides a more direct and model-free measure of concordance with respect to the identity line $U_S = U_Y$, combining both correlation and systematic deviation. Nevertheless, the ICC(2,1) remains a valid complementary reliability-based summary of agreement.

S2 Web Appendix B: Simulation Study

S2.1 Meta-analysis model specification and estimation

Here, we investigate the impact of the meta-analysis model specification and estimation on the false positive rate control from the resulting hypothesis tests. In particular, we examine the choice to perform random-effects as opposed to fixed-effect meta-analysis. Separately, for random-effects meta-analysis, we evaluate the use of the HKSJ adjustments for the variance estimates of the pooled effects as opposed to the conventional normal approximation.

Results are given in Web Figure 4. The fixed-effect model approach exceeds the nominal FPR for all but the smallest combinations of heterogeneity parameters. The random-effects model combined with the conventional approach to τ^2 estimation controls the FPR relatively well, but exceeds it for low to moderate numbers of studies and

heterogeneity. The random-effects model combined with the HKSJ approach controls the FPR well across all settings.

Overall, these results support the use of the random-effects model combined with the HKSJ approach when the number of studies is low-to-moderate, and when there is any doubt about the assumptions underlying the fixed-effect model.

S2.2 Distribution of true mean

As briefly explained in the main text, as our TOST equivalence testing procedure corresponds to testing a region of the parameter space, not all true parameter values will produce equal results. This is why, when investigating the test calibration, we generated invalid surrogates having true means at the boundaries of the valid region, corresponding to the least favourable configuration (LFC). It follows that, for invalid surrogates with means further into the invalid region of the parameter space, the empirical FPR will be at most as conservative as the same setting under the LFC. For the investigation of the power, we allowed the true means to vary uniformly across the valid region.

In this section, we investigate the properties of the test under different settings of true mean generation. Web Figure 5 presents the empirical FPR as a function of the true mean of invalid surrogate generation for 3 values of maximum between-study variance. As the true mean moves further into the invalid region, the empirical FPR quickly drops to 0. This is a consequence of the equivalence test setup, which must be calibrated to control the FPR in the LFC.

Next, we investigated empirical power when fixing the true mean of valid surrogate generation. The results are presented in Web Figure 6. As expected, empirical power peaks when the true mean is generated at the greatest absolute distance from the equivalence region bounds i.e. $\mu = 0$, and goes to 0 as the mean approaches the bounds. The power improves for increasing numbers of studies.

S2.3 Nonparametric simulation

As noted in the main text, permutation-based simulations are useful for assessing test behaviour when the parametric model is uncertain and the data-generating distribution is unknown. The aim is to construct a permutation scheme on a real dataset that breaks the associations of interest while preserving key features of the observed data.

We used a subset of the Immune Signatures 2 dataset from the high-dimensional application in the main manuscript. This multi-study dataset contains paired pre- and post-influenza vaccination measurements. The primary endpoint was the mean antibody response across the three vaccine strains within each individual and timepoint. The surrogate candidates were the expression levels of 10,086 genes, analysed at the gene-level rather than the geneset-level. The primary endpoint was measured at day 0 and day 28 ($\pm$7 days), whereas the surrogate candidates were measured at day 0 and day 7. Although day 1 was used in the main application, day 7 was chosen here because it yielded a larger pool of available studies.

The permutation scheme was designed to break the association between the paired primary endpoint measurements and the paired surrogate measurements. Within each study, we permuted the paired primary endpoint values across individuals while leaving the surrogate pairs unchanged. This preserves study-level effects, within-individual correlation between pre- and post-vaccination measurements of the same variable, and between-marker correlation within individuals. Under this construction, the permuted datasets contain only true negatives, so this setting was used to assess false positive rate (FPR) control only.

Two design features were varied: the number of samples per study and the number of studies. These were controlled by subsampling from the permuted datasets. To study the effect of sample size:

1. Define the sample size grid: $n^* \in \{5, 10, \ldots, 25\}$.

2. Restrict to studies with at least 25 individuals: $M^* = \{m : n_m \geq 25\}$.

3. For $b = 1, \ldots, 1000$:

- subsample n^* individuals per study;
- apply the within-study permutation scheme;
- apply RISE-meta to obtain p-values $\{p_j^{(b)}\}_{j=1}^J$;
- compute the empirical FPR

$$\widehat{\text{FPR}}_b(n^*) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\left(p_j^{(b)} < \alpha\right).$$

A maximum sample size of 25 was chosen so that the same set of studies remained available across the grid, leaving 6 studies.

To study the effect of the number of studies:

1. Restrict to studies with at least 15 individuals: $M^* = \{m : n_m \geq 15\}$.
2. Define the study grid: $m^* \in \{2, \ldots, 10\}$.
3. For $b = 1, \ldots, 1000$:
   - sample a subset of studies $S_b \subset M^*$ with $|S_b| = m^*$;
   - subsample 15 individuals per study;
   - apply the within-study permutation scheme;
   - apply RISE-meta to obtain p-values $\{p_j^{(b)}\}_{j=1}^J$;
   - compute the empirical FPR

$$\widehat{\text{FPR}}_b(m^*) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\left(p_j^{(b)} < \alpha\right).$$

Here, 15 individuals per study was chosen so that up to 10 studies were available.

Web Figure 8 shows the results for the sample-size investigation. The average empirical FPR remained well below the nominal level $\alpha = 0.05$ across all sample sizes, and decreased towards 0 as the number of observations increased. This conservatism is expected: unlike in the parametric simulations, the invalid surrogates in this real-data setting are unlikely

to lie at the least favourable configuration used in the parametric case, and their true mean is probably farther into the invalid region (see Web Figure 5 for a demonstration).

Web Figure 7 shows the results for the study-number investigation. False positive rate control was maintained across all values of m^*. For the smallest number of studies, the empirical FPR occasionally exceeded the nominal level, with a maximum of 0.10 when only 2 studies were used. As the number of studies increased, the empirical mean again approached 0.

Overall, these results support the statistical validity of the equivalence testing in RISE-Meta in realistic scenarios under unknown data generating processes.

S3 Web Appendix C: Data applications

S3.1 Low-dimensional examples

S3.1.1 Description of bivariate joint modelling approach

Buyse et al. (2000) and Molenberghs et al. (2002) describe an approach using bivariate mixed-effects models to assess trial-level surrogacy in the meta-analytic framework. Using our notation for an arbitrary surrogate candidate S for individual i in trial m, they consider models

$$S_{i,m} = \mu_S + m_{S,m} + (\alpha + a_m)A_{i,m} + \epsilon_{S_{i,m}},$$

$$Y_{i,m} = \mu_Y + m_{Y,m} + (\beta + b_m)A_{i,m} + \epsilon_{Y_{i,m}}$$

where μ_S, μ_Y are fixed intercepts, $m_{S,m}, m_{Y,m}$ are trial-specific random intercepts, α, β are fixed treatment effects, and a_m, b_m are trial-specific random treatment effects. The random-effects $\boldsymbol{V} = (m_{S,m}, m_{Y,m}, a_m, b_m)^T$ are assumed to be multivariate normally distributed $V \sim \mathcal{N}(\mathbf{0}, D)$, where

$$D = \begin{pmatrix} d_{SS} & d_{SY} & d_{Sa} & d_{Sb} \\ d_{SY} & d_{YY} & d_{Ya} & d_{Yb} \\ d_{Sa} & d_{Ya} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Yb} & d_{ab} & d_{bb} \end{pmatrix}$$

and the error terms $\boldsymbol{\epsilon}_{i,m} = (\epsilon_{S_{i,m}}, \epsilon_{Y_{i,m}})^T$ are bivariate normally distributed $\boldsymbol{\epsilon}_{i,m} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{SY} \\ \sigma_{SY} & \sigma_{YY} \end{pmatrix}$$

This gives the trial-level surrogacy metric

$$R^2_{trial} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}$$

However, as described in Burzykowski, Molenberghs, and Buyse (2005) and Tibaldi et al. (2003), this approach often results in computational problems. Indeed, in our two data examples from the main manuscript, we were unable to fit these models using the `BimixedContCont()` function from the `Surrogate` package, either as described above or in a form with reduced complexity of the random-effects structure. This was due to numerical convergence issues arising from the estimation of D as a singular covariance matrix. An alternative approach is to instead fit a simplified modelling strategy using a two-stage bivariate fixed-effect modelling approach, implemented through the `BifixedContCont()` function.

In the first stage, this uses bivariate fixed-effect model

$$S_{i,m} = \mu_{S,m} + \alpha_m A_{i,m} + \epsilon_{S_{i,m}},$$

$$Y_{i,m} = \mu_{Y,m} + \beta_m A_{i,m} + \epsilon_{Y_{i,m}}$$

where $\mu_{S,m}, \mu_{Y,m}$ are trial-specific fixed intercepts, α_m, β_m are trial-specific fixed treatment effects, and the error terms are distributed as above.

In the second stage, the trial-specific treatment effects on the primary endpoint are modelled as a function of the trial-specific surrogate intercepts and treatment effects,

estimated from the first stage, as

$$\widehat{\beta_m} = \kappa_0 + \kappa_1 \widehat{\mu}_{S,m} + \kappa_2 \widehat{\alpha}_m + \epsilon_i$$

This model is fit using weighted least squares, where the weights are the within-study sample sizes. Then, the R^2_{trial} is taken as the coefficient of determination from this weighted linear regression fit. A 95% confidence interval around this quantity can be estimated using the delta method.

S3.1.2 Discussion on analysis choices

A few choices in the low-dimensional data analysis examples merit discussion. We note that we are using these data sets as illustrations for benchmarking the qualitative conclusions of the two methods, and the resulting analysis should be interpreted with caution. In addition, in these analyses, our analysis choices follow the lead of previous works which have already used these datasets (or sub/supersets thereof) (Burzykowski et al., 2005; Buyse et al., 2000; Molenberghs et al., 2002; Tibaldi et al., 2003).

Firstly, we discuss the fact that we choose to use the treatment center as the unit of analysis for the analysis of the data. This choice enables application of the meta-analytic frameworks, which require between-unit variability. In the case of the `ARMD` data, the data arise from a single RCT, and therefore cannot be analysed with meta-analytical methods without further choosing a smaller investigative unit. The `Ovarian` dataset comes from 4 RCTs, but information on these is not available in the dataset accessible from the `Surrogate` package. Further discussion on the choice of units for meta-analytical evaluation of surrogate markers can be found in Abrahantes et al. (2004).

Another point of discussion is the treatment of survival times as continuous outcomes. Here, we used log transformations on the survival times to render them continuous and unbounded. However, we note that an approach tailored to survival data, and in particular taking into account censoring, is a more appropriate way to analyse these data.

S3.1.3 Evaluation metrics and confidence intervals

We estimate a R^2_{trial} statistic for the RISE-Meta results in a similar spirit to the bivariate fixed-effect modelling approach, in two stages. First, we compute estimates $\widehat{\boldsymbol{U}}_{\boldsymbol{Y}} = (\widehat{U}_{Y,1}, \ldots, \widehat{U}_{Y,M})^T$ and $\widehat{\boldsymbol{U}}_{\boldsymbol{S}} = (\widehat{U}_{S,1}, \ldots, \widehat{U}_{S,M})^T$ and then fit a model

$$\widehat{U}_{Y,m} = \kappa_0 + \kappa_1 \widehat{U}_{S,M} + \epsilon_i$$

with weighted least squares, where the weights are the sample sizes n_m, taking the R^2 statistic from this fit.

In the case of the bivariate fixed-effects model approach and the R^2_{trial} statistic, the 95% CI could be extracted directly from the `BifixedContCont()` function, and is estimated via the delta method. Otherwise, the variance of the R^2_{trial} for RISE-Meta and the CCC for both methods is estimated with bias-corrected and accelerated bootstrapping (Efron, 1987). Specifically, we generate $B = 2000$ datasets by re-sampling with replacement M pairs $(U_{Y,m}, U_{S,m})$.

S3.2 High-dimensional example

In this section, we discuss the high-dimensional data application example, including a fuller description of the dataset, as well as sensitivity and supplementary analyses.

S3.2.1 Description of the Immune Signatures 2 dataset and preprocessing

Here we give a fuller description of the subset of the Immune Signatures 2 (IS2) dataset considered as an illustrative example for the application of RISE-Meta. The Immune Signatures 2 dataset is a release of the Immune Signatures Data Resource (ISDR), which is a standardised compendium of systems vaccinology datasets. A description of the full dataset can be found in Diray-Arce et al. (2022). For our purposes, we considered only datasets on inactivated influenza vaccination, since this vaccine had available data from multiple studies, necessary for meta-analysis. All studies consisted of longitudinal measurements of gene expression markers. Therefore, it was necessary to make a choice

of the post-vaccination timepoint used for the analysis. A selection of the candidate timepoints available for each of the trivalent inactivated influenza studies, and the valid samples at each, is presented in Web Figure 1. Day 1 was chosen as the post-treatment timepoint of interest due to our *a priori* knowledge and experience with these datasets suggesting the most promising signal would likely be found soon after vaccination. A flowchart showing the preprocessing steps taken to filter the individuals from the IS2 dataset is given in Web Figure 2. This includes the removal of individuals without gene expression measurements at both baseline and day 1, immune response measurements at both baseline and day 28 ± 7 days, and studies containing 5 or fewer individuals. The latter filtering stage is important as, below a certain number of individuals, the RISE method cannot estimate the sampling variability of the surrogacy parameter δ.

S3.2.2 Geneset pre-processing

Initially, the BTMs consisted of $J = 346$ genesets. For better interpretability, we preprocessed the BTM genesets to contain only modules with informative titles, which involved the removal of any geneset titled "TBA" (to be attributed). In addition, one geneset that had no member genes in common with the $10,086$ initial input features was removed. After these filtering steps, we obtained $J = 258$ features as inputs into RISE-Meta.

S3.2.3 Fixed-effect analysis

Here, we perform the same analysis as in the main manuscript, but instead do fixed-effect meta-analysis. Web Figure 9 shows the results of the marker-level meta-analysis. The standard errors for each marker are smaller compared to the random-effects meta-analysis, leading to more significant markers (16 compared to 7). Web Figure 10 shows the results of the evaluation of the 16-marker signature on the 33% of unseen data. Interestingly, the point estimates are less consistent using this 16-marker signature compared to the 7-marker signature identified in the RE analysis. In addition, the CI is wider and the CCC is smaller. This is possibly a reflection of the additional noise incorporated into the

signature due to the more relaxed significance criteria when using the FE method. Overall, however, the qualitative conclusions of the markers forming this signature stay the same as in the RE analysis, with this signature being dominated by antiviral, dendritic cell, and monocyte signals.

S3.2.4 Impact of geneset choice and aggregation method

Here, we perform supplementary analyses to evaluate the qualitative impact of the choice of geneset definition and aggregation method. In particular, we investigate the use of a different set of genesets, called the BloodGen3Modules (BG3M), as presented in Rinchai et al. (2021). The BG3Ms were processed in a similar way to the BTMs, where an initial number of $J = 382$ genesets was filtered by removing sets with uninformative titles (i.e. genesets titled "TBD"), resulting in $J = 205$ inputs to RISE-Meta. In this analysis, we continue to use the mean as the aggregation function.

Web Figure 11 presents the results of the marker-level meta-analysis when using the BG3M genesets. Here, only 1 geneset is significant after multiplicity correction. Qualitatively, the top of this list is still dominated by interferon signals, with several other innate immune modules further down (monocytes, inflammation). Web Figure 12 shows the evaluation of this single marker on unseen data, where the null hypothesis for non-equivalence of the pooled effect is rejected.

In addition, we performed RISE-Meta when choosing not to aggregate at all to the geneset-level. Web Figure 13 shows the top results for the marker-level meta-analysis in this analysis. Although many markers have pooled effects well within the equivalence margin, due to the severe multiplicity penalty after performing over 10,000 hypothesis tests, no marker is significant after adjustment. This points to an advantage of the geneset approach, which allows for a prior dimension reduction of the candidate biomarkers.

Supplementary Material References

Abrahantes, José Cortiñas, Geert Molenberghs, Tomasz Burzykowski, Ziv Shkedy, Ariel Alonso Abad, and Didier Renard (Oct. 2004). "Choice of units of analysis and modeling strategies in multilevel hierarchical models". In: *Computational Statistics and Data Analysis* 47.3, pp. 537–563. ISSN: 0167-9473. DOI: 10.1016/j.csda.2003.12.003.

Burzykowski, Tomasz, Geert Molenberghs, and Marc Buyse (2005). *The Evaluation of Surrogate Endpoints*. Ed. by Tomasz Burzykowski, Geert Molenberghs, and Marc Buyse. Springer New York. ISBN: 9780387270807. DOI: 10.1007/b138566.

Buyse, Marc, Geert Molenberghs, Tomasz Burzykowski, Didier Renard, and Helena Geys (Apr. 2000). "Statistical Validation of Surrogate Endpoints: Problems and Proposals". In: *Drug Information Journal* 34.2, pp. 447–454. ISSN: 2164-9200. DOI: 10.1177/009286150003400213.

Diray-Arce, Joann, Helen E. R. Miller, Evan Henrich, Bram Gerritsen, Matthew P. Mulè, Slim Fourati, et al. (Oct. 2022). "The Immune Signatures data resource, a compendium of systems vaccinology datasets". In: *Scientific Data* 9.1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01714-7.

Efron, Bradley (Mar. 1987). "Better Bootstrap Confidence Intervals". In: *Journal of the American Statistical Association* 82.397, pp. 171–185. ISSN: 1537-274X. DOI: 10.1080/01621459.1987.10478410.

Hartung, Joachim and Guido Knapp (Dec. 2001). "A refined method for the meta-analysis of controlled clinical trials with binary outcome". In: *Statistics in Medicine* 20.24, pp. 3875–3889. ISSN: 1097-0258. DOI: 10.1002/sim.1009.

Hughes, Arthur, Layla Parast, Rodolphe Thiébaut, and Boris P. Hejblum (Sept. 2025). "RISE: Two-Stage Rank-Based Identification of High-Dimensional Surrogate Markers Applied to Vaccinology". In: *Statistics in Medicine* 44.20–22. ISSN: 1097-0258. DOI: 10.1002/sim.70241.

Koo, Terry K. and Mae Y. Li (2016). "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". In: *Journal of Chiropractic Medicine* 15.2, pp. 155–163. ISSN: 1556-3707. DOI: 10.1016/j.jcm.2016.02.012.

Lin, Lawrence I-Kuei (Mar. 1989). "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* 45.1, p. 255. ISSN: 0006-341X. DOI: 10.2307/2532051.

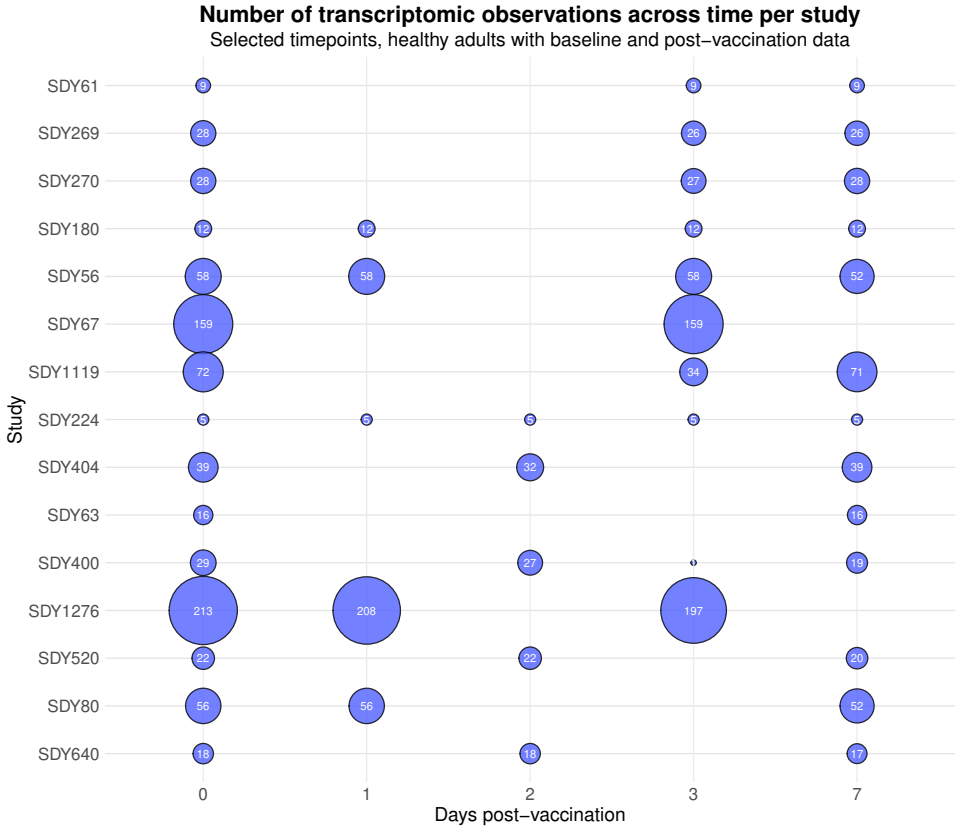
Molenberghs, Geert, Marc Buyse, Helena Geys, Didier Renard, Tomasz Burzykowski, and Ariel Alonso (Dec. 2002). "Statistical challenges in the evaluation of surrogate endpoints in randomized trials". In: *Controlled Clinical Trials* 23.6, pp. 607–625. ISSN: 0197-2456. DOI: 10.1016/s0197-2456(02)00236-2.

Parast, Layla, Tianxi Cai, and Lu Tian (Jan. 2024). "A rank-based approach to evaluate a surrogate marker in a small sample setting". In: *Biometrics* 80.1. ISSN: 1541-0420. DOI: 10.1093/biomtc/ujad035.

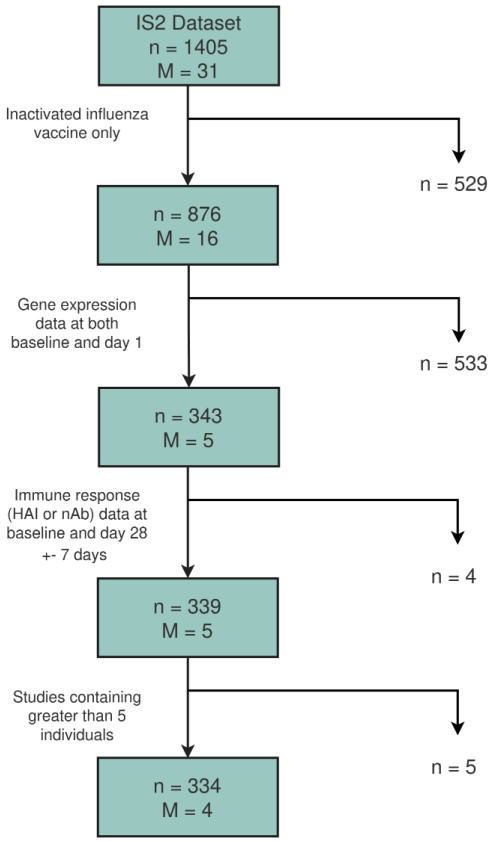
Rinchai, Darawan, Jessica Roelands, Mohammed Toufiq, Wouter Hendrickx, Matthew C Altman, Davide Bedognetti, et al. (Feb. 2021). "BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R". In: *Bioinformatics* 37.16. Ed. by Pier Luigi Martelli, pp. 2382–2389. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btab121.

Tibaldi, Fabián, José Cortiñas Abrahantes, Geert Molenberghs, Didier Renard, Tomasz Burzykowski, Marc Buyse, et al. (Sept. 2003). "Simplified hierarchical linear models for the evaluation of surrogate endpoints". In: *Journal of Statistical Computation and Simulation* 73.9, pp. 643–658. ISSN: 1563-5163. DOI: 10.1080/0094965031000062177.

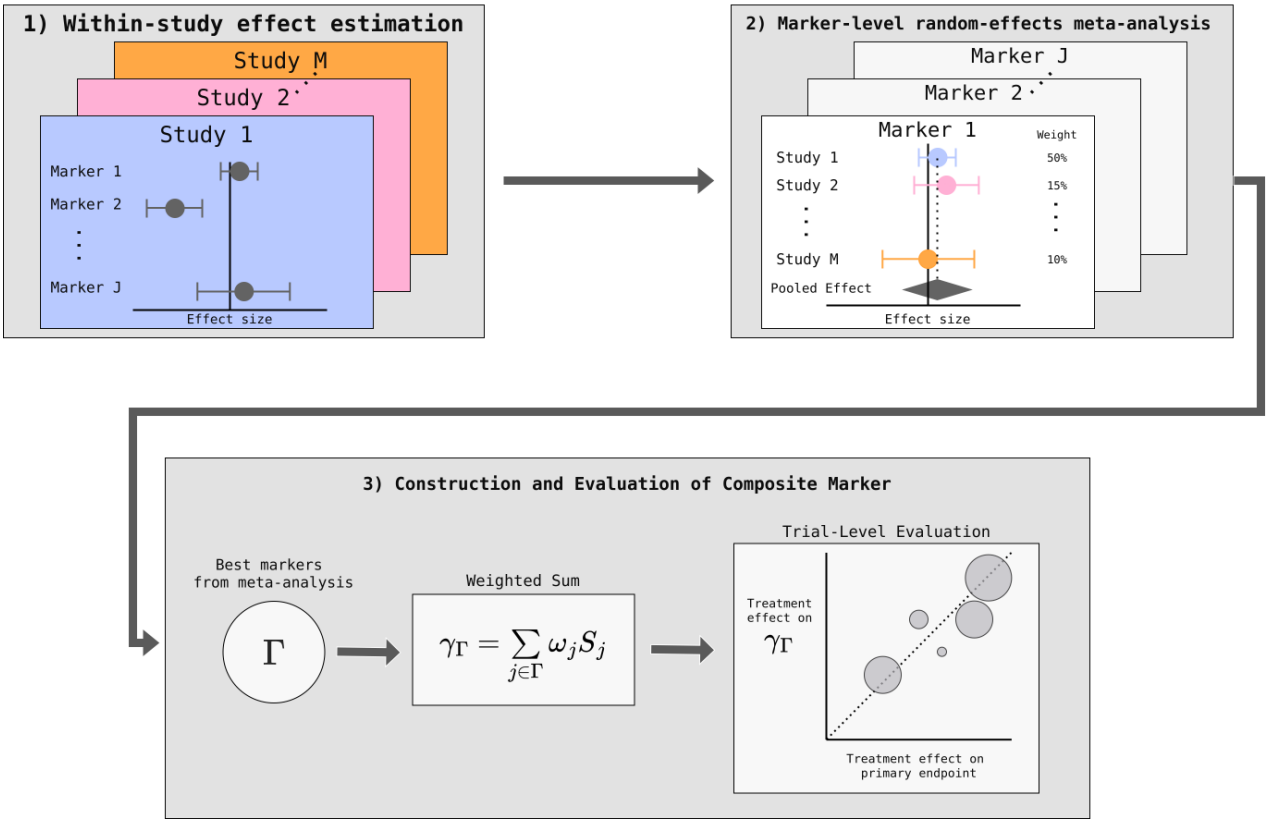
S4 Web Appendix D: Supplementary Figures



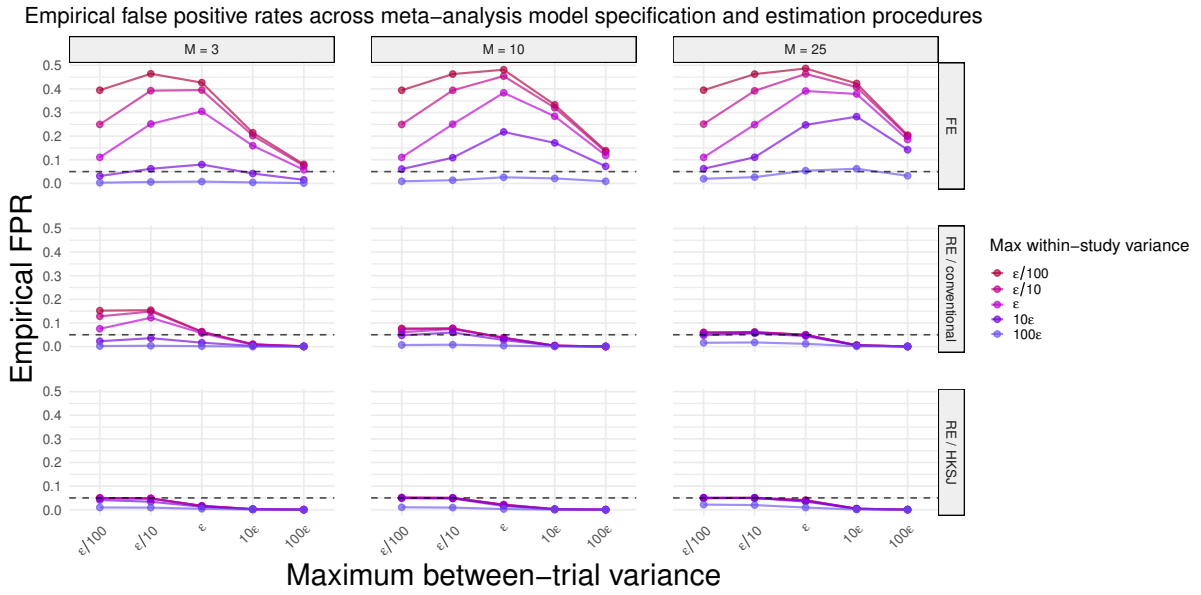
Web Figure 1: **Available timepoints and sample sizes per inactivated influenza study in the IS2 data**. The sample sizes denote the unique individuals with available immune response data (nAb or HAI) at each timepoint.



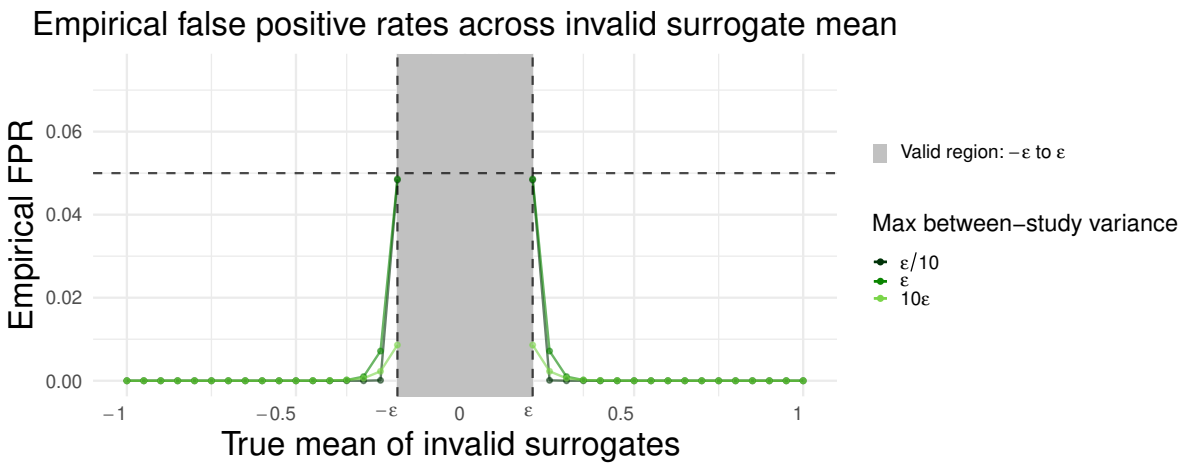
Web Figure 2: **Flowchart showing the filtration of the IS2 dataset prior to analysis**. n denotes the total number of unique individuals and M the number of studies in the dataset at each stage.



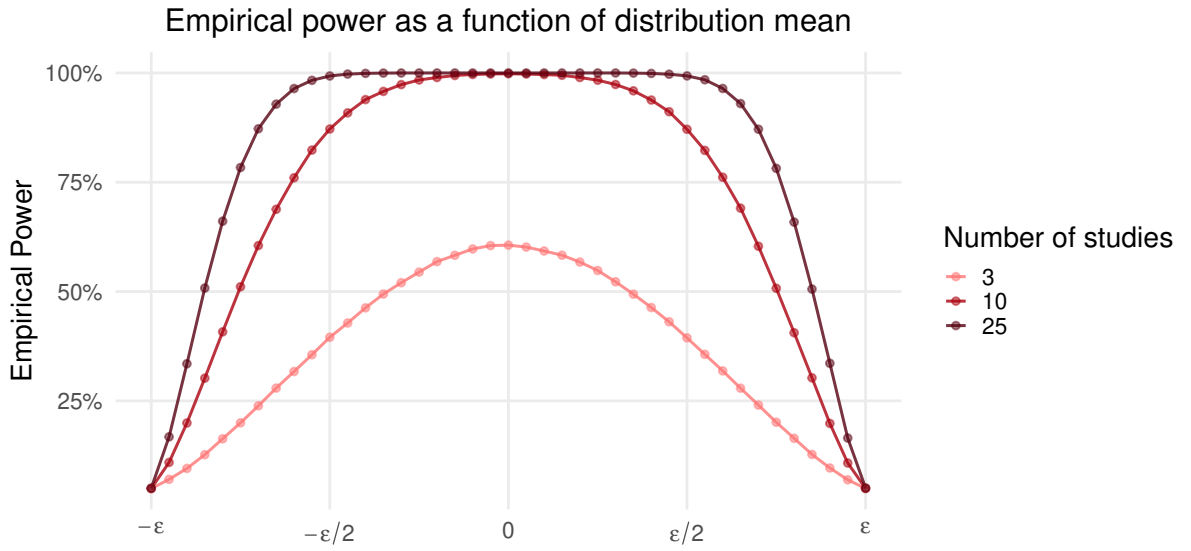
Web Figure 3: **Graphical overview of RISE-Meta**.



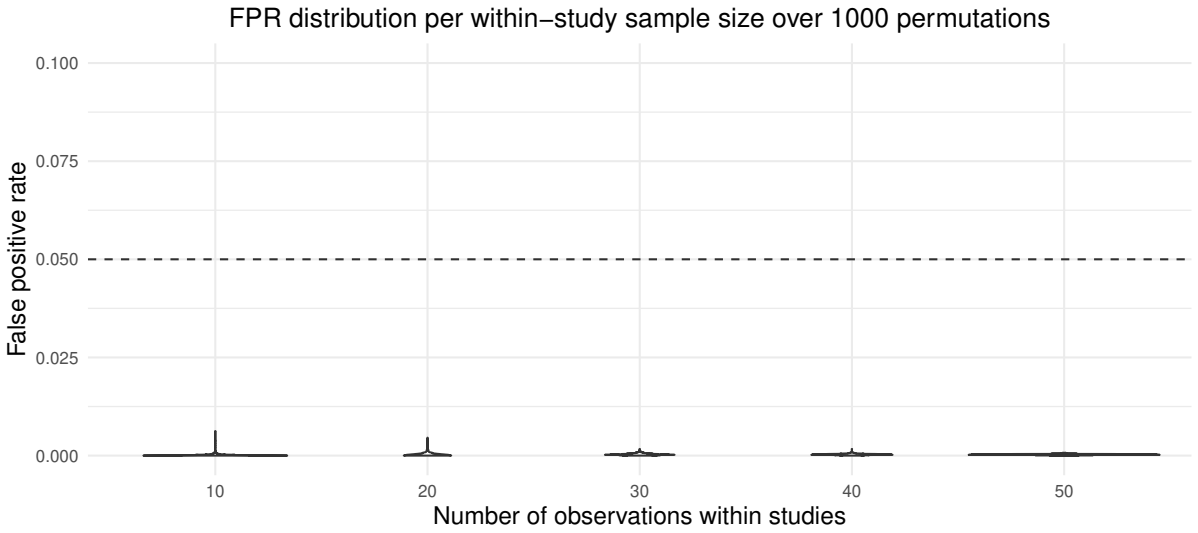
Web Figure 4: **Empirical false positive rates for different meta-analysis model specifications and estimation procedures.** Each plot in the grid gives results for one combination of model specification/estimation procedure for the between-study variance τ^2 and number of trials M. *FE* and *RE* refer to fixed and random-effects models. *Conventional* refers to the conventional estimator of τ^2 using the normal approximation whereas *HKSJ* refers to the Hartung-Knapp-Sidik-Jonkman approach. For each plot, the empirical FPR is plotted against the maximum between-trial variance parameter $u_{\tau^2,max}$ where the coloured lines refer to different values of the maximum within-trial variance parameter $u_{\nu,max}$. Both of these parameters are given relative to a fixed value of $\varepsilon = 0.1$, which defines the equivalence margin for the TOST. The nominal false positive rate $\alpha = 0.05$ is plotted as a dashed line.



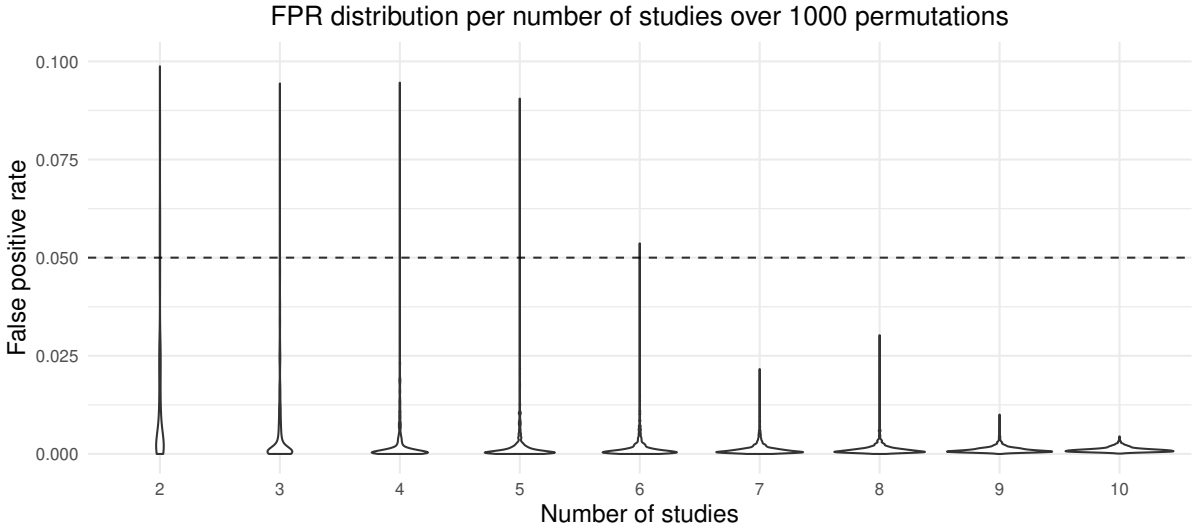
Web Figure 5: **Empirical false positive rates across different true mean values for invalid surrogate generation.** The grey area denotes the valid region $(-\varepsilon, \varepsilon)$, whereas the rest corresponds to the invalid region. True mean generation on these boundaries corresponds to the least favourable configuration investigated in other simulations. The line colours correspond to different values of $u_{\tau^2,max}$ determining the maximum between-study variance. The fixed parameter values were $u_{\tau^2,max} = \varepsilon/10$, $\alpha = 0.05$, $M = 25$, $n_m = 250$, and $\varepsilon = 0.2$.



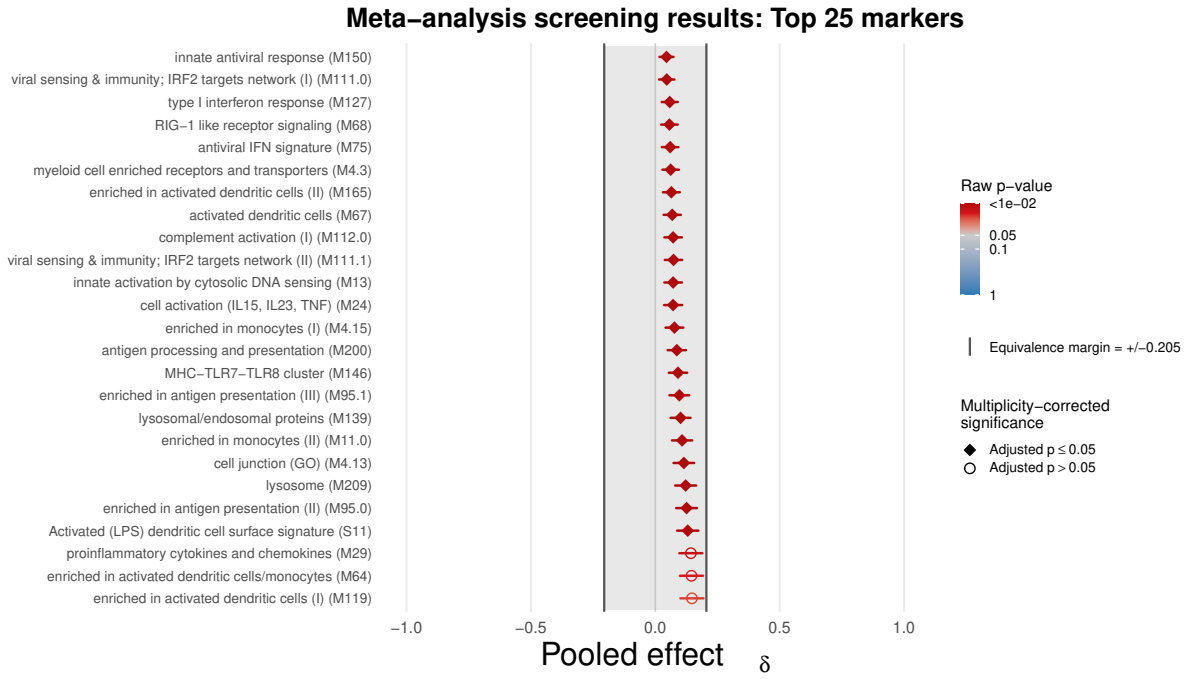
Web Figure 6: **Empirical power across different true mean values for valid surrogate generation.** The x axis corresponds to different values in the valid region $(-\varepsilon, \varepsilon)$ for true mean generation. The line colours correspond to different values of the number of studies M. The fixed parameter values were $u_{\tau^2,max} = u_{\nu,max} = \varepsilon/10$, $\alpha = 0.05$, $n_m = 50$, and $\varepsilon = 0.2$.



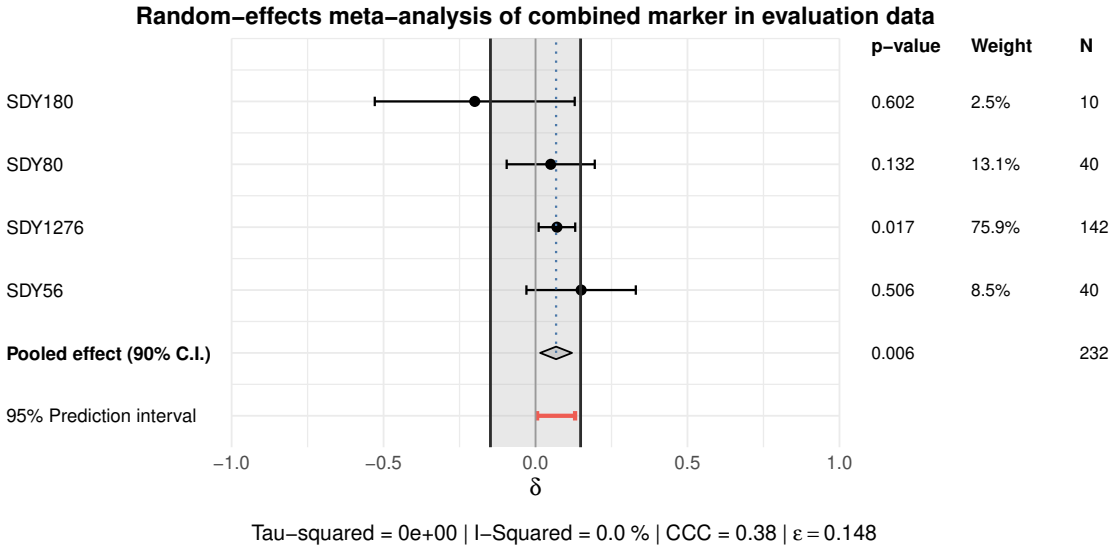
Web Figure 7: **Nonparametric simulation investigating the impact of the number of samples within each study.** The mean empirical false positive rates across 1,000 permuted datasets is shown as violin plots. Each value on the x-axis corresponds to the number of subsampled observations for a fixed pool of 6 studies containing at least 50 observations. The horizontal dashed line corresponds to the nominal false positive rate $\alpha = 0.05$.



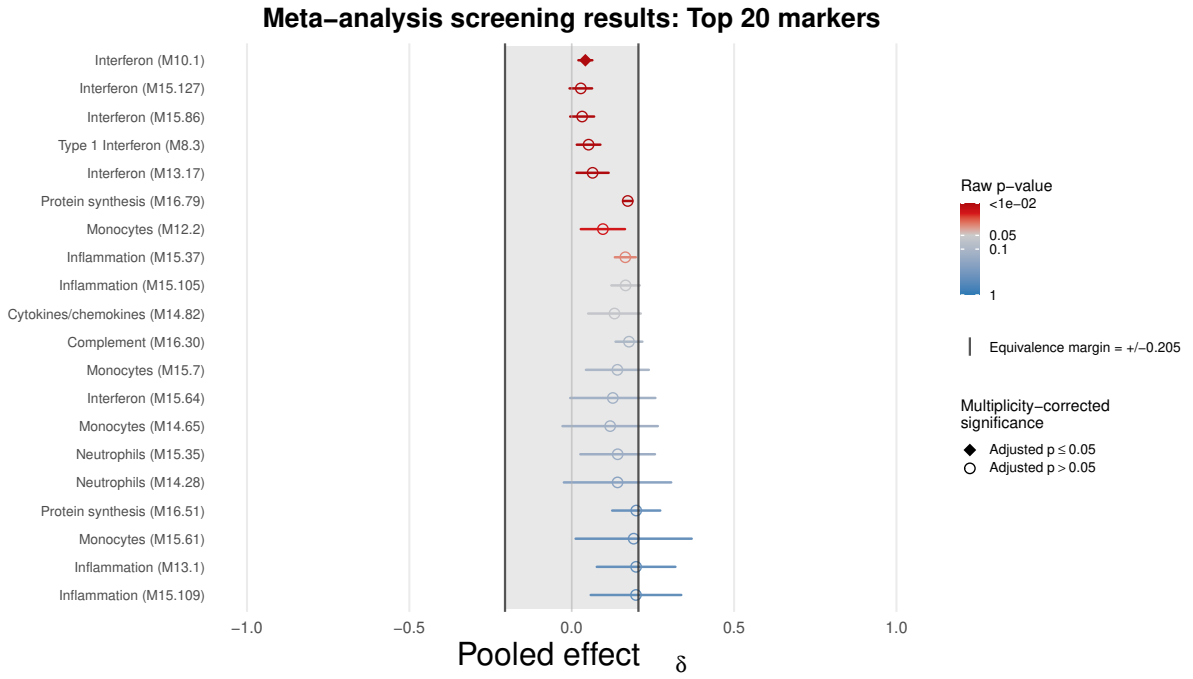
Web Figure 8: **Nonparametric simulation investigating the impact of the number of studies for meta-analysis.** The mean empirical false positive rates across 1,000 permuted datasets is shown as violin plots. Each value on the x-axis corresponds to the number of subsampled studies from a fixed pool of 10 studies with at least 25 individuals. The horizontal dashed line corresponds to the nominal false positive rate $\alpha = 0.05$.



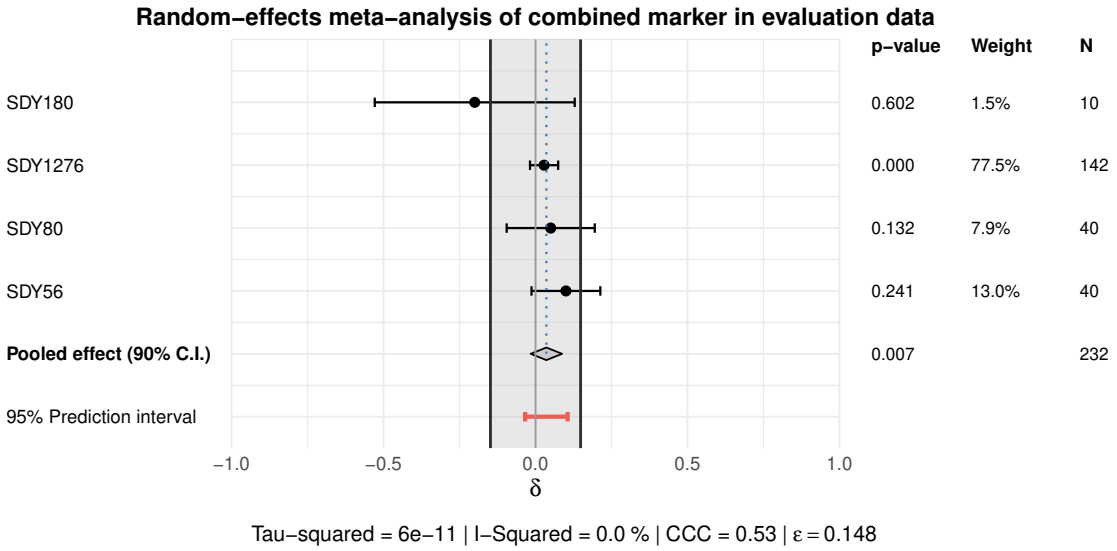
Web Figure 9: **Results from the marker-level meta-analysis when using a fixed-effect model.** This shows the pooled effect estimates for the top 25 markers by raw p-value. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.205$. The colour corresponds to the raw p-value, where shades of red indicate points whose 90% CIs are contained entirely within these bounds. The shape of the point denotes the significance decision after adjusting for test multiplicity, which includes the top 16 markers.



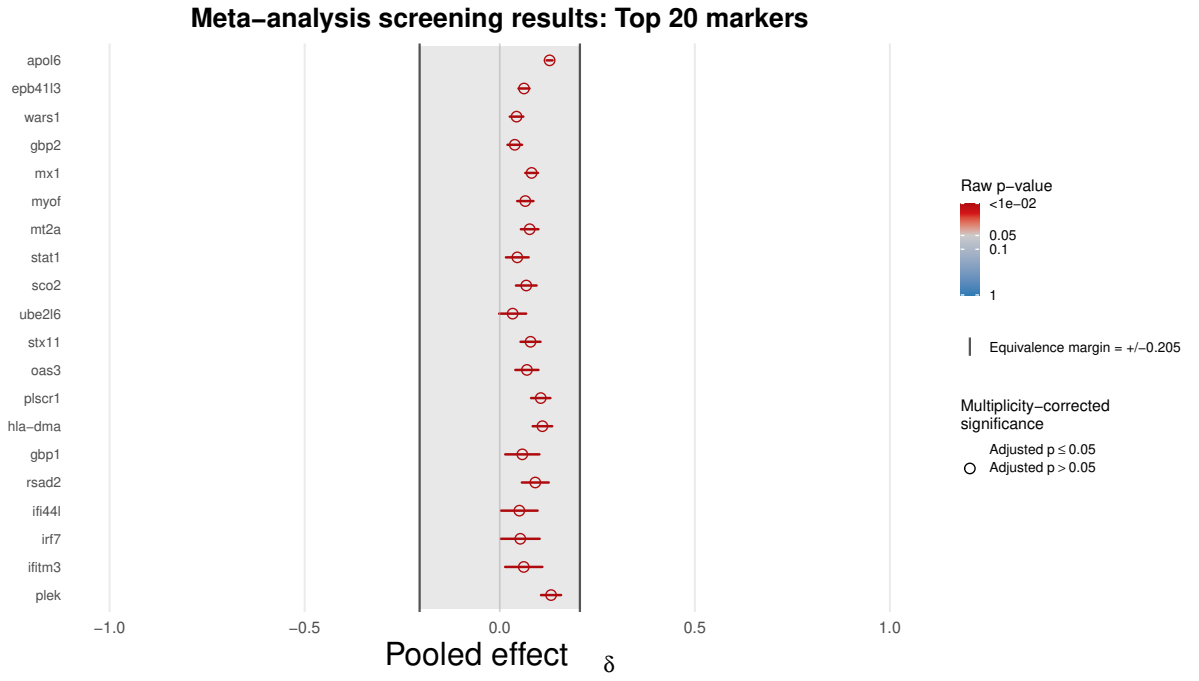
Web Figure 10: **Results from the evaluation of the 16-marker composite signature using a fixed-effect model.** This shows the effect estimates δ for each of the 4 studies, and the pooled effect estimate. The within-study p-values, meta-analysis weights, and number of observations N are given on the right hand side. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.148$. The pooled effect and its 90% CI are contained with this region, corresponding to a TOST p-value of ≈ 0.006.



Web Figure 11: **Results from the marker-level meta-analysis when using the BloodGen3Modules.** This shows the pooled effect estimates for the top 20 markers by raw p-value. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.205$. The colour corresponds to the raw p-value, where shades of red indicate points whose 90% CIs are contained entirely within these bounds. The shape of the point denotes the significance decision after adjusting for test multiplicity, which includes only the top marker.



Web Figure 12: **Results from the evaluation of the single-marker (Interferon - M10.1) surrogate signature using the BloodGen3Modules.** This shows the effect estimates δ for each of the 4 studies, and the pooled effect estimate. The within-study p-values, meta-analysis weights, and number of observations N are given on the right hand side. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.148$. The pooled effect and its 90% CI are contained with this region, corresponding to a TOST p-value of ≈ 0.013.



Web Figure 13: **Results from the marker-level meta-analysis when not aggregating to the geneset-level.** This shows the pooled effect estimates for the top 20 markers by raw p-value. The shaded section is the equivalence region, delimited by the equivalence margins $\varepsilon \approx \pm 0.205$. The colour corresponds to the raw p-value, where shades of red indicate points whose 90% CIs are contained entirely within these bounds. The shape of the point denotes the significance decision after adjusting for test multiplicity, which includes no markers.