Polygenic prediction of treatment efficacy with causal transfer learning

Jiacheng Miao[1,2,*], Jin Mu[3,*], Xiaoyu Yang[3], Jason M. Fletcher[4,5], Lauren L. Schmitz[4], and Qiongshi Lu[3,6,†]

[1]Department of Genetics, Stanford University
[2]Department of Biomedical Data Science, Stanford University
[3]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison
[4]Robert M. La Follette School of Public Affairs, University of Wisconsin–Madison
[5]Department of Population Health Science, University of Wisconsin–Madison
[6]Department of Statistics, University of Wisconsin–Madison
[*]These authors contributed equally to this work
[†]Correspondence to qlu@biostat.wisc.edu

Software

Abstract

Therapeutic interventions often exhibit heterogeneous treatment effects (HTE) across individuals. A central goal of precision medicine is to enable personalized treatment recommendations based on patients' measurable characteristics. Identifying factors that explain HTE is therefore essential. However, detecting HTE remains challenging due to limited sample size in randomized controlled trials (RCTs), often-missing baseline information, and suboptimal statistical methods with limited power. Here, we introduce a principled statistical framework named M-Learner to identify genetically-driven HTE. This approach leverages genetic variation involved in diverse biological pathways influencing drug response, integrates insights from two decades of complex trait genetics, and employs causal transfer learning applicable to both individual-level data and summary statistics. Applying M-Learner to multiple RCTs, we found low bone mineral density as a key determinant of secukinumab efficacy in ankylosing spondylitis, and identified smoker subpopulations adversely affected by a bronchodilator treatment. Together, our findings demonstrate the utility of genetic variation in HTE inference and make important advances toward the promise of precision medicine.

1 Introduction

Heterogeneous treatment effect (HTE) refers to the variation in the effect of a treatment across individuals or subpopulations [1, 2]. Pinning down the source of HTE will have profound implications in precision medicine, which seeks to customize healthcare based on individual characteristics [3, 4]. With a full understanding of how an individual's (pre-treatment) characteristics affect treatment response, healthcare providers can optimize treatment strategies for each patient based on these characteristics, potentially enhancing treatment efficacy and minimizing adverse effects [5].

For example, secukinumab (Cosentyx) is an anti-interleukin-17 (IL-17) monoclonal antibody widely used for treating inflammatory diseases [6]. It has broad applicability across several conditions, but is also known to exhibit considerably variable therapeutic outcomes—response rates

range from 81.6% for psoriasis [7], 62.6% for psoriatic arthritis [8], 60.5% for ankylosing spondylitis [9], to only 30.7% for rheumatoid arthritis [10]. These moderate treatment response rates demonstrate the critical need for more individually tailored therapeutic approaches. Efforts have been made to identify factors that can explain the variation in secukinumab treatment response, but have had limited success [11].

This reflects a broader challenge in precision medicine research: detecting HTE is inherently difficult. Randomized controlled trials (RCTs) often have limited sample size and are primarily designed to estimate average treatment effects rather than HTE. In addition, variables driving the heterogeneity in treatment response are unknown prior to the trial and can be high-dimensional. Therefore, these variables are often unavailable in RCT data, even though the trials may be sufficiently powered to identify HTE.

The challenge to identify HTE is also partly due to a lack of effective methodology. The two primary methods to estimate HTE are subgroup analysis and risk score analysis [12]. Subgroup analysis stratifies the study sample based on their characteristics such as age and sex, followed by statistical comparisons of treatment effects across these groups [5]. This approach is often statistically underpowered due to the limited size of stratified samples, especially when multiple characteristics are used to define subgroups, due to the limited size of stratified samples. Considering each characteristic individually improves power but introduces new challenges in clinical decision making, since an individual may belong to multiple subgroups with varying treatment effects. Alternatively, risk score analysis combines multiple variables to predict the trial outcome, followed by assessing how the treatment effects change across risk levels defined by this score [5]. This approach avoids the multiplicity issue inherent in subgroup analysis. However, risk scores directly predicting treatment effects are rarely available in practice. Instead, researchers often rely on the assumption that some existing score for other outcomes (e.g., risk to develop the disease) can explain HTE. This strategy has major limitations since the risk score developed without using any information on the treatment is not guaranteed to correlate with treatment effectiveness.

In this paper, we introduce two key advances to the current practice of HTE inference. First, we leverage genome-wide genetic variation to predict HTE—a paradigm we term "polygenic efficacy". A main reason why we envision genetics playing a key role in HTE inference is that germline DNA genotypes are fixed at conception. Therefore, post-trial genetic measures are identical to the baseline measures since the treatment will not change DNA genotypes. This means that, with proper consenting in place, it is possible to retrospectively sequence RCT participants, creating high-dimensional "baseline" genetic measures that are known to associate with nearly every human trait, characteristic, and behavior, which can be informative on HTE.

However, genetic data are high-dimensional, with each genetic variant explaining a tiny fraction of variation of most human traits [13]. Therefore, such information needs to be leveraged in conjunction with advanced new methods that can handle infinitesimal genetic effects. Only few studies have explored using genetic information in HTE inference [14]. These studies have largely adopted the conventional methods. For example, a recent study applied one-variable-at-a-time subgroup analysis and risk score analysis to test whether genetic variables modify the therapeutic effects of secukinumab in four inflammatory diseases [11]. Their subgroup analysis at the individual genetic variant level was underpowered due to the small trial size, modest variant effects, and multiple testing. The risk score analysis aggregated genome-wide genetic variation into polygenic risk scores (PRS) [15–17], which predict the risk of developing immune diseases rather than treatment response, and tested whether PRS modifies treatment efficacy [11, 18–20]. Both analyses produced null results, highlighting the limitation of conventional approaches in real-world RCTs [11]. The second key advance in this study is to introduce a principled statistical approach, named

M-Learner, to quantify genetic influences on HTE. It includes methods to handle summary-level association statistics as well as a causal transfer learning framework that enables HTE estimation when individual-level genetic data is available. We demonstrate its utility using RCTs on anti-IL17 monoclonal antibodies for inflammatory diseases and an RCT on a bronchodilator treatment for better lung function in smokers. These approaches align with the growing trend towards more personalized healthcare solutions, promising to enhance patient care by tailoring treatment options to the unique genetic makeup of each individual.

2 Results

Statistical formulation of polygenic efficacy

We first introduce the concept of polygenic efficacy, defined as the aggregated contribution of genome-wide genetic variation to treatment response. To formulate this concept, we adopt a polygenic genotype-by-treatment interaction (G×T) model [21]:

$$Y_i = \sum_{j=1}^{J} G_{ij}\beta_{Gj} + T_i\beta_T + \sum_{j=1}^{J} G_{ij}T_i\beta_{Ij} + \epsilon_i,$$

where Y_i is the treatment outcome (which can be a measure of treatment efficacy or adverse effect, depending on the application) for individual i, G_{ij} is the standardized genotype at variant j (mean 0 and variance 1), T_i is the standardized treatment indicator with treatment probability p, β_T is the treatment effect, β_{Gj} and β_{Ij} denote the additive and interation effects for variant j, and ϵ_i is the error term. Within this framework, we define the polygenic efficacy score (PES) for individual i as the weighted sum of allele counts with weights given by their G×T effects.

$$\text{PES}_i = \sum_{j=1}^{J} G_{ij}\beta_{Ij}$$

Next, we illustrate the connection of PES with the conditional average treatment effect (CATE), a standard metric for quantifying HTE, defined as the expected outcome difference between treatment and control groups given a set of covariates [22]. In our study, the covariates are the allele counts of millions of single-nucleotide polymorphisms (SNPs) in the genome. Under our polygenic G×T model, the CATE for individual i is a linear function of the PES, linking genome-wide GxT interactions directly to HTE:

$$\text{CATE}_i = \text{ATE} + C \times PES_i,$$

where ATE is the average treatment effect, PES_i is the PES for individual i, and $C = (\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}})$ is a positive constant related to the probability p of being assigned to the treatment group in the trial. Therefore, PES explains all the variability in treatment response that is attributable to genetic variation. Assuming that larger Y_i indicates better outcomes, individuals with higher PES_i values are expected to experience greater treatment benefit. The concept of polygenic efficacy also naturally extends to polygenic safety, which captures the aggregated genetic contribution to the risk of developing adverse events or treatment-related harms. We also present a more general, non-linear extension of this framework in the Methods section.

From these derivations, it follows that estimating genetically-driven HTE is equivalent to estimating the PES. However, this task can be challenging in an RCT because of limited sample size

and the typically small GxT effect at the SNP level. As we will illustrate later, directly estimating SNP weights in a PES from RCT data is nearly impossible. To address this challenge, we introduce M-Learner, a causal transfer learning framework that leverages the power of PRS models pre-trained from numerous well-powered GWAS.

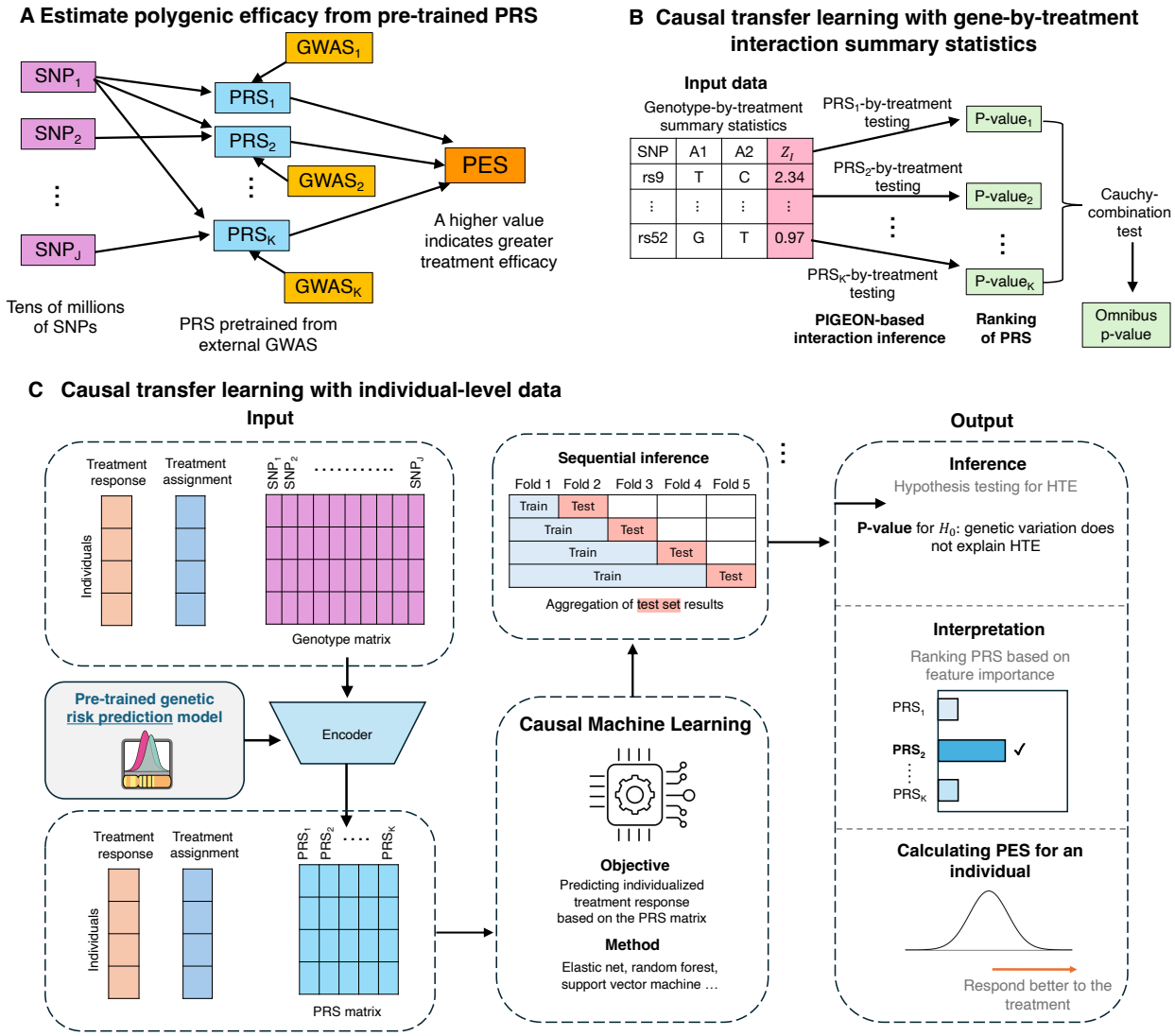

Figure 1: **M-Learner framework for modeling polygenic efficacy and HTE. (A) Genetic representation of polygenic efficacy.** Instead of relying on candidate variants or a single PRS for the trial outcome, M-Learner uses many PRS derived from external GWAS to capture the complex genetic mechanisms underlying HTE. **(B) M-Learner based on summary statistics (M-Learner-S).** Using GxT GWIS summary statistics, M-Learner tests many PRS' interactions with the treatment using PIGEON and integrates them with the Cauchy combination test. The output includes a p-value for testing the HTE and a ranked list of PRS that are informative on the HTE. **(C) M-Learner based on individual-level data (M-Learner-I).** M-Learner applies causal transfer learning to fine-tune pre-trained PRS models for individualized treatment prediction. Sequential inference ensures valid statistical testing. The outputs include a P-value for detecting HTE, a ranked list of PRS contributing to HTE, and individualized predictions of treatment response.

An overview of M-Learner

M-Learner is built on two key ideas: (1) it models polygenic efficacy as a function of latent genetic representations derived from high-dimensional PRS variables (Figure 1A), and (2) it employs a causal transfer learning framework that fine-tunes pre-trained PRS models with RCT data to predict treatment response. This design moves beyond conventional strategies that rely on a handful of candidate variants or a single PRS on the trial outcome, which does not fully capture the complex role of genetic background in HTE. By compressing millions of SNPs into PRS variables, M-Learner leverages external information from exceptionally-powered GWAS through transfer learning, and effectively reduces dimensionality and alleviates the computational challenge of modeling high-dimensional genetic data, thereby addressing the limited statistical power inherent to small RCTs.

We introduce two versions of M-Learner to accommodate different types of input data: M-Learner-S, which uses GxT summary statistics from genome-wide interaction studies (GWIS) in settings with only privacy-preserved summary-level data (Figure 1B), and M-Learner-I, which leverages individual-level genotype and treatment data for more sophisticated investigation (Figure 1C). Briefly, M-Learner-S first employs a polygenic interaction inference approach to conduct a hypothesis-free scan of PRS-by-treatment interactions from summary statistics [21]. The resulting interaction p-values can identify PRS that are informative on the HTE. Results across multiple PRS are then combined using the Cauchy combination test [23] to yield a global P-value. Alternatively, M-Learner-I first compresses millions of SNPs into a matrix of pre-trained PRS models [24, 25]. These PRS, together with treatment assignments and outcomes in the RCT, are then used to fit sophisticated machine learning models for predicting treatment efficacy. We use sequential inference to ensure valid statistical testing while avoiding overfitting [26, 27]. This method produces a global P-value for genetic modification of treatment response, identifies informative PRS through feature importance analysis, and calculates PES that predict how well each patient is expected to respond to the treatment which can be used to design personalized treatment strategies [28].

Scaling laws for HTE estimation with genetic information

Next, we derive power scaling laws for estimating HTE using genetic variation, focusing on the PES×T interaction coefficient. Here, PES is first estimated through machine learning and then included in the regression to assess its interaction with the treatment, using an independent RCT subsample for evaluation. Intuitively, when the estimated PES×T coefficient is close to zero, there is little statistical power to detect HTE; when it approaches one, the predictive model is well calibrated. The scaling laws provide theoretical guidance on the feasibility of detecting polygenic efficacy in RCTs and highlight the need to leverage pre-trained PRS models for reliable inference.

We compare three strategies for estimating PES: a within-sample GWIS approach, a single-outcome PRS approach, and a multi-PRS ensemble approach (detailed derivations are presented in the Methods section). The within-sample GWIS approach estimates SNP-by-treatment interaction coefficients directly within the RCT and uses them as SNP weights for the PES; naturally, limited sample size in the RCT makes these estimates extremely noisy, yielding lower power. The single-outcome PRS approach uses a pre-trained PRS for the trial outcome to assess the interaction with the treatment; its performance depends on the GWAS sample size which determines measurement error in the PRS, as well as the correlation between this PRS and the true PES (which is also equivalent to the correlation between $\beta_G j$ and $\beta_I j$). Finally, the multi-PRS ensemble approach leverages a high-dimensional PRS matrix pre-trained from many GWAS, fine-tuned in the RCT sample, to estimate a function that best predicts treatment response. This strategy effectively

combines multiple PRS that are informative on the PES and provides improved power to detect genetically-driven HTE.

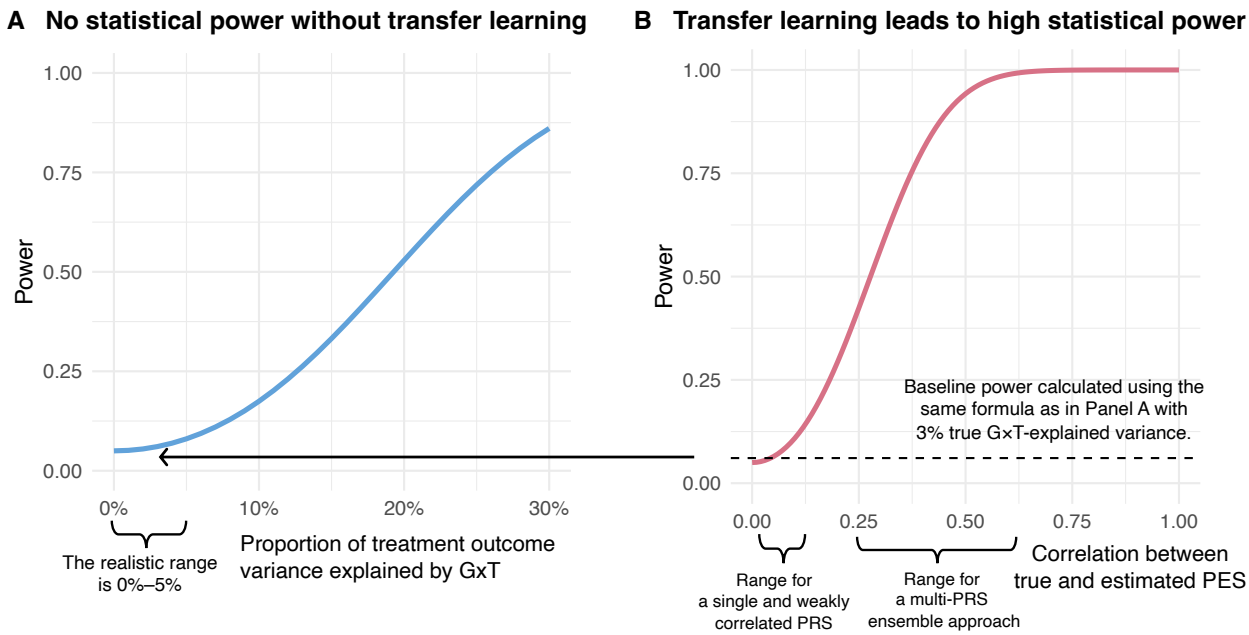

Figure 2: **Power scaling laws for detecting genetically-driven HTE. (A)** For analyses without transfer learning, the graph plots statistical power (y-axis) against the proportion of treatment response variance explained by gene-by-treatment interaction (x-axis). At the realistic range of this value(i.e., 0–5%), statistical power for detecting HTE is negligible. **(B)** With transfer learning, the graph plots power (y-axis) against the correlation between the true and estimated PES (x-axis). When the correlation reaches 0.5, we achieve >90% power. M-Learner achieves high power by aggregating information from a large number of pre-trained PRS.

Figure 2 illustrates these power scaling laws under realistic parameter settings: an RCT training and testing sample size of 2,500, a GWAS training sample size of 400,000, a heritability of 0.3 for the trial outcome, and an effective number of 60,000 SNPs. We also fix the type-I error rate at 0.05 in these curves. Panel A shows that without transfer learning (i.e., the within-sample GWIS approach), statistical power depends entirely on the proportion of treatment response variance explained by G×T. Because we expect this parameter to be small (around 0–5% in gene–environment interaction studies), statistical power of this approach is negligible. Panel B fixes the true G×T-explained variance at 3% and examines the effect of transfer learning. Here, power rises steeply as the estimated PES approaches its true value (quantified by the correlation between true and estimated PES), reaching 100% once the correlation reaches moderate levels. If a single PRS weakly correlated with the true PES is used, the analysis will have little power. In contrast, a multi-PRS ensemble approach aggregates information across multiple PRS to obtain a better estimate of the PES and, consequently, achieve higher power. In practice, M-Learner achieves this through an ensemble of PRS pre-trained from external GWAS, enabling detection of genetically-driven HTE even when the G×T signal is weak.

Simulation studies

We performed extensive simulations using genotype data from the UK Biobank [29] to compare M-Learner with alternative methods. These simulations were designed to reflect realistic genetic architectures and treatment response models (Methods) [30]. We assessed each method's ability to detect HTE, reporting statistical power and type I error rates.

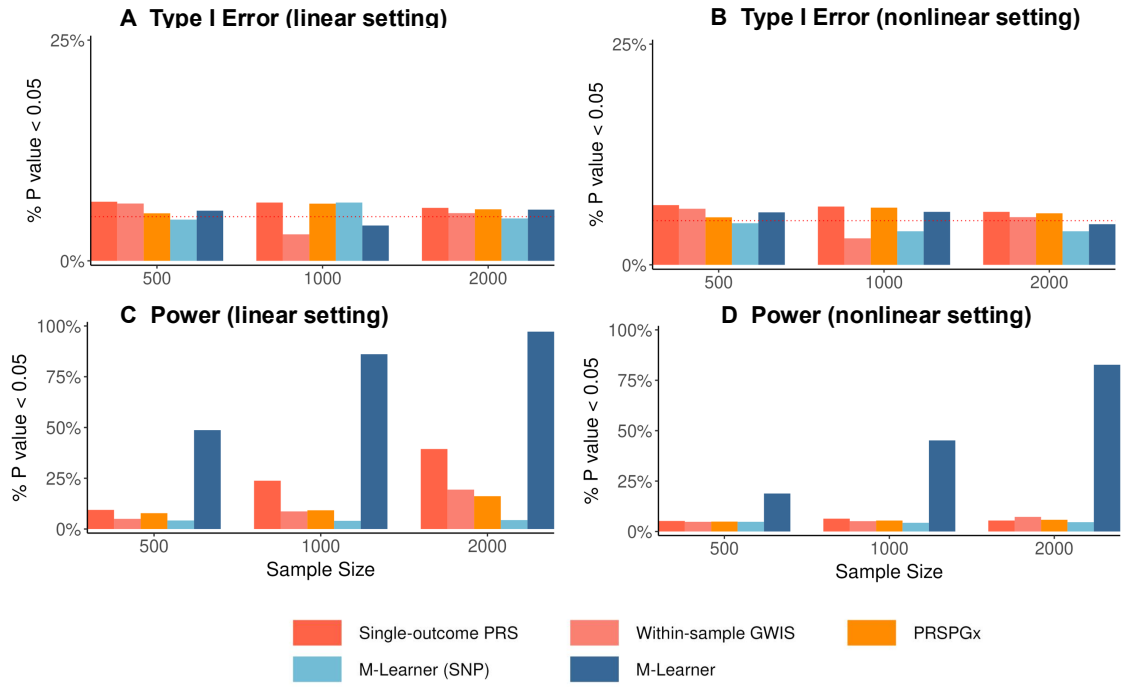

Figure 3: **Simulation results comparing M-Learner with alternative methods for detecting HTE.** Panels A and B show that all methods maintain well-calibrated type I error around the nominal 0.05 level under both linear (A) and nonlinear (B) settings. Panels C and D demonstrate that M-Learner achieves substantially higher power than alternative approaches across increasing sample sizes, with consistent gains under both linear (C) and nonlinear (D) settings. Competing methods exhibit limited power, especially in nonlinear scenarios, underscoring the advantage of M-Learner in detecting genetically driven heterogeneous treatment effects.

Specifically, we simulated a quantitative treatment outcome under a polygenic G×T interaction model with 200,000 SNPs. SNP effect sizes were drawn jointly from a multivariate normal distribution to capture linkage disequilibrium patterns observed in real data. The G×T term was specified in both linear and nonlinear forms. RCT sample sizes of 500, 1,000, and 2,000 individuals were examined. Each simulation scenario was repeated 500 times. Power was defined as the proportion of replicates rejecting the null hypothesis of no HTE at a significance level of 0.05, while type I error was evaluated under settings with no true HTE. We considered four alternative methods: (i) Single-outcome PRS, which uses the trial outcome PRS to approximate PES; (ii) Within-sample GWIS, which constructs the PES directly from the SNP-by-treatment interaction coefficients; (iii) PRSPGx [31], which constructs the PES as a weighted sum of the outcome PRS and GWIS, where the weights are optimized to maximize the prediction of treatment response in hold-out samples. (iv) M-Learner (SNP): which adapts M-Learner framework for estimating CATE, directly applied to the SNP genotype data without transfer learning through PRS. Here,

methods (i) and (ii) are two approaches we have compared with in the scaling law discussion, showcasing the performance of using either a single PRS on the outcome or a (noisy) PES estimated from RCT as the genetic modifier of treatment effect. Method (iii) is an approach in the literature which is a combination of methods (i) and (ii) without using high-dimensional PRS information or machine learning. Method (iv) explores the performance of M-Learner without external PRS models. This is a comprehensive list of available strategies for estimating HTE from genetic data, and serves as an important benchmark for M-Learner's performance.

All methods maintained well-calibrated type I error rates in both linear and nonlinear settings (Figure 3A-B), but M-Learner achieved substantially higher statistical power than other approaches in all settings. Among the alternatives, the single-outcome PRS approach demonstrated modest power in the linear setting but failed to capture nonlinear effects. The other three methods showed low power in all scenarios. Together, these findings demonstrate M-Learner's ability to identify genetically-driven HTE in small samples that are comparable to real-world RCTs, especially under a non-linear mapping from PRS to treatment effectiveness, where other approaches fail. These results also clearly suggest that the strategy to incorporate many PRS from external GWAS is a key methodological advance that improves HTE inference.

Genetic factors explain the HTE of secukinumab in ankylosing spondylitis

Next, we revisit the heterogeneous response to the anti-IL17 monoclonal antibody secukinumab in inflammatory diseases. A recent study failed to identify genetically-driven HTE, but released GWIS summary statistics for SNP-by-treatment interactions on four diseases (i.e., psoriasis, psoriatic arthritis, rheumatoid arthritis, and ankylosing spondylitis) from 19 RCTs (total $N_{treated}$=4,063, $N_{control}$=1,151; Supplementary Table 1) [11]. We applied M-Learner-S to investigate whether the HTE of secukinumab can be explained by genetic variation. Following previous work, we investigated four continuous measures of disease activity as the primary outcomes, including the disease activity score 28 with C-reactive protein (DAS28-CRP) for psoriatic arthritis and rheumatoid arthritis, the ankylosing spondylitis disease activity score with C-reactive protein (ASDAS-CRP), and the psoriasis area and severity index (PASI) score [11]. We used 70 external GWAS for a wide range of complex traits as inputs to M-Learner-S (Supplementary Table 2).

We replicated the previous study using conventional one-SNP-at-a-time subgroup analysis and risk score analysis based on the PRS of 15 immune diseases and found null results (Supplementary Figure 1-3 and Figure 4A). Next, we employed M-Learner to identify genetically-driven HTE. We identified significant HTE driven by genetic variation in the primary outcome for ankylosing spondylitis, i.e., ASDAS-CRP, after Bonferroni correction (omnibus P = 0.01) (Supplementary Figure xx and Figure 4). To follow up on this primary finding, we employed M-learner to examine two secondary outcomes for ankylosing spondylitis in the RCTs, i.e., C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR), and found highly significant HTE on CRP (P = 0.0001 for CRP and P = 0.086 for ESR). We did not identify significant HTE for psoriasis, psoriatic arthritis, or rheumatoid arthritis.

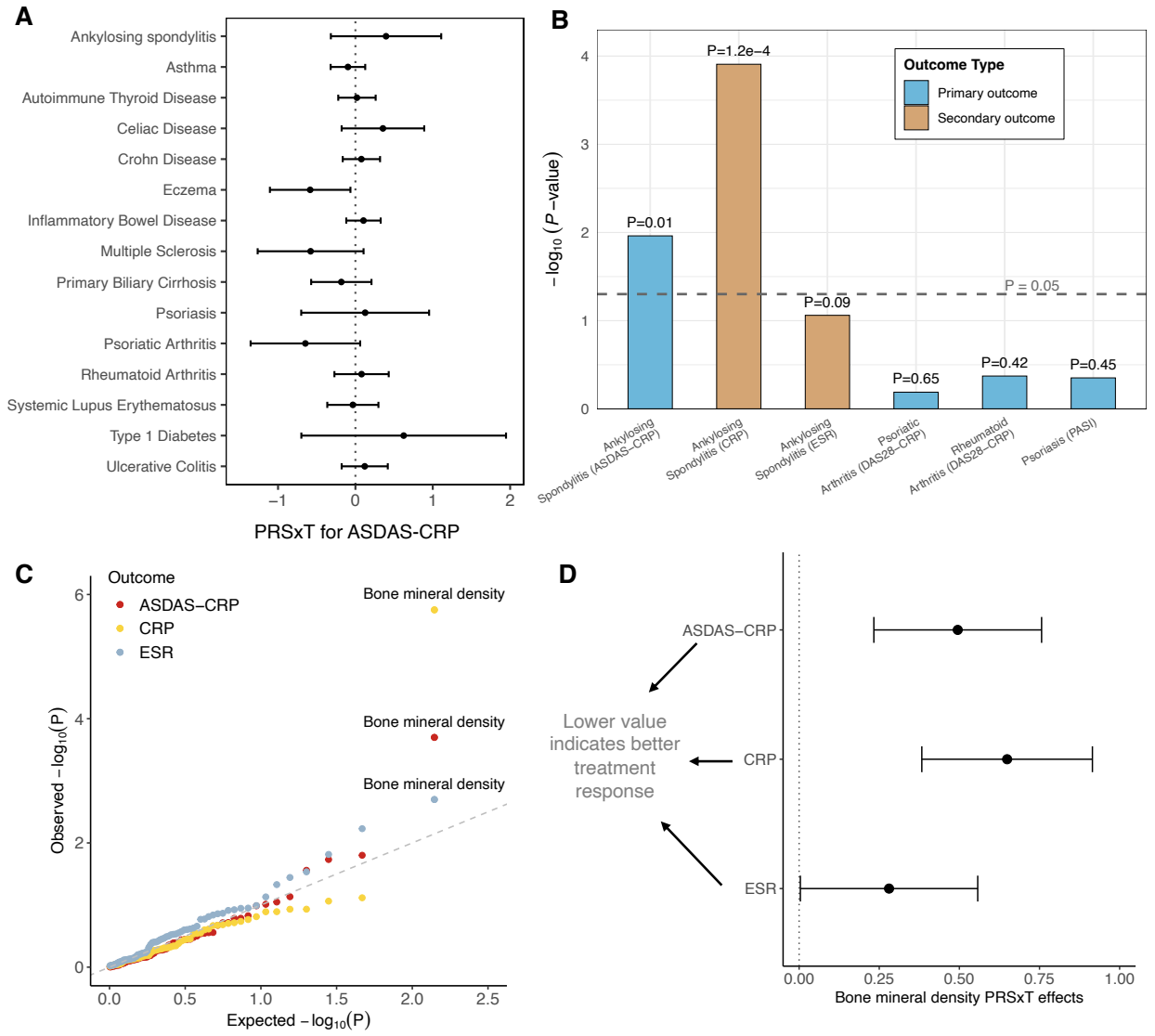

Figure 4: **M-Learner identifies the PRS of BMD as the top predictor of secukinumab HTE in ankylosing spondylitis. (A)** Risk score analysis based on PRS of 15 immune diseases yielded null results. **(B)** M-Learner detects significant genetically driven HTE for ankylosing spondylitis (P = 0.01 for ASDAS-CRP and P = 1.2e-4 for CRP) (Supplementary Table 3). **(C)** QQ plot for the PRSxT interaction p-values on three outcomes of ankylosing spondylitis. BMD PRS was a consistent strong predictor for HTE in all three outcomes. **(D)** Individuals genetically predisposed to lower BMD show greater benefit from anti-IL17 treatment for ankylosing spondylitis. ASDAS-CRP: Ankylosing spondylitis disease activity score with CRP; CRP: C-reactive protein levels; ESR: erythrocyte sedimentation rate.

Next, we sought to identify specific genetic traits that can explain the observed HTE in ankylosing spondylitis. The highest-ranked PRS was bone mineral density (BMD). This analysis suggests that ankylosing spondylitis patients who had lower PRS of bone mineral density (BMD) showed better outcomes after treatment, quantified by a greater reduction of ASDAS-CRP (estimated PRSxT coefficient = 0.495, P = 0.0002) (Figure 4 and Supplementary Table 3). This interaction was also found in the two secondary outcomes CRP (estimated PRSxT coefficient = 0.649,

P = 1.8e-6 and Supplementary Table 4) and ESR (estimated PRSxT coefficient = 0.281, P = 0.048) (Figure 4C and Supplementary Table 5). Interestingly, we note that the HTE of secukinumab on ankylosing spondylitis was not explained by the genetic predisposition to ankylosing spondylitis itself or by the trial outcome CRP. PRS of ankylosing spondylitis or CRP did not show significant interactions with the treatment on primary or secondary outcomes Supplementary Table 3-5). The interaction between BMD PRS and treatment remained significant after conditioning on the PRS of ankylosing spondylitis (estimated PRSxT coefficient = 0.477, P = 0.0007) and the PRS of CRP (estimated PRSxT coefficient = 0.903, P = 5.2e-5). This clearly demonstrates an important point we have made throughout the paper—risk score analysis based on a single PRS relies on a strong assumption that the PRS included in the analysis is correlated with the PES underlying HTE. In this example, although it seemed natural to use the PRS of ankylosing spondylitis (the disease of interest) or CRP (the trial outcome) as the risk score, they do not explain the HTE. The multi-PRS strategy implemented in M-Learner was able to leverage information from many other traits, and in this case, identified a specific interactor BMD that significantly modifies the effect of secukinumab on ankylosing spondylitis. We also note that this is a finding specific to ankylosing spondylitis. BMD did not modify the treatment response in the other three inflammatory diseases in our analysis (Figure 4).

The findings suggest that the higher treatment benefits to anti-IL17 therapy in patients genetically predisposed to lower bone mineral density (BMD) may be explained by the central role of IL17 in osteoimmunology. IL17 amplifies bone resorption by upregulating RANKL expression in osteoblasts and stromal cells, inducing downstream cytokines such as TNF, IL-1β, and IL-6, and reducing osteoprotegerin (OPG), thus increasing the RANKL:OPG ratio and promoting osteoclastogenesis [32]. Concurrently, IL17 perturbs osteoblast regulation by suppressing the Wnt inhibitors DKK1 and sclerostin, which enhances osteogenic differentiation but in an aberrant, inflammatory context that favors pathological syndesmophyte formation rather than healthy bone accrual [33]. Patients with genetic predispostion to low BMD—characterized by variants that increase RANKL or decrease OPG, and by heightened activity of Wnt inhibitors—are therefore particularly vulnerable to IL17–driven imbalances in bone remodeling [34]. In such individuals, IL17 inhibition not only suppresses the inflammatory cytokine network that fuels osteoclast activation but also restores normal regulation of osteoblast differentiation, leading to both improved systemic bone density and attenuation of pathological new bone formation. This mechanistic framework provides a plausible explanation for the observed HTE, wherein genetically low-BMD patients derive disproportionate clinical and skeletal benefit from anti-IL17 therapy.

Genetically-derived HTE of the bronchodilator treatment in the Lung Health Study

The Lung Health Study was an RCT conducted across multiple centers in the United States and Canada to evaluate whether a smoking intervention and the long-term use of an inhaled bronchodilator could slow the progressive decline of lung function in smokers with early-stage chronic obstructive pulmonary disease [35]. The trial concluded that smoking cessation substantially reduced the age-related loss of lung function, whereas the bronchodilator did not slow the decline of forced expiratory volume in one second (FEV1), suggesting the limited effectiveness of pharmacologic therapy compared to behavioral intervention [36].

We applied M-Learner to the Lung Health Study to revisit its null finding on the bronchodilator. We included 2,441 participants with both genotype and outcome data, comprising 1,249 in the treatment group (smoking intervention plus bronchodilator) and 1,192 in the control group (smoking intervention plus placebo). The outcome was defined as the cumulative change in

post-bronchodilator FEV1 over five years. We employed the single-outcome PRS approach and PRSPGx to benchmark their performance against M-Learner in identifying genetic factors explaining HTE. In addition, we applied M-Learner to non-genetic variables in the RCT to see if non-genetic factors can explain any HTE in the same trial.

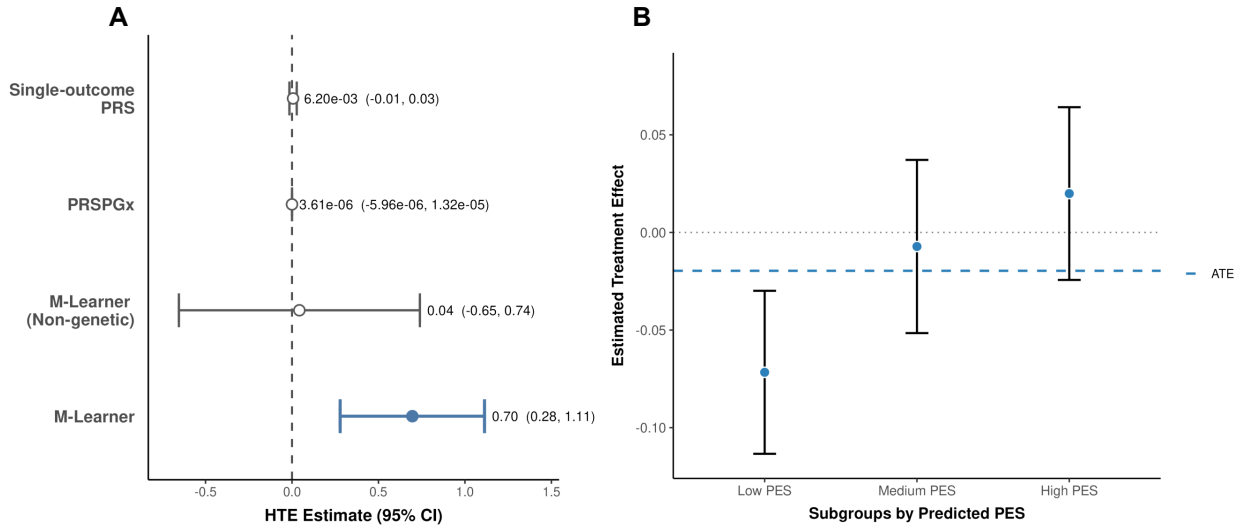

Figure 5: **M-Learner identifies HTE of the bronchodilator treatment in smokers. (A)** M-Learner improves the detection of HTE compared to alternative methods. M-Learner identified a significant HTE (P = 0.001). The single-outcome PRS approach, PRSPGx, and M-Learner using non-genetic variables all failed to identify HTE. **(B)** The average treatment in each subgroup defined by PES strata demonstrate a clear separation: individuals with low PES show negative treatment effects, while those with high PES show a positive response. In particular, individuals in the lowest PES tercile showed a significant but negative treatment effect (P = 0.0008), suggesting potential harm of the bronchodilator treatment in this group.

M-Learner identified significant evidence of HTE (P = 0.001), suggesting that genetic background can modify the response to the bronchodilator therapy (Figure 5A). To ensure robustness, we compared multiple learners—including support vector machine, random forest, elastic net, and neural network—within the same estimation framework (Supplementary Table 6). Support vector machine with a nonlinear radial basis function kernel achieved the best performance, highlighting the importance of employing nonlinear causal machine learning for HTE estimation. In comparison, the SNPxT GWIS, single-outcome PRS, and PRSPGx produced null results (Figure 5A and Supplementary Figure 4). When we applied M-Learner to only non-genetic features, the result was also null (Figure 5A and Supplementary Table 7).

In addition to detecting overall HTE, we stratified individuals according to their predicted PES and subsequently estimated the average treatment effect within each subgroup.(Figure 5B, Supplementary Table 8). The subgroup with the lowest PES exhibited a surprising but statistically significant treatment effect, which led to worse outcomes (P = 0.0008), suggesting potential harm of the bronchodilator therapy in this group. Although not all subgroups reached statistical significance, the treatment effect estimates in all three strata reflects a consistent trend, indicating that M-Learner effectively captured the variation in treatment response that was masked in the RCT analysis that only focused on the average treatment effect.

To further interpret the HTE, we examined the PRS that contributed strongly to M-Learner predictions. The PRS for standing height, FEV1, ever smoker, and alcohol use disorder, emerged as the most significant modifiers of bronchodilator treatment response. These findings indicate that genetic predispositions linked to lung function, substance use, and related physiological traits contribute to the variation in bronchodilator treatment efficacy. These results demonstrate that M-Learner extends beyond global detection of heterogeneity to pinpoint specific genetic mechanisms and pathways that can shape treatment outcomes.

3 Discussion

Understanding HTE is crucial for advancing precision medicine, but remains a challenge due to small RCT size, insufficient baseline measures of informative variables, and suboptimal statistical methodology. In this work, we envisioned and demonstrated the potential of using genetic variation to improve HTE inference. We also introduced M-Learner, a causal transfer learning framework that leverages many PRS models pre-trained on external GWAS data to detect and interpret HTE. Through theoretical derivations, numerical simulations, and applications to several RCTs with summary-level or individual-level genetic data, we demonstrated that M-Learner brings important advances to precision therapeutics.

We proposed two key advances to the current practice of HTE analysis. First, we argue that future RCTs should routinely incorporate genome-wide genetic variation of trial participants as predictive features, rather than solely relying on non-genetic covariates or a handful of candidate variants—a paradigm we term polygenic efficacy. Our vision is partly based on the fact that the cost for whole-genome genotyping or sequencing continues to drop. But more importantly, genetic information has two unique features that are not shared by other types of data: 1) it is fixed at conception and thus not affected by the treatment; 2) it is known to associate with a large collection of human traits, diseases, and behavior. The first feature is important because it enables systematic re-analysis of past RCTs. If genotype data can be generated for trial participants, even if this is done through a new wave of bio-sample collection, genetic variables generated from these samples can be treated as baseline, pre-treatment measures. This will benefit many RCTs, including the ones that have failed in the past due to HTE. If a subgroup of treatment responders can be identified based on genetic information, it could potentially revive failed trials and benefit patients in urgent need for more treatment options. The second feature of genetic information directly addresses the lack of variables that are informative on HTE in RCT data. The idea is that we can use genetic data to "impute" a large collection of human characteristics. Many PRS models have been developed in the literature which can nicely serve this purpose. From a machine learning perspective, this is an excellent example of transfer learning—we can borrow information from the numerous, well-powered GWAS to improve the learning task on HTE. This brings us to the second key advance we have introduced in this study. We argue that the estimation of HTE from genetic data should be grounded in a causal transfer learning framework that compresses millions of SNPs into PRS representations, which can then be fine-tuned with trial outcomes to capture the complex genetic modification of treatment response.

To operationalize these advances, we developed M-Learner in two forms: M-Learner-I, which uses individual-level genotypes, and M-Learner-S, which relies on summary statistics. M-Learner-I employs sophisticated machine learning to improve the estimation of HTE, even with limited RCT samples. M-Learner-S has broad applications when individual-level genotype data are unavailable from the trial, which is a common struggle in the field. We have demonstrated that through joint consideration of many pre-trained PRS, M-Learner maintains type I error control

while achieving markedly higher power compared to conventional approaches.

Applying M-Learner to RCTs on secukinumab revealed a striking role of BMD genetics in modifying the treatment effect for ankylosing spondylitis. Our results suggest that ankylosing spondylitis patients who are genetically predisposed to lower BMD may receive significantly greater benefit from the anti-IL17 therapy. We believe this is a profound finding for several reasons. First, it provides a clear example where the genetic trait that modifies treatment efficacy is not the genetic predisposition of the primary outcome. We have recently made similar comments on gene-environment interaction research [21], and are now making this observation for HTE inference. Many studies make the (incorrect) assumption that the risk score for the disease being studied is also the modifier of treatment efficacy. Future precision medicine research should expand the search for genetic modifiers into including a broader collection of complex trait candidates. The M-Learner framework provides a principled strategy to implement this idea in real-world RCTs. Second, BMD exhibits a highly specific modifier effect of secukinumab efficacy on ankylosing spondylitis, but not other inflammatory diseases. This observation is consistent with IL17's central role in osteoimmunology and provides a plausible mechanistic link between inflammatory cytokine signaling and pathological bone remodeling, offering a genetic rationale for prioritizing IL17 blockade in patients with specific skeletal risk profiles. Third, this finding was made using (summary-level) data produced from previous attempts that have generated largely null results. This highlights the importance of genomic data sharing as well as the need for advanced methods in precision medicine research, especially new strategies that can better handle summary-level genetic association data. Beyond secukinumab, M-Learner also revisited the Lung Health Study, identifying polygenic modifiers of bronchodilator response related to lung function, smoking, and anthropometric traits. Importantly, M-Learner produced PES that can be used to identify responder subgroups for the treatment (in this case, it was a subgroup that showed worse clinical outcomes). These results highlight the potential of using PES to inform trial design, improve sample recruitment, and more broadly, facilitate evidence-based clinical decision making. We believe this is an important step toward a future of healthcare many in the field have envisioned, in which personalized treatment recommendation becomes routine.

Taken together, we present a study that highlights the importance of adopting genetically-driven, hypothesis-free strategies in precision medicine research, and introduce a rigorous statistical framework named M-Learner to facilitate this type of investigation. We expect our findings to motivate important follow-up clinical investigations. We also anticipate M-Learner to have broad applications in genomic medicine and benefit numerous future studies.

4 Methods

Connection between the polygenic GxT model and CATE

We consider the following model for polygenic GxT effects [21]:

$$Y_i = \sum_{j=1}^{J} G_{ij}\beta_{Gj} + T_i\beta_T + \sum_{j=1}^{J} G_{ij}T_i\beta_{Ij} + \epsilon_i,$$

where Y_i is the treatment outcome for individual i, G_{ij} is the standardized genotype at variant j (mean 0, variance 1), $T_i \in \{\sqrt{\frac{1-p}{p}}, -\sqrt{\frac{p}{1-p}}\}$ is the standardized treatment indicator with treatment probability p, β_T is the treatment effect, β_{Gj} and β_{Ij} denote the additive and interaction effects for variant j, and ϵ_i is the error term.

Under this model, the ATE in an RCT can be denoted as

$$\begin{aligned}\text{ATE} &= \mathbb{E}[Y_i|T_i = \sqrt{\frac{1-p}{p}}] - \mathbb{E}[Y_i|T_i = -\sqrt{\frac{p}{1-p}}] \\ &= (\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}})\beta_T\end{aligned}$$

In addition, the CATE can be denoted as

$$\begin{aligned}\text{CATE}_i &= \mathbb{E}[Y_i|\mathbf{G}_i, T_i = \sqrt{\frac{1-p}{p}}] - \mathbb{E}[Y_i|\mathbf{G}_i, T_i = -\sqrt{\frac{p}{1-p}}] \\ &= (\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}})\beta_T + (\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}})\sum_{j=1}^{J} G_{ij}\beta_{Ij} \\ &= \text{ATE} + C \times \text{PES}_i,\end{aligned}$$

where $C = (\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}})$ and $\text{PES}_i = \sum_{j=1}^{J} G_{ij}\beta_{Ij}$. Therefore, the CATE is a linear function of PES. Inferring genetically-driven HTE is equivalent to estimating the PES under the polygenic GxT model.

M-Learner framework

We introduce a general framework beyond the linear GxT model we have shown above for assessing HTE using causal transfer learning. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control for i-th individual, respectively [37]. Let $T_i \in \{0, 1\}$ represent the binary treatment assignment. The covariate vector contains J SNPs denoted as $\mathbf{G}_i$. In M-learner, we will compress those J SNPs into K pre-trained PRS, $\widehat{\mathbf{PRS}}_i = \left(\widehat{\text{PRS}}_{1i}, \ldots, \widehat{\text{PRS}}_{Ki}\right) \in \mathbb{R}^K$, where each $\widehat{\text{PRS}}_{ki}$ summarizes the polygenic genetic predisposition for a k-th trait or disease. PRS can be constructed from external GWAS summary statistics using standard practice [16, 17]. We assume that treatment is randomly assigned conditional on $\mathbf{G}_i$, such that the propensity score $e(\mathbf{G}_i) = P(T_i = 1 \mid \mathbf{G}_i)$ is known and bounded away from 0 and 1. In this general setting, CATE for i-th individual is defined as

$$\tau(\mathbf{G}_i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{G}_i]$$

Since individual-level treatment effects are unobservable, we estimate $\tau(\mathbf{G}_i)$ using ML models combined with causal inference techniques [27].

We adopt a two-stage procedure to estimate CATE. In the first stage, we fit a predictive model using $\widehat{\mathbf{PRS}}_i$ as input to generate individual-level CATE estimates $\hat{\tau}(\mathbf{G}_i)$. This can be interpreted as the PES plus ATE under this general framework. Candidate learners include penalized linear models, support vector machines, random forests, and neural networks, chosen to flexibly capture nonlinear relationships between PRS and treatment response. In the second stage, we evaluate the predictiveness of these estimates using out-of-fold inference. This ensures that each $\hat{\tau}(\mathbf{G}_i)$ is obtained from a model trained without using the individual's own outcome data. This cross-fitting strategy mitigates overfitting and yields valid out-of-sample estimates of heterogeneity.

We introduce a pseudo outcome $Y_i^H = \frac{T_i Y_i}{e(\mathbf{G}_i)} - \frac{(1-T_i)Y_i}{1-e(\mathbf{G}_i)}$. It is well established that $\mathbb{E}\left[Y_i^H \mid \mathbf{G}_i\right] = \tau(\mathbf{G}_i)$ [1, 27], making Y_i^H an unbiased proxy for the CATE function. We then estimate predictive models using sequential cross-fitting rather than conventional K-fold cross-fitting [26]. Following the sequential validation framework of [38] and [26], the data are partitioned into ordered folds. On fold k, we fit the ML model using only the preceding folds $1, \ldots, k-1$ and then predict $\hat{\tau}(\mathbf{G}_i)$ on the current fold k. This sequential approach ensures that each prediction is strictly out-of-sample while pooling information across folds with statistical rigor.

To quantify predictiveness and correct for potential bias, we perform a best linear projection (BLP) of Y_i^H onto $\hat{\tau}(\mathbf{G})$

$$Y_i^H = \alpha + \beta \hat{\tau}(\mathbf{G}_i) + \varepsilon_i.$$

Here, the intercept α absorbs the overall ATE component, while the slope β captures how much variation in the pseudo-outcomes is explained by the predicted CATE.

- If $\hat{\tau}(\mathbf{G}_i)$ perfectly recovers the true CATE, then $\beta \approx 1$.
- If $\hat{\tau}(\mathbf{G}_i)$ contains only noise, then $\beta \approx 0$.
- Rejection of the null hypothesis $\beta = 0$ provides statistical evidence for HTE explained by genetic variation.

To examine how treatment effects differ across genetically defined risk profiles, we sort and stratify individuals by individuals by their predicted $\hat{\tau}(\mathbf{G}_i)$ and compute ATEs in each subgroup defined by quantiles of $\hat{\tau}(\mathbf{G}_i)$.

$$\tau_q = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid \hat{\tau}(\mathbf{G}_i) \in Q_q\right]$$

where Q_q denotes the q-th quantile bin of estimated $\hat{\tau}(\mathbf{G}_i)$. Increasing (or decreasing) τ_q across quantiles indicates a stronger (or weaker) benefit among individuals with higher predicted responses.

For interpretability, we investigate how each PRS feature individually relates to HTE. After obtaining out-of-fold $\hat{\tau}(\mathbf{G}_i)$, we perform a series of single-variable regressions:

$$\hat{\tau}(\mathbf{G}_i) = a_{0k} + a_k \widehat{\mathrm{PRS}}_{ki} + \eta_{ki}, \quad k = 1, \ldots, K.$$

Here, a_k captures the marginal association between $\widehat{\mathrm{PRS}}_{ki}$ and the predicted treatment effect. To quantify the strength of this association, we define the PRS importance score as the absolute

t-statistic from the regression:

$$\texttt{FeatureImportance}\left(\widehat{\text{PRS}}_{ki}\right) = \frac{|a_k|}{\text{SE}\left(a_k\right)}.$$

Larger values of $\texttt{FeatureImportance}\left(\widehat{\text{PRS}}_{ki}\right)$ indicate stronger evidence that $\widehat{\text{PRS}}_{ki}$ explains HTE. Ranking PRS by $\texttt{FeatureImportance}\left(\widehat{\text{PRS}}_{ki}\right)$ allows us to identify genetic traits most strongly associated with the variation in treatment response.

The M-Learner algorithm proceeds with individual-level data as follows:

Algorithm 1: M-Learner-I Algorithm using Individual-level data as input

---

Input: Dataset $\{(\mathbf{G}_i, T_i, Y_i)\}_{i=1}^n$ with genotype $\mathbf{G}_i$; propensity model $e(\cdot)$; number of folds K; a machine learning learner $\mathcal{L}$

Output: P-value for testing the null hypothesis H_0; The average treatment in each subgroups by predicted PES; PRS feature importance scores

Step 1: Calculate PRS from SNP data to obtain $\widehat{\mathbf{PRS}}_i = \left(\widehat{\text{PRS}}_{1i}, \ldots, \widehat{\text{PRS}}_{Ki}\right) \in \mathbb{R}^K$ for each individual.

Step 2: Pseudo-outcomes. Compute

$$Y_i^H \;=\; \frac{T_i Y_i}{e(\mathbf{G}_i)} - \frac{(1-T_i)Y_i}{1-e(\mathbf{G}_i)}.$$

Step 3: Sequential cross-fitting over a K-fold partition.
Randomly partition $\{1, \ldots, n\}$ into K folds $\{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$, then, for $k = 2, \ldots, K$:

1. *Train on past folds:* fit the ML model on $\{(\widehat{\mathbf{PRS}}_i, Y_{ji}^H) : j \in \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_{k-1}\}$ to obtain $f_{\hat{\theta}}^{(1:(k-1))}$.
2. *Predict on current fold:* for each $i \in \mathcal{I}_k$, set $\hat{\tau}(\mathbf{G}_i) \leftarrow f_{\hat{\theta}}^{(1:(k-1))}(\text{PRS}_i)$.

Step 4: Assess heterogeneity by the best linear prediction. Regress Y_i^H on $\hat{\tau}(\mathbf{G}_i)$:

$$Y_i^H = \alpha + \beta\hat{\tau}(\mathbf{G}_i) + \varepsilon_i.$$

Test $H_0 : \beta = 0$.

Step 5: Treatment effect stratification. Sort individuals by $\hat{\tau}(\mathbf{G}_i)$; divide into quantile bins $\{Q_q\}$; compute:

$$\tau_q = \mathbb{E}[Y_i(1) - Y_i(0) \mid \hat{\tau}(\mathbf{G}_i) \in Q_q].$$

Step 6: PRS importance. For each PRS_k:

1. Fit single-variable regression

$$\hat{\tau}(\mathbf{G}_i) = a_{0k} + a_k\widehat{\text{PRS}}_{ki} + \eta_{ki}.$$

2. Compute importance score $\texttt{FeatureImportance}(\widehat{\text{PRS}}_{ki}) = |a_k|/\text{SE}(a_k)$.

The M-Learner algorithm proceeds with summary statistics as follows:

Algorithm 2: M-Learner-S Algorithm using GWIS summary statistics as input

Input: Z-score from GWIS summary statistics $\texttt{GWIS} = \{Z_{Ij} \text{ for } j = 1, \cdots, J\}$, Z-score from $k = 1, \cdots, K$ GWAS summary statistics $\texttt{GWAS}_k = \{Z_{kj} \text{ for } j = 1, \cdots, J\}$. Here, J is the total number of SNPs

Output: P-value for testing the null hypothesis H_0: treatment effect does not differ from genetic variation. A ranked list of PRS (GWAS) that explains the HTE.

Step 1: Apply PIGEON to each GWAS summary staistics $(k = 1, \cdots, K)$ **to compute the PRS×T interaction P-value:**

$$\text{P-value}_k = \texttt{PIGEON}(\texttt{GWIS}, \texttt{GWAS}_k) \text{ for } k = 1, \cdots, K$$

Step 2: Apply Cauchy-combination test to obtain the aggregated p-value
Define the Cauchy combination statistic

$$T_K = \sum_{k=1}^{K} \frac{1}{K} \tan\left[\left(\frac{1}{2} - \text{P-value}_k\right)\pi\right].$$

Under the global null (no HTE), the combined P-value is

$$\text{P-value}_{\text{CCT}} = \frac{1}{2} - \frac{1}{\pi}\arctan\left(T_K\right).$$

Step 3: Ranking the PRS based on the P-value
Let $\text{P-value}_1, \text{P-value}_2, \ldots, \text{P-value}_K$ be the set of PRSxT interaction P-values obtained from Step 1. Define the ordering function q such that

$$\text{P-value}_{q(1)} \leq \text{P-value}_{q(2)} \leq \cdots \leq \text{P-value}_{q(K)}$$

Then the ranked list of PRS corresponds to

$$\text{GWAS}_q(1), \text{GWAS}_q(2), \ldots, \text{GWAS}_{q(K)},$$

where $\text{GWAS}_q(1)$ is the top-ranked GWAS (smallest P-value) and GWAS $q(K)$ is the lowest-ranked.

Derivation for the scaling laws for HTE estimation with genetic variations

We derive the power scaling laws based on the polygenic GxT model. Suppose we have N_{test} testing samples for quantifying the existence of HTE by linear interaction regression

$$Y_i \sim S_i\alpha_S + T_i\alpha_T + S_iT_i\alpha_I + \delta_i,$$

where S_i is the estimated PES. We use a general notation for S_i to emphasize that the PES can be estimated in various ways as described below.

We consider different S_i and evaluate the expected effect size magnitude and statistical power

of hypothesis testing of $H_0 : \alpha_I = 0$ in the test data. We have

$$\begin{aligned}
\alpha_I &= \frac{\mathrm{Cov}(Y_i, S_i T_i)}{\mathrm{Var}(S_i T_i)} \\
&= \frac{\mathrm{Cov}(PES_i T_i, S_i T_i)}{\mathrm{Var}(S_i T_i)} \\
&= \frac{\mathbb{E}[PES_i S_i]}{\mathrm{Var}(S_i)} \\
&= \frac{\mathrm{Cov}(PES_i, S_i)}{\mathrm{Var}(S_i)}
\end{aligned}$$

We adopted the random-effects assumption used in the PIGEON paper [21], which is a standard approach in GWAS and G×E methodology studies. This assumption is required only for deriving the theoretical power-scaling laws, but not for other analyses in this manuscript.6 The variance-covariance matrix between true PES PES_i, the estimated PES using within-sample GWIS approach $\widehat{PES}_i^{\texttt{within-sample GWIS}}$ with N_{RCT} samples, and the the estimated PES using a single PRS approach $\widehat{PES}_i^{\texttt{single-PRS}} = \widehat{PRS}_{1i} := PRS_{1i} + \gamma$ are

$$\mathrm{Var}\left(\begin{bmatrix} PES_i \\ \widehat{PES}_i^{\texttt{within-sample GWIS}} \\ \widehat{PES}_i^{\texttt{single-PRS}} \end{bmatrix}\right) = \begin{bmatrix} \sigma_I^2 & \sigma_I^2 & r_{GI1}\sqrt{\sigma_I^2\sigma_G^2} \\ \sigma_I^2 & \sigma_I^2 + \frac{M_e}{N_{RCT}} & r_{GI1}\sqrt{\sigma_I^2\sigma_G^2} \\ r_{GI1}\sqrt{\sigma_I^2\sigma_G^2} & r_{GI1}\sqrt{\sigma_I^2\sigma_G^2} & \sigma_{G1}^2 + \frac{M_e}{N_{G1}} \end{bmatrix},$$

where σ_I^2 is the proportion of the variance of the treatment response explained by GxT interactions, σ_G^2 is the heritability for the outcome trait, r_{GI1} is the correlation between true PES, PES_i, and PRS, PRS_{1i}, and M_e is the effective number of SNPs, γ is the measurement error in estimating PRS.

Then, if we plug in the estimated PES (i.e. use the SNPs as input and use the estimated interaction coefficient to weight the SNP), we have

$$\mathbb{E}[\widehat{\alpha}_I^{\texttt{within-sample GWIS}}] = \frac{\sigma_I^2}{\sigma_I^2 + \frac{M_e}{N_{RCT}}}$$

Suppose the $N_{RCT} = 5000$, $\sigma_I^2 = 0.01$, and the tyical number for M_e is 60k.

If we plug in the estimated PRS, we have

$$\mathbb{E}[\widehat{\alpha}_I^{\texttt{single-PRS}}] = \frac{r_{GI1}\sqrt{\sigma_I^2\sigma_G^2}}{\sigma_{G1}^2 + \frac{M_e}{N_{G1}}}$$

If we have additional K-1 PRS, i.e., $\widehat{PRS}_{2i}, \cdots, \widehat{PRS}_{Ki}$, under the same random effect assumption, we have

$$\mathbb{E}[\widehat{\alpha}_I^{\texttt{M-Learner}}] = \sqrt{\frac{\Sigma_{WX}^T \Sigma_{XX}^{-1} \Sigma_{WX}}{\sigma_W^2}},$$

where

$$\text{Var}\left(\begin{bmatrix} PES_i \\ \widehat{PRS}_{1i} \\ \vdots \\ \widehat{PRS}_{Ki} \end{bmatrix}\right) = \begin{bmatrix} \sigma_W^2 & \Sigma_{WX}^{\mathrm{T}} \\ \Sigma_{WX} & \Sigma_{XX} \end{bmatrix}$$

We then have $\mathbb{E}[\alpha_I^{\texttt{M-Learner}}] \geq \mathbb{E}[\alpha_I^{\texttt{single-PRS}}]$.

The statistical power for a two-sided test for $H_0 : \alpha_I = 0$ can be derived as

$$\text{Power} = 1 - \Phi(z_{1-\alpha/2} - \lambda) + \Phi(-z_{1-\alpha/2} - \lambda),$$

where $\lambda = \dfrac{\sqrt{N_{test}}\sigma_I^2}{\sqrt{\sigma_I^2 + \frac{M_e}{N_{RCT}}}}$ for estimated PES, $\lambda = \dfrac{\sqrt{N_{test}} r_{GI1}\sqrt{\sigma_I^2\sigma_G^2}}{\sqrt{\sigma_{G1}^2 + \frac{M_e}{N_{G1}}}}$ for estimated PRS, and $\lambda = \dfrac{\sqrt{N_{test}}\Sigma_{YX}^T\Sigma_{XX}^{-1}\Sigma_{YX}}{\sigma_Y^2}$ for M-Learner.

Simulations

Each simulation scenario was repeated 500 times for robust inference. We restricted the analysis to autosomal SNPs that exist in the HapMap3 panel and 1000 Genomes Project. We further filtered SNPs with an imputation quality score greater than 0.9, minor allele frequency (MAF) $\geq$ 0.05, missing call rate $\leq$ 0.01, and Hardy–Weinberg equilibrium test p-value $\geq$ 1.0e-6. We simulated treatment outcome data using 200,000 randomly selected single-nucleotide polymorphisms (SNPs), with effect sizes drawn jointly from a normal distribution. A quantitative phenotype was generated according to

$$Y_i = \text{PRS}_i^{\text{baseline}} + \text{PES}_i T_i^{\text{std}} + \sqrt{0.01} T_i^{\text{std}} + \epsilon_i$$

where the binary treatment $T_i \sim \text{Bernoulli}(0.5)$ was standardized to T_i^{std} to have a mean of 0 and variance of 1, and ϵ_i denotes Gaussian noise scaled such that $\text{Var}(Y_i) = 1$. The term $\text{PRS}_i^{\text{baseline}} = \sum_j^J \beta_{Gj} G_{ij}$ represents the polygenic additive effect that contribute 10% of the treatment outcome, modeled as a single baseline PRS, whereas PES_i specifies the true PES, which can be linear or non-linear function of the SNP. To simulate realistic genetic correlation structures, we further constructed 25 auxiliary true PRS, PRS_{ij} for $j = 1, ..., 25$. Each PRS is simulated with heritability from $N(0.5, 0.05)$, and the 25 PRS have correlation coefficients ranging linearly from -0.8 to 0.8. Each PRS is then estimated from a GWAS of 200,000 individuals. The sample size of RCT varies across three scenarios: 500, 1000, and 2000.

1. Linear setting: the true PES was modeled as a weighted sum of the baseline and auxiliary PRS:

$$\text{PES}_i = \sqrt{0.1}\left(0.1\text{PRS}_i^{\text{baseline}} + \sum_{k=1}^{25} \alpha_k \text{PRS}_{ki}\right),$$

   where the weights α_k were equally spaced between -0.6 and 0.6.

2. Nonlinear setting: we replaced the linear combination of auxiliary PRS with a nonlinear transformation:

$$\mathrm{PES}_i = \sqrt{0.1}\left(0.1\mathrm{PRS}_i^{\mathrm{baseline}} + \texttt{scale}(f)\right),$$

where $f = 5\mathrm{PRS}_1 + \mathrm{PRS}_2\mathrm{PRS}_3 + \sin(\mathrm{PRS}_4) + \cos(\mathrm{PRS}_5) + e^{\mathrm{PRS}_6} + \log(|\mathrm{PRS}_7| + 1) + \mathrm{PRS}_8\mathrm{PRS}_9 - \mathrm{PRS}_{10}^3 + 0.1\sum_{j=11}^{25}\mathrm{PRS}_j^2$, and $\texttt{scale}(f)$ indicates scale f to have a mean of 0 and variance of 1. This formulation captures complex, nonlinear SNP-by-treatment interactions that mimic realistic biological heterogeneity.

We evaluated both statistical power (the probability of rejecting the null hypothesis of no HTE at $\alpha = 0.05$) and type I error control under null scenarios with no heterogeneity.

Analysis of HTE for anti-IL17 therapy in inflammatory diseases

We reanalyzed the SNP-by-treatment GWIS summary statistics derived from clinical and genetic data from 19 RCTs of the anti-IL17 monoclonal antibody secukinumab. The trials covered four indications: psoriasis, psoriatic arthritis, ankylosing spondylitis, and rheumatoid arthritis . After quality control, the dataset included 5,218 individuals (4,063 treated with secukinumab and 1,151 placebo controls).

The primary outcomes were continuous disease activity measures specific to each indication: PASI for psoriasis, DAS28-CRP for psoriatic arthritis and rheumatoid arthritis, and ASDAS-CRP for ankylosing spondylitis. For each participant, we used the change from baseline in outcome score at the primary assessment time point (week 12 for psoriasis; week 16 for psoriatic arthritis, ankylosing spondylitis, and rheumatoid arthritis). Secondary outcomes included laboratory measures, clinician- and patient-reported assessments, and composite indices as described in the original study [11].

For each indication, the original study performed a GxT genome-wide interaction study (GWIS) using linear regression. Models included treatment (secukinumab vs placebo), genotype dosage, the SNP-by-treatment interaction (the primary term of interest), and covariates (age, sex, BMI, baseline disease activity, three ancestry principal components, and trial-specific design variables). For the present study, we reanalyzed the GWIS summary statistics generated in the original analyses rather than individual-level genotype data. Summary statistics included variant-level regression coefficients, standard errors, and p-values for the GxT interaction terms across all four indications.

Data analysis in Lung Health Study

We analyzed individual-level clinical and genetic data from the Lung Health Study (LHS), which originally enrolled 5,887 smokers aged 35-60 years with early-stage chronic obstructive pulmonary disease [35]. For our analysis, we included 2,441 participants who had both genotype data and complete five-year follow-up of postbronchodilator FEV1. Participants with missing genetic information or incomplete outcome data were excluded, and the separate LHS cohort without smoking intervention was not considered. The primary outcome was defined as the cumulative change in post-bronchodilator FEV1 over five years.

We compared M-Learner to three baseline approaches commonly used for evaluating heterogeneity in treatment response in RCT data.

- a single-outcome PRS approach. We fit a regression model including treatment assignment, the PRS derived from GWAS of FEV1 (the same outcome measured in the trial), and their

interaction:

$$Y_i = \alpha + \beta_T \text{Treatment}_i + \beta_G \text{PRS}_i^{FEV1} + \beta_I \left(\text{Treatment}_i \times \text{PRS}_i^{FEV1} \right) + \varepsilon_i,$$

This model tests whether genetic liability to FEV1 modifies treatment response.

- We also applied PRS-PGx [31], a recently developed framework for pharmacogenomic prediction. At the SNP level, PRS-PGx assumes the model

$$Y_i = \alpha + \beta_T T_i + \beta_G G_i + \beta_I (G_i \times T_i) + \varepsilon_i$$

where G_i is genotype for i-th individual, β_G and β_I are the main-effect and interaction-effect coefficients estimated from GWIS. PRS-PGx then aggregates these SNP-level effects into two polygenic scores: the prognostic score, defined as the weighted sum of genotypes using β_G, which captures genetic predisposition to the outcome independent of treatment; and the predictive score, defined as the weighted sum of genotypes using β_I, which captures genetic predisposition to differential treatment response. These two scores are conceptually similar to PRS and PES we have discussed throughout the paper. The two scores are then incorporated into the following regression model

$$Y_i = \alpha + \beta_T T_i + \beta_P \text{ProgScore}_i + \beta_I (T_i \times \text{PredScore}_i) + \varepsilon_i,$$

which jointly evaluates the contribution of genetic background to both baseline prognosis and treatment modification. We implemented the GWIS and obtained summary statistics using plink 2.0 [30].

- M-Learner (Non-genetic). As a baseline, we implemented M-Learner using only non-genetic features. In this setting, the predictors included approximately 40 baseline demographic and clinical variables, such as age, sex, baseline lung function measures, respiratory symptom scores, and other spirometry- and questionnaire-based indicators of respiratory health. These features were combined with treatment assignment to model heterogeneous treatment effects, without incorporating any genetic information.

In addition to these baselines, we applied M-Learner using 4,739 pre-trained PRS from the PGS Catalog [24] as high-dimensional representations of germline genetic variation. These PRS were integrated with randomized treatment assignment and outcome data through sequential cross-validation to estimate heterogeneous treatment effects and identify the most predictive features.

Code availability

M-Learner is publicly available at https://github.com/qlu-lab/mlearner

Data availability

The GWAS and GWIS summary statistics used in this study are publicly available at https://qlu-lab.org/data.html. The SNP-by-treatment GWIS summary statistics for the anti-IL17 monoclonal antibody secukinumab in inflammatory diseases are available in the GWAS Catalog (https://www.ebi.ac.uk/gwas/, accession numbers GCST90274734 to GCST90274752). Data from the Lung Health studies are available by application at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000291.v2.p1.

Author contributions

J.Miao., J.Mu., and Q.L. conceived and designed the study.
J.Miao. and J.Mu. developed the statistical framework.
J.Miao., J.Mu., and X.Y. performed the data analysis.
J.F. and L.S. advised on interaction inference and result interpretation.
Q.L. advised on statistical and genetic issues.
J.Miao., J.Mu., and Q.L. wrote the manuscript.
All authors contributed to manuscript editing and approved the manuscript.

Acknowledgments

We acknowledge research support from National Institutes of Health (NIH) grant U01 HG012039, the University of Wisconsin-Madison Office of the Chancellor, and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (WARF).

References

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

[2] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[3] Issa J Dahabreh and Dhruv S Kazi. Toward personalizing care: assessing heterogeneity of treatment effects in randomized trials. *JAMA*, 329(13):1063–1065, 2023.

[4] Jesse J Swen, Cathelijne H van der Wouden, Lisanne EN Manson, Heshu Abdullah-Koolmees, Kathrin Blagec, Tanja Blagus, Stefan Böhringer, Anne Cambon-Thomsen, Erika Cecchin, Ka-Chun Cheung, et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. *The Lancet*, 401(10374):347–356, 2023.

[5] Ewan C Goligher, Patrick R Lawler, Thomas P Jensen, Victor Talisa, Lindsay R Berry, Elizabeth Lorenzi, Bryan J McVerry, Chung-Chou Ho Chang, Eric Leifer, Charlotte Bradbury, et al. Heterogeneous treatment effects of therapeutic-dose heparin in patients hospitalized for covid-19. *Jama*, 329(13):1066–1077, 2023.

[6] Jillian Frieder, Dario Kivelevitch, and Alan Menter. Secukinumab: a review of the anti-il-17a biologic for the treatment of psoriasis. *Therapeutic advances in chronic disease*, 9(1):5–21, 2018.

[7] Richard G Langley, Boni E Elewski, Mark Lebwohl, Kristian Reich, Christopher EM Griffiths, Kim Papp, Lluís Puig, Hidemi Nakagawa, Lynda Spelman, Bárður Sigurgeirsson, et al. Secukinumab in plaque psoriasis—results of two phase 3 trials. *New England Journal of Medicine*, 371(4):326–338, 2014.

[8] Philip Mease, Désirée van der Heijde, Robert Landewé, Shephard Mpofu, Proton Rahman, Hasan Tahir, Atul Singhal, Elke Boettcher, Sandra Navarra, Karin Meiser, et al. Secukinumab improves active psoriatic arthritis symptoms and inhibits radiographic progression: primary results from the randomised, double-blind, phase iii future 5 study. *Annals of the rheumatic diseases*, 77(6):890–897, 2018.

[9] Karel Pavelka, Alan Kivitz, Eva Dokoupilova, Ricardo Blanco, Marco Maradiaga, Hasan Tahir, Luminita Pricop, Mats Andersson, Aimee Readie, and Brian Porter. Efficacy, safety, and tolerability of secukinumab in patients with active ankylosing spondylitis: a randomized, double-blind phase 3 study, measure 3. *Arthritis research & therapy*, 19(1):285, 2017.

[10] Dominique Baeten, Joachim Sieper, Jürgen Braun, Xenofon Baraliakos, Maxime Dougados, Paul Emery, Atul Deodhar, Brian Porter, Ruvie Martin, Mats Andersson, et al. Secukinumab, an interleukin-17a inhibitor, in ankylosing spondylitis. *New England journal of medicine*, 373(26):2534–2548, 2015.

[11] Cong Zhang, Konstantin Shestopaloff, Benjamin Hollis, Chun Hei Kwok, Claudia Hon, Nicole Hartmann, Chengeng Tian, Magdalena Wozniak, Luis Santos, Dominique West, et al. Response to anti-il17 therapy in inflammatory disease is not strongly impacted by genetic background. *The American Journal of Human Genetics*, 110(10):1817–1824, 2023.

[12] Guanbo Wang, Patrick J Heagerty, and Issa J Dahabreh. Using effect scores to characterize heterogeneity of treatment effects. *JAMA*, 331(14):1225–1226, 2024.

[13] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[14] Richard M Weinshilboum and Liewei Wang. Pharmacogenomics: precision medicine and drug response. In *Mayo Clinic Proceedings*, volume 92, pages 1711–1722. Elsevier, 2017.

[15] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, 2018.

[16] Jiacheng Miao, Hanmin Guo, Gefei Song, Zijie Zhao, Lin Hou, and Qiongshi Lu. Quantifying portable genetic effects and improving cross-ancestry genetic prediction with gwas summary statistics. *Nature Communications*, 14(1):832, 2023.

[17] Zijie Zhao, Tim Gruenloh, Meiyi Yan, Yixuan Wu, Zhongxuan Sun, Jiacheng Miao, Yuchang Wu, Jie Song, and Qiongshi Lu. Optimizing and benchmarking polygenic risk scores with gwas summary statistics. *Genome Biology*, 25(1):260, 2024.

[18] Amy Damask, P Gabriel Steg, Gregory G Schwartz, Michael Szarek, Emil Hagström, Lina Badimon, M John Chapman, Catherine Boileau, Sotirios Tsimikas, Henry N Ginsberg, et al. Patients with high genome-wide polygenic risk scores for coronary artery disease may receive greater clinical benefit from alirocumab treatment in the odyssey outcomes trial. *Circulation*, 141(8):624–636, 2020.

[19] Pradeep Natarajan, Robin Young, Nathan O Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F Reilly, Adam Butterworth, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135(22):2091–2101, 2017.

[20] Nicholas A Marston, Frederick K Kamanu, Francesco Nordio, Yared Gurmu, Carolina Roselli, Peter S Sever, Terje R Pedersen, Anthony C Keech, Huei Wang, Armando Lira Pineda, et al. Predicting benefit from evolocumab therapy in patients with atherosclerotic disease using a genetic risk score: results from the fourier trial. *Circulation*, 141(8):616–623, 2020.

[21] Jiacheng Miao, Gefei Song, Yixuan Wu, Jiaxin Hu, Yuchang Wu, Shubhashrita Basu, James S Andrews, Katherine Schaumberg, Jason M Fletcher, Lauren L Schmitz, et al. Pigeon: a statistical framework for estimating gene–environment interaction for polygenic traits. *Nature Human Behaviour*, pages 1–15, 2025.

[22] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[23] Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[24] Samuel A Lambert, Laurent Gil, Simon Jupp, Scott C Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nature genetics*, 53(4):420–425, 2021.

[25] Buu Truong, Leland E Hull, Yunfeng Ruan, Qin Qin Huang, Whitney Hornsby, Hilary Martin, David A Van Heel, Ying Wang, Alicia R Martin, S Hong Lee, et al. Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genomics*, 4(4), 2024.

[26] Stefan Wager. Sequential validation of treatment heterogeneity. *arXiv preprint arXiv:2405.05534*, 2024.

[27] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica*, 93(4):1121–1164, 2025.

[28] Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping, et al. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592(7855):629–633, 2021.

[29] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[30] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

[31] Song Zhai, Hong Zhang, Devan Mehrotra, and Judong Shen. Pharmacogenomics polygenic risk score for drug response prediction using prs-pgx methods. *Nature Communications*, 13:5278, 09 2022.

[32] Danping Lin, Lu Li, Ying Sun, Weidong Wang, Xiaoqian Wang, Yu Ye, Xu Chen, and Yan Xu. Interleukin-17 regulates the expressions of rankl and opg in human periodontal ligament cells via traf 6/tbk 1-jnk/nf-κb pathways. *Immunology*, 144(3):472–485, 2015.

[33] Mengjia Tang, Lingyun Lu, and Xijie Yu. Interleukin-17a interweaves the skeletal and immune systems. *Frontiers in Immunology*, 11:625034, 2021.

[34] Nicole M Warrington, John P Kemp, Kate Tilling, Jonathan H Tobias, and David M Evans. Genetic variants in adult bone mineral density and fracture risk genes are associated with the rate of bone mineral density acquisition in adolescence. *Human molecular genetics*, 24(14):4158–4166, 2015.

[35] John E. Connett, John W. Kusek, William C. Bailey, Peggy O'Hara, and Margaret Wu. Design of the lung health study: A randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Controlled Clinical Trials*, 14(2, Supplement):3–19, 1993.

[36] Nicholas R Anthonisen, John E Connett, James P Kiley, Murray D Altose, William C Bailey, A Sonia Buist, William A Conway, Paul L Enright, Richard E Kanner, Peggy O'hara, et al. Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of fev1: the lung health study. *Jama*, 272(19):1497–1505, 1994.

[37] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

[38] Alexander R. Luedtke and Mark J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2), April 2016.

Supplementary Information

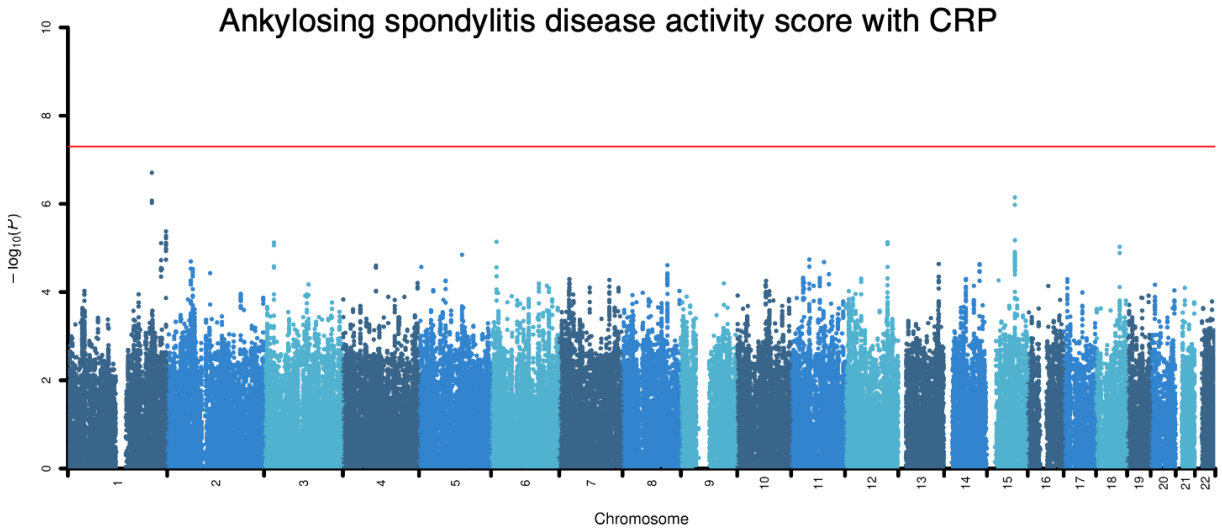

Figure S1. Manhattan plot for genome-wide genotype-by-treatment interaction results. The treatment is anti-IL17 therapy. The outcome is the ankylosing spondylitis disease activity score with CRP. The red line means P = 5e-8.

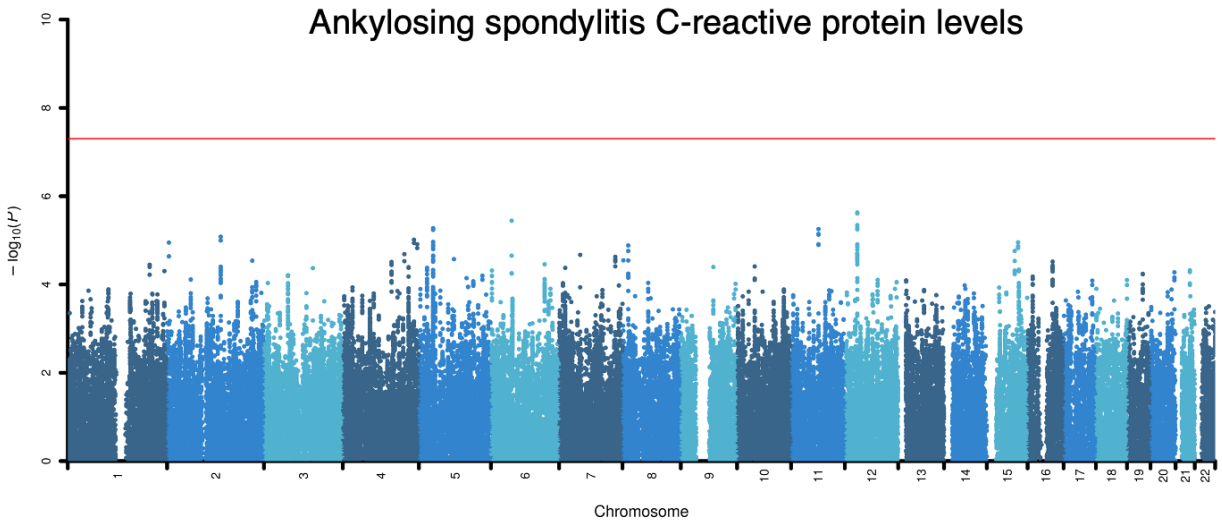

Figure S2. Manhattan plot for genome-wide genotype-by-treatment interaction results. The treatment is anti-IL17 therapy. The outcome is the ankylosing spondylitis C-reactive protein levels. The red line means P = 5e-8. CRP: C-reactive protein levels

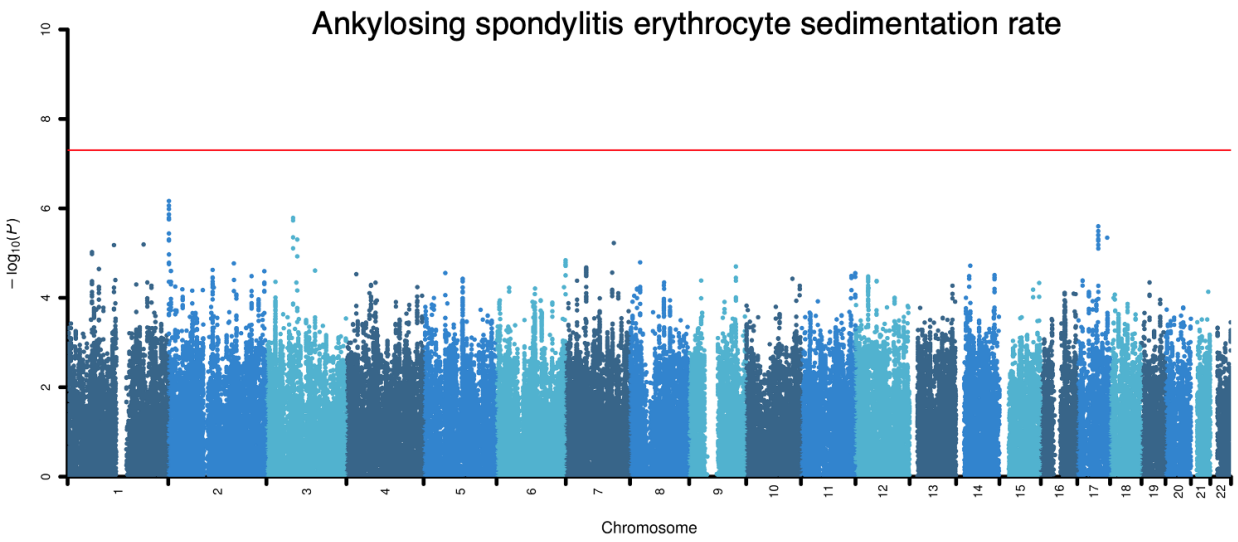

Figure S3. Manhattan plot for genome-wide genotype-by-treatment interaction results. The treatment is anti-IL17 therapy. The outcome is the ankylosing spondylitis erythrocyte sedimentation rate. The red line means P = 5e-8.

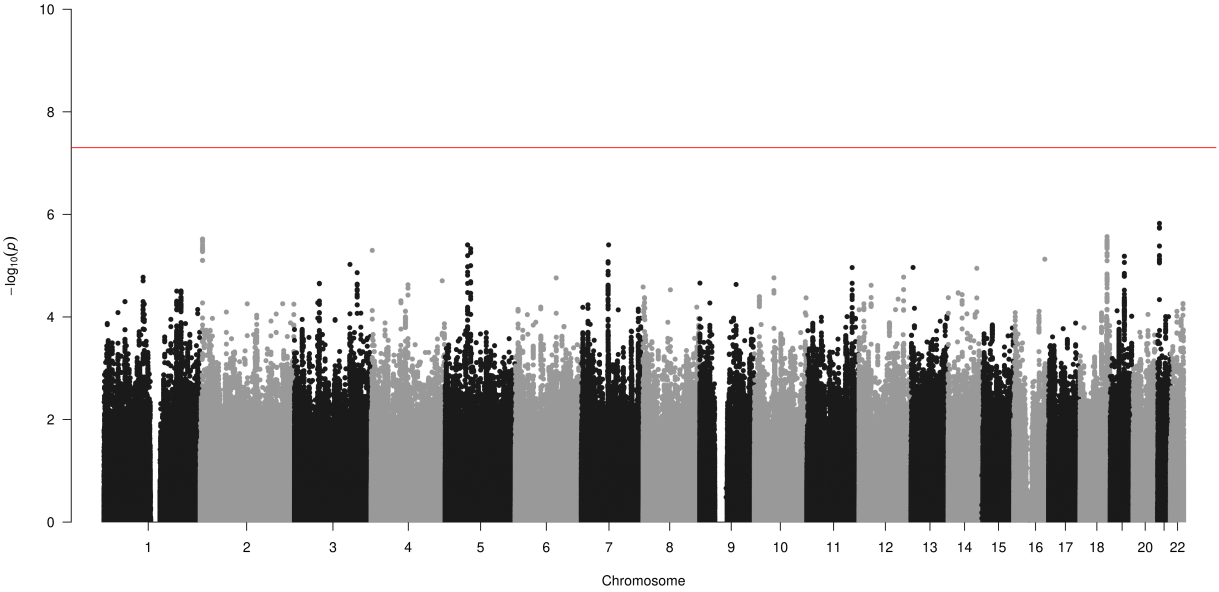

Figure S4. Manhattan plot for genome-wide genotype-by-treatment interaction results in Lung Health Study. The treatment is bronchodilator. The outcome is the changes of FEV1 (Volume that has been exhaled at the end of the first second of forced expiration) from baseline. The red line means P = 5e-8.