Beyond Reweighting: On the Predictive Role of Covariate Shift in Effect Generalization

Ying Jin*[1], Naoki Egami[2], and Dominik Rothenhäusler[3]

[1]Data Science Initiative & Department of Health Care Policy, Harvard University
[2]Department of Political Science, Columbia University
[3]Department of Statistics, Stanford University

December 13, 2024

Abstract

Many existing approaches to generalizing statistical inference amidst distribution shift operate under the covariate shift assumption, which posits that the conditional distribution of unobserved variables given observable ones is invariant across populations. However, recent empirical investigations have demonstrated that adjusting for shift in observed variables (covariate shift) is often insufficient for generalization. In other words, covariate shift does not typically "explain away" the distribution shift between settings. As such, addressing the unknown yet non-negligible shift in the unobserved variables given observed ones (conditional shift) is crucial for generalizable inference.

In this paper, we present a series of empirical evidence from two large-scale multi-site replication studies to support a new role of covariate shift in "predicting" the strength of the unknown conditional shift. Analyzing 680 studies across 65 sites, we find that even though the conditional shift is non-negligible, its strength can often be bounded by that of the observable covariate shift. However, this pattern only emerges when the two sources of shifts are quantified by our proposed standardized, "pivotal" measures. We then interpret this phenomenon by connecting it to similar patterns that can be theoretically derived from a random distribution shift model. Finally, we demonstrate that exploiting the predictive role of covariate shift leads to reliable and efficient uncertainty quantification for target estimates in generalization tasks with partially observed data. Overall, our empirical and theoretical analyses suggest a new way to approach the problem of distributional shift, generalizability, and external validity.

Keywords: Generalizability, external validity, distribution shift, replication studies.

1 Introduction

Distribution shift is a central issue in generalizing statistical evidence from an observed (source) population to a new, at most partially observed (target) population, with significant implications in many domains. For instance, in the medical and social sciences, researchers/policymakers seek to leverage existing randomized control trials (RCTs) to estimate the treatment effect on a new cohort to guide clinical decisions or policy making (Shadish et al., 2002; Hotz et al., 2005; Imai et al., 2008; Cole and Stuart, 2010; Tipton, 2013; Bareinboim and Pearl, 2016; Deaton and Cartwright, 2018). However, the challenge lies in whether statistical methods can capture the changes between populations to produce credible predictions of target effects.

To address the generalizability question, many statistical methods operate under assumptions positing that observed variables capture all distributional differences between populations. These assumptions can often be described as *covariate shift*, that is, the distribution of covariates observed in both populations can change, while the conditional distribution of the outcomes (unobserved in the target population) given the observed covariates remains invariant. For example, the distribution of age, gender, and education can

*Reproduction code for data processing and analysis is available at https://github.com/ying531/predictive-shift.

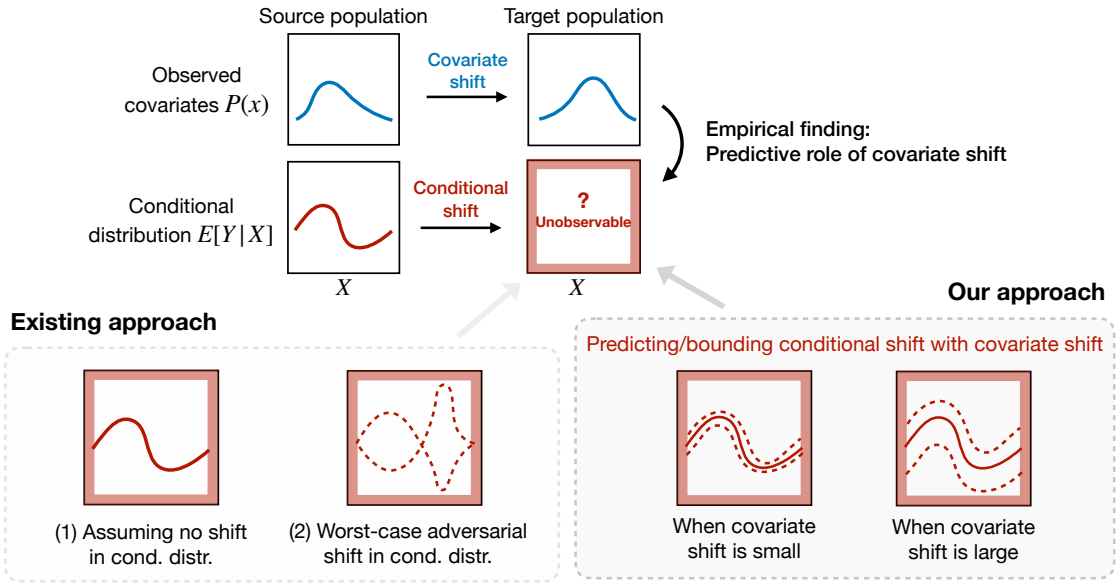

Figure 1: **Overview of the problem and our approach:** *Effect generalization from source and target populations needs to address the distribution shift consisting of the observed covariate shift and unobserved conditional shift. We argue a novel predictive role of covariate shift in bounding the strength of unknown conditional shift, which is supported by our empirical findings and leads to reliable and efficient generalization.*

differ across populations (e.g., due to convenience sampling), but the conditional treatment effect is the same for individuals with the same covariate profiles. Under this common assumption, adjusting for shift in the observed covariates, either by reweighting based on density ratios or estimating the heterogeneous covariate-outcome relationship (Stuart et al., 2011; Tipton et al., 2014; Miratrix et al., 2018; Dahabreh et al., 2019; Egami and Hartman, 2023), is sufficient for unbiased estimation of the target parameters. This common approach highlights the role of covariate shift in *explaining away* the distribution shift.

Given its popularity, a series of recent papers (Cai et al., 2023; Jin et al., 2023; Lu et al., 2023) have empirically evaluated the performance of generalization estimators based on the covariate shift assumption by comparing them against experimental benchmark estimates. Although each paper focuses on different domains, a common yet somewhat surprising finding is that observed covariate shift often can only explain a small proportion of the distributional shift in real-world applications. This implies two pessimistic messages: (1) adjusting for observed covariate shift may be insufficient for generalization, and (2) the remaining, unobserved conditional shift (i.e., shift in the conditional distribution of the outcomes given the observed covariates) is "larger" than the observed covariate shift. As such, it remains unclear how the conditional shift may be addressed for effect generalization in practice even in well-controlled settings.

1.1 This work: the predictive role of covariate shift

In this paper, we introduce a different role of covariate shift in *predicting* the unknown shift in the conditional distribution for generalization (Figure 1). The distribution shift between the source and target populations consists of the observed covariate shift and unobserved conditional shift, the latter being a key challenge in a generalization task. In contrast to existing approaches that either (i) assume there is no conditional shift, or (ii) establish worst-case bounds based on adversarial shift in the conditional distribution, we argue that the strength of covariate shift can *bound* that of the unknown conditional shift. Exploiting this bounding relationship is useful in effect generalization with improved validity and efficiency.

Our proposal is supported by empirical evidence from two well-known, large-scale multi-site replication projects—the Pipeline project (Schweinsberg et al., 2016) and the Many Labs 1 project (Klein et al., 2014)—

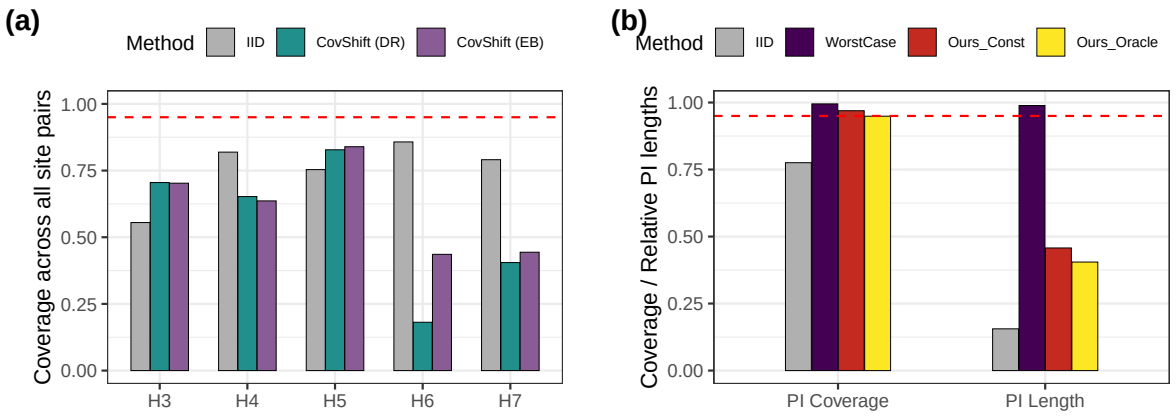

Figure 2: Preview of results. **[Left] Insufficient explanatory role of covariate shift:** *Empirical coverage of prediction intervals based on i.i.d. assumption (grey) and covariate shift assumption (green and purple), showing covariate shift cannot explain away distribution shift across sites.* **[Right] Reliable and efficient effect generalization based on the predictive role of covariate shift:** *Empirical coverage of prediction intervals based on i.i.d. assumption (grey), worst-case bounds (dark blue), and our method with the belief that conditional shift is bounded by covariate shift (red) or with knowledge of their relative strength (yellow).*

from the social sciences, analyzing a total of 680 studies across 65 sites examining 25 hypotheses.[1] To ensure faithful evaluation, since we have no access to the underlying population parameters, we build prediction intervals—based on various distribution shift assumptions—for *estimators* in target populations (including our proposed ones built upon empirical findings), and use their empirical coverage to examine the plausibility of the assumptions they are based upon. Figure 2 previews our empirical results.

We begin by examining common approaches that either ignore distribution shift or assume covariate shift (Section 2.3). In the two replication projects, the *explanatory* role of covariate shift is limited as evident from the low coverage of prediction intervals, complementing existing work that either examine pairs of studies (Jin et al., 2023) or mean squared errors (Lu et al., 2023; Kern et al., 2016). As shown in Panel (a) of Figure 2, even for controlled multi-site replication studies, distribution shifts across sites are not negligible (methods that assume no distributional shift (`IID`) do not achieve valid coverage). Furthermore, observed covariate shift cannot explain away the total distributional shift, as methods that only adjust for observed covariate shift (`CovShift`) do not achieve valid coverage, either.

We then proceed to compare the strengths of the observed covariate shift and the conditional shift (Section 3.2). In stark contrast with the pessimistic conjectures in previous works, we find that conditional shift is often *smaller* than covariate shift across different applications and comparisons. However, this empirical pattern became clear only after we measured covariate and conditional shifts with proper standardization.

We interpret our empirical findings by connecting them to similar patterns that can be theoretically derived under a recently proposed random distribution shift model (Jeong and Rothenhäusler, 2022, 2024; Bansak et al., 2024) (Section 3.3). Under this model, one expects to observe smaller conditional shift than covariate shift when the probability space is randomly perturbed in a way that does not favor any direction yet some component of the observed data, which is the treatment assignment here, is kept invariant. This model describes scenarios where the difference between the source and target distributions is not adversarial but is contributed by many small and random factors. Such scenarios are common in collaborative replication studies and potentially other carefully controlled studies where replicators try their best to mimic the original study design and population, but they have to deviate due to logistical and other constraints.

Finally, we demonstrate the effectiveness of exploiting this predictive role in effect generalization, again

[1] Note that not all sites examine all hypotheses.

(for evaluation purposes) by examining the empirical coverage of prediction intervals that aim to address the unknown conditional shift (Section 4). Panel (b) of Figure 2 previews key takeaway messages. Prediction intervals based on the novel predictive role maintain valid coverage while significantly shortening the intervals. This reveals that the predictive role is stable across contexts and permits effective empirical calibration. In contrast, existing methods assuming worst-case conditional shift (`WorstCase`) achieve valid coverage when the worst-case shift strength is (unrealistically) calibrated by data, but at the expense of too wide intervals.

Overall, our empirical and theoretical analyses suggest a new way to approach the problem of distributional shift, generalizability, and external validity. Most existing methods either (i) assume no shift in the unobserved conditional shift or (ii) assume shift in the unobserved conditional shift is bounded, and search for the worst-case scenarios that tend to be extremely adversarial. Instead, we offer a data-adaptive middle ground—shift in the unobserved conditional shift is non-negligible but is predictable from the observed covariate shift. Our results shall serve as the empirical and conceptual basis for developing new methods and models beyond the covariate shift assumption.

1.2 Scope of the paper

We note with caution that the main goal of this paper is to provide empirical and theoretical evidence for a new way of understanding real-world distribution shifts. The random distribution shift modeling assumption offers a perspective to justify our empirical findings, yet we do not anticipate it to be universally grounded. In particular, we limit the interpretation of our results to contexts similar to multi-site replication studies where data are collected in a "natural" manner, meaning that experimenters try to maintain consistency without adversarial patterns. In other words, the two projects provide a testbed for distribution shifts that emerge due to inevitable deviations despite well-controlled experimental settings (Stroebe and Strack, 2014; Hudson, 2023). Counter-examples include studies where the recruitment strategy changes. As an example, one study may be conducted on university students, whereas the second study may recruit only middle-aged participants. In this case, the random shift assumption may not be appropriate.

We also note that our evaluation mainly focuses on uncertainty quantification, that is, whether statistical methods can produce reliable prediction intervals for the actual estimates from data in the target population. Focusing on prediction intervals is inevitable since the underlying super-population parameter is not accessible for evaluation purposes. In addition, uncertainty quantification offers a more comprehensive assessment than evaluating the consistency or unbiasedness of point estimates (see Section 1.3 for more discussion).

1.3 Related work

Re-weighting in causal inference. Using re-weighting to generalize from one population to another population has a long history in causal inference. Early examples include Horvitz-Thompson (Horvitz and Thompson, 1952) and Hájek's estimator. Inverse probability weights are often unstable in practice. This has spurred the development of procedures that use outcome models to reduce variance (Robins et al., 1994) and balancing weight procedures that penalize the weights (Deville and Särndal, 1992; Hainmueller, 2012). Modern re-weighting procedures were used to generalize the results of experiments from one site to another (e.g., Cole and Stuart, 2010; Stuart et al., 2011; Tipton, 2013; Hartman et al., 2015; Buchanan et al., 2018; Dahabreh et al., 2019, 2020; Egami and Hartman, 2021). See Degtiar and Rose (2023) and Colnet et al. (2024) for recent reviews.

Empirical evaluation of generalization. This work adds to several recent works empirically evaluating generalization procedures that use unit-level data to generalize from one site to another. Cai et al. (2023) diagnose how much of the drop of prediction performance can be attributed to covariate shift vs concept $Y|X$ shift. Jin et al. (2023) and Lu et al. (2023) investigate how much of the discrepancy between causal effect estimates in different sites is due to unit-level covariates, among other factors. In welfare-to-work experiments, Lu et al. (2023) found that less than 10% of discrepancies between sites is explained by changes in covariate distributions. This work echoes these works on the insufficient explanatory role of covariate shift. An important distinction is that our evaluation leverages the coverage of prediction intervals over many

replication studies, which offers more comprehensive and faithful evaluation than methods that evaluate one pair of studies for a hypothesis (Jin et al., 2023; Cai et al., 2023) or examine the mean squared errors (Lu et al., 2023; Kern et al., 2016). For example, while Kern et al. (2016) find in another multi-site replication dataset that covariate adjustment leads to *unbiased* estimators (with bias averaged over multiple sites) for target estimates, it may still underestimate the variability if the conditional shift leads to discrepancies that are mean zero when averaged over studies but have non-negligible magnitude. More importantly, we also investigate a novel predictive role of covariate shift that can inform reliable generalization in practice.

Heterogeneity and meta-analysis in replicability. Multi-site replication projects have been used to examine the heterogeneity in effect estimates across sites (Klein et al., 2018; Coppock et al., 2018; McShane et al., 2022; Delios et al., 2022; Holzmeister et al., 2024). A prominent distinction is that these works often measure certain global notions of heterogeneity via meta-analysis (McShane et al., 2022), while we focus on generalization from one site to another. Methodologically, our generalization methods are applicable when data from only the source and target sites are available, whereas meta-analysis needs data from many sites. In addition, these works provide echoing messages for weak explanatory roles of observed factors (Klein et al., 2018; Delios et al., 2022) or complementary messages for design and estimation uncertainty (Krefeld-Schwalb et al., 2024; Holzmeister et al., 2024); the latter may be interpreted as "random" shifts if not documented.

Covariate and conditional shift in machine learning. The term covariate shift was first introduced by Shimodaira (2000), and has become one of the standard domain adaptation models, see Quinonero-Candela et al. (2008) and Pan and Yang (2009). Most commonly, covariate shift is addressed via importance weighting with the density ratio, which can be estimated directly, e.g., via a classifier (Bickel et al., 2007). Similarly, density ratio reweighting is a standard approach to addressing covariate shift for statistical estimation and inference. The conditional shift we study is related to the notion of concept drift in machine learning (Gama et al., 2014; Lu et al., 2018). The techniques for addressing these shifts in prediction problems serve distinct goals than our estimation and inference problems.

2 Motivating Applications and Methodological Problem

We introduce our motivating applications and illustrate the core methodological challenges in generalization.

2.1 Motivating Applications: Multi-Site Replication Projects

In this paper, we use two large-scale multi-site replication projects from the social sciences to empirically investigate the role of covariate shifts in generalization. The Many Labs 1 project (Klein et al., 2014) evaluates the replicability of 13 classic and contemporary experimental findings in the social sciences, ranging from gain versus loss framing (Tversky and Kahneman, 1981) to sex differences in implicit attitudes toward math (Nosek et al., 2002), across 36 independent data collection sites. Similarly, in the Pipeline project (Schweinsberg et al., 2016), 25 laboratories across the world (contributing 29 populations) independently replicate experiments for 10 scientific hypotheses concerning moral judgment, which is a well-known theory in psychology. Combining the two replication projects, we analyze 680 studies across 65 sites, examining 25 research hypotheses. This scale and diversity allow us to assess the proposed new role of covariate shifts across diverse empirical settings.

Several features of these multi-site replication projects make them suitable for evaluating distribution shifts in generalization. First, we can mimic the real-world generalization task by generalizing an effect estimate from one source site to another target site. Unlike the real generalization task, we have access to the effect estimate from the target site, and therefore, we can empirically evaluate the performance of common generalization estimators based on the covariate shift assumption and our proposed estimator, without simulating data from the artificial data-generating process. Second, in these replication projects, multiple laboratories follow the same experimental process as much as they can, known as direct replications. As a result, the measurement of the outcome variable and treatment variable is consistent across sites, and the interpretation of the covariate shift and the unobserved conditional shift becomes clearer. Finally, the

two replication projects differ in how laboratories are recruited. In the Pipeline project (Schweinsberg et al., 2016), laboratories are invited by the project lead because they had "access to a subject population in which the original finding was theoretically expected to replicate using the original materials" (p 57). Therefore, sites were selected such that distributional shifts between them are expected to be small or negligible. On the other hand, in the Many Labs 1 project (Klein et al., 2014), laboratories voluntarily participated in the project without specific eligibility criteria related to whether each site was expected to replicate the original finding. Here, sites were selected conveniently but "naturally" without explicit intention. This variation in site selection enables us to empirically evaluate distributional shifts in diverse scenarios.

The datasets are processed based on the raw data and scripts published by the original authors. In both projects, the covariates include demographic variables such as political ideology, gender, age, education and income. See Appendix A for details about the datasets and data pre-processing.

2.2 Notation and Setup

To formally discuss the generalization problem, we introduce some notation. While we tailor our notation to the two projects above for concrete presentation, the same general framework can be applied to any generalization setting across sites.

We first index the hypotheses by $k \in \{1, \ldots, K\}$ and the sites by $j \in \{1, \ldots, N\}$. Each hypothesis k is tested by a randomized experiment in a subset of sites $\mathcal{J}_k \subseteq \{1, \ldots, N\}$, following the same experimental protocol. Each site $j \in \mathcal{J}_k$ independently collects $n_j^{(k)} \in \mathbb{N}$ participants and collects data $\mathcal{D}_j^{(k)} = \{X_i^{(j,k)}, T_i^{(j,k)}, Y_i^{(j,k)}\}_{i=1}^{n_j^{(k)}}$, where X_i is the covariates, $T_i \in \{0, 1\}$ is the binary treatment, and Y_i is the outcome(s). Then, within each site, we can define the parameter of interest $\theta_j^{(k)}$ and its consistent and asymptotically normal estimator $\widehat{\theta}_j^{(k)}$, which is a function of $\mathcal{D}_j^{(k)}$. In our applications, most of them consider the average treatment effect (ATE) as $\theta_j^{(k)}$ and use a t-test that compares the sample mean of treated and control groups as $\widehat{\theta}_j^{(k)}$. Some hypotheses are tested with $\theta_j^{(k)}$ being the mean of outcomes and $\widehat{\theta}_j^{(k)}$ being a paired t-test comparing two outcomes. The specific hypotheses and tests are summarized in Tables 2 and 4.

We assume $\mathcal{D}_j^{(k)}$ are drawn i.i.d. from an underlying (hypothetical) super-population $\mathcal{S}_j^{(k)}$, and datasets are independent across sites $j \in \mathcal{J}_k$ for each hypothesis k. Importantly, the underlying data generating process $\mathcal{S}_j^{(k)}$ may vary across sites $j \in \mathcal{J}_k$ since there might exist distribution shifts.

We consider the generalization of estimates from site j_1 to j_2 for all pairs $j_1, j_2 \in \{1, \ldots, N\}$, $j_1 \neq j_2$, in each application. In general, we call the population in site j_1 as the source population P and the population in site j_2 as the target population Q. As typically the case in practice, for a *generalization task*, we assume all data from P are observed while only covariates X are observed from Q. When we *evaluate* the performance of various generalization estimators, we will use the full data in the target population Q to empirically evaluate how well the generalization estimators approximate the benchmark estimates in Q.

2.3 Challenge: Covariate Shift Cannot Explain Away Distributional Shift

The vast majority of existing methods for generalization assume that accounting for distributional shifts in observed covariates is sufficient, known as the covariate shift assumption. For example, when researchers want to generalize causal effects in one site to another site in the Pipeline project, they may assume that adjusting for observed characteristics of respondents, such as political ideology, gender, age, and education, is sufficient for generalization (consistent estimation and valid inference for the parameter in the target site).

However, in line with recent empirical evaluations (Cai et al., 2023; Jin et al., 2023; Lu et al., 2023), we find that this common assumption of covariate shift is often insufficient to explain away distributional shifts in the real-world applications. Figure 3 examines existing procedures that adjust for shift in observed covariates. We consider generalizing treatment effects from one site to another, using two commonly used estimators—the doubly robust (DR) estimator (Robins et al., 1994; Dahabreh et al., 2019) and the entropy balancing (EB) estimator (Särndal et al., 2003; Hainmueller, 2012)—to construct point estimates that are consistent

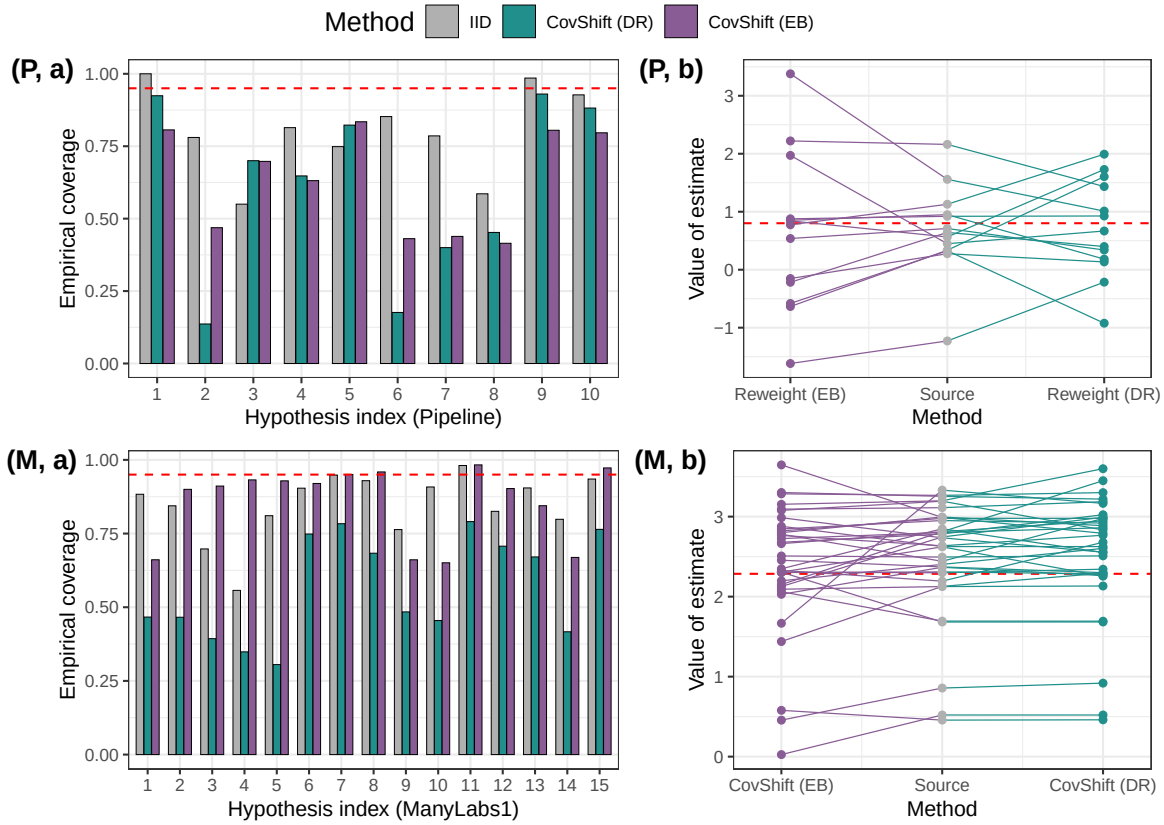

Figure 3: **Insufficient explanatory role of covariate shift**. [**Left**]: *Under-coverage of* 95% *prediction intervals based on the i.i.d. assumption (grey) and covariate shift assumption adjusted via doubly robust estimator (green) and entropy balancing (purple), averaged over all pairs of sites within each hypothesis for the Pipeline project* (P, a) *and the ManyLabs 1 data* (M, a)*, respectively. The red dashed line is the nominal level.* [**Right**]: *Estimates based on existing approaches (via doubly robust estimator (green) and entropy balancing (purple)) do not bring the source estimates (grey) closer to the target estimate (red dashed line). As illustrative examples, we show results when generalizing from all other sites to site 5 (raw ID) in hypothesis 5 in the Pipeline data* (P, b) *and when generalizing from all other sites to site 4 in hypothesis 4 in ManyLabs 1 data* (M, b)*. The segments connect estimates for the same pairs of sites.*

for the target *parameter* under the covariate shift assumption. Then, we follow Jin and Rothenhäusler (2024) to construct prediction intervals that would cover the target *estimator* with probability $1 - \alpha$ under covariate shift, and evaluate their empirical coverage.[2] As a simple baseline, we also compute prediction intervals based on the i.i.d. assumption that assumes no distribution shift between sites. Detailed estimation procedures are deferred to Appendix B.2. Figure 3 highlights two key findings:

(i) *Adjusting for distribution shift is necessary*, as prediction intervals based on the assumption of no distribution shift (denoted as `IID`) do not deliver valid coverage (grey bars in panel (a)).

(ii) *The explanatory role of covariate shift is insufficient.* This is evident from the under-coverage in panel (a) of both of the two `CovShift` methods. The coverage is sometimes even lower than `IID`; this is because the uncertainty that remains after adjustment is under-estimated. When comparing the estimates in the source population and generalization estimates in panel (b), we see that adjusting for covariate shift does not necessarily bring the estimators closer to the target estimate.

[2] We use prediction intervals rather than the conventional confidence intervals because we only have access to target population estimates (instead of the underlying parameters) for rigorous evaluation purposes.

3 The Predictive Role of Covariate Shift

In this paper, we highlight a new role of covariate shifts: observed covariate shifts can be used to *predict* unobserved shifts in the conditional distribution of Y given X, even though covariate shifts cannot fully explain the total distributional shift. We first propose standardized measures of distributional shifts, and then provide empirical and theoretical evidence for the predictive role of covariate shift.

3.1 Comparing the Strength of Covariate Shift and Conditional Shift

We begin by defining our measures of the two sources of distribution shifts: (i) the covariate shift in X (the part commonly addressed in existing methods) and (ii) the conditional shift—the shift in the conditional distribution of Y given X (the part assumed away under the covariate shift assumption). Our construction is based on two simple principles:

- *Scale invariance.* We would like our measures to reflect the strength of perturbations to the probability space, hence they should be invariant under linear scalings of the variables.
- *Numerical stability.* We would like our measures to be useful in guiding real generalization tasks, hence they should permit stable estimation.

Throughout the paper, we suppose the goal is to understand how causal effects change across sites, and we have two randomized experiments with treatment assignment probability π (most studies in our datasets are of this form). We can write the the difference of the causal effects across sites as

$$\theta_Q - \theta_P = \mathbb{E}_Q[\phi(T,Y)] - \mathbb{E}_P[\phi(T,Y)]$$
$$\text{where } \phi = \frac{T}{\pi}Y - \frac{1-T}{1-\pi}Y,$$

where $\mathbb{E}_P$ and $\mathbb{E}_Q$ are expectations over the source and target distribution. While we focus our discussion on causal effects in this paper for the sake of clear presentation, our proposed approach is applicable to any parameter of interest by redefining ϕ. For example, some studies in the Pipeline project use a one-sample t-test, in which case the parameter of interest is the mean of the outcome and $\phi = Y$.

We begin by conceptually decomposing the impact of overall distribution shift on the parameter of interest $(\theta_Q - \theta_P)$ to measure the shifts in X and ϕ given X separately:

$$\mathbb{E}_Q[\phi] - \mathbb{E}_P[\phi] = \underbrace{\left\{\mathbb{E}_Q[\phi_P(X)] - \mathbb{E}_P[\phi_P(X)]\right\}}_{=: \text{ Covariate shift}} + \underbrace{\left\{\mathbb{E}_Q[\phi - \phi_P(X)]\right\}}_{=: \text{ Conditional shift}}, \tag{1}$$

where $\phi_P(X) := \mathbb{E}_P[\phi|X]$ is the conditional expectation of the influence function in the source distribution. When the parameter of interest is the average treatment effect (ATE), we have $\phi_P(X) = \mathbb{E}_P[Y(1) - Y(0) \mid X]$, the conditional ATE function. In Jin et al. (2023), the decomposition (1) is used to diagnose the roles of different distribution shifts on the discrepancy of effect estimates between a pair of studies.

The first "Covariate shift" term in the decomposition (1) captures the shift in the observed covariates X. Intuitively, it measures how much the estimate can be brought closer to the target by adjusting for the shift in X. This term becomes larger when the strength of shift between $P(X)$ and $Q(X)$ is larger. Importantly, it also depend on the heterogeneity in $\phi_P(X)$, that is, how much the parameter of interest varies with the covariates. Our proposed distribution shift measures will remove the impact of such heterogeneity (sensitivity) on our measure of the strength of distribution shift to ensure interpretability and scale invariance.

The second term in (1) is equal to $\mathbb{E}_Q[\phi_Q(X) - \phi_P(X)]$, which captures the shift in the conditional expectation $\mathbb{E}[\phi|X]$ between the source and target distribution. For example, when the parameter of interest is the average treatment effect (ATE), this part captures how much the conditional ATE changes between the source and target distribution. Similarly, it not only depends on the strength of conditional shift but also the heterogenity in $\phi - \phi_P(X)$; again, the latter will be removed in our meausures.

The common assumption of covariate shift essentially assumes away the second shift in the conditional distribution and only accounts for the first term. We formalize the covariate shift assumption as follows.

Assumption 3.1 (Covariate Shift)**.** $P(\phi \mid X) = Q(\phi \mid X)$ *holds* P_X*-almost surely.*

If $\phi = Y$, this assumption is the classical covariate shift assumption in machine learning. For experiments, Assumption 3.1 is satisfied if the treatment probabilities do not change and the conditional distribution of the potential outcomes is invariant, i.e., if $P(Y(1), Y(0)|X) = Q(Y(1), Y(0)|X)$.

Assumption 3.1 implies the second term in (1) is zero, and thus it suffices to adjust for the shift in observed covariates (the first term). While this is a commonly imposed assumption for the identifiability of target parameters, as discussed in Section 2.3, it is often violated in practice, which implies that the conditional shift (the second term) is often nonzero in real-world applications. Therefore, instead of assuming away the conditional shift, we are to carefully investigate the relationship between these two shifts to offer new insights for moving beyond the covariate shift assumption in practice.

We define our distribution shift measures by rescaling the two terms in (1) by their standard deviation to ensure scale invariance:

$$\text{Relative conditional shift} = \frac{|\mathbb{E}_Q[\phi - \phi_P(X)]|}{\text{sd}_P(\phi - \phi_P(X))}, \tag{2}$$

$$\text{Relative covariate shift} = \frac{|\mathbb{E}_Q[\phi_P(X)] - \mathbb{E}_P[\phi_P(X)]|}{\text{sd}_P(\phi_P(X))}. \tag{3}$$

We will measure the strength of the conditional shift by the "relative conditional shift" (2). However, an issue with the "relative covariate shift" measure (3) is numerical instability whenever $\text{sd}_P(\phi_P(X))$ is close to zero. This might be problematic in social science applications where the explanatory power of covariates X can be low. To address this issue, we will use a Mahalanobis-type, "stabilized" measure instead:

$$\text{Stabilized covariate shift measure} = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L} \frac{(\mathbb{E}_Q[X_\ell] - \mathbb{E}_P[X_\ell])^2}{\text{Var}_P(X_\ell)}}, \tag{4}$$

where L is the number of covariates. We justify this covariate shift measure from a theoretical perspective in Section 3.3. Importantly, this measure is also invariant under the scaling of features.

Remark 3.2. We note that both (i) rescaling by standard deviation for scale invariance and (ii) adopting stabilized measure of covariate shift are crucial for interpretable and robust empirical insights. We illustrate the importance of these considerations through an ablation study in Appendix D which explores alternative distribution shift measures without these elements. These alternative distribution shift measures either fail to induce the predictive role or lead to much wider intervals in generalization due to numerical instablity.

In our evaluations, the two population measures will be replaced by their estimators. The estimation details are deferred to Appendix B with specific references in the corresponding parts of the paper.

3.2 Empirical Evidence: Covariate Shift Can Bound Conditional Shift

Using data from both the Pipeline project and the ManyLabs1 project, we establish empirical evidence that with our distribution shift measures, the covariate shift can *bound* the conditional shift, even though the strength of both may change across hypotheses and sites. Because the covariate shift is estimable in common generalization tasks, researchers can use this bounding relationship to *predict* the conditional shift, which is usually unobserved. We provide theoretical justification for the empirical findings in the next subsection.

We estimate the two distribution shift measures for any pair of sites for each hypothesis in the Pipeline project and the ManyLabs1 project. For any given hypothesis k, we define ϕ following the original analysis (c.f. Tables 2 and 4 for details), and $P = \mathcal{S}_{j_1}^{(k)}$, $Q = \mathcal{S}_{j_2}^{(k)}$ for all site pairs (j_1, j_2) and hypothesis index k.

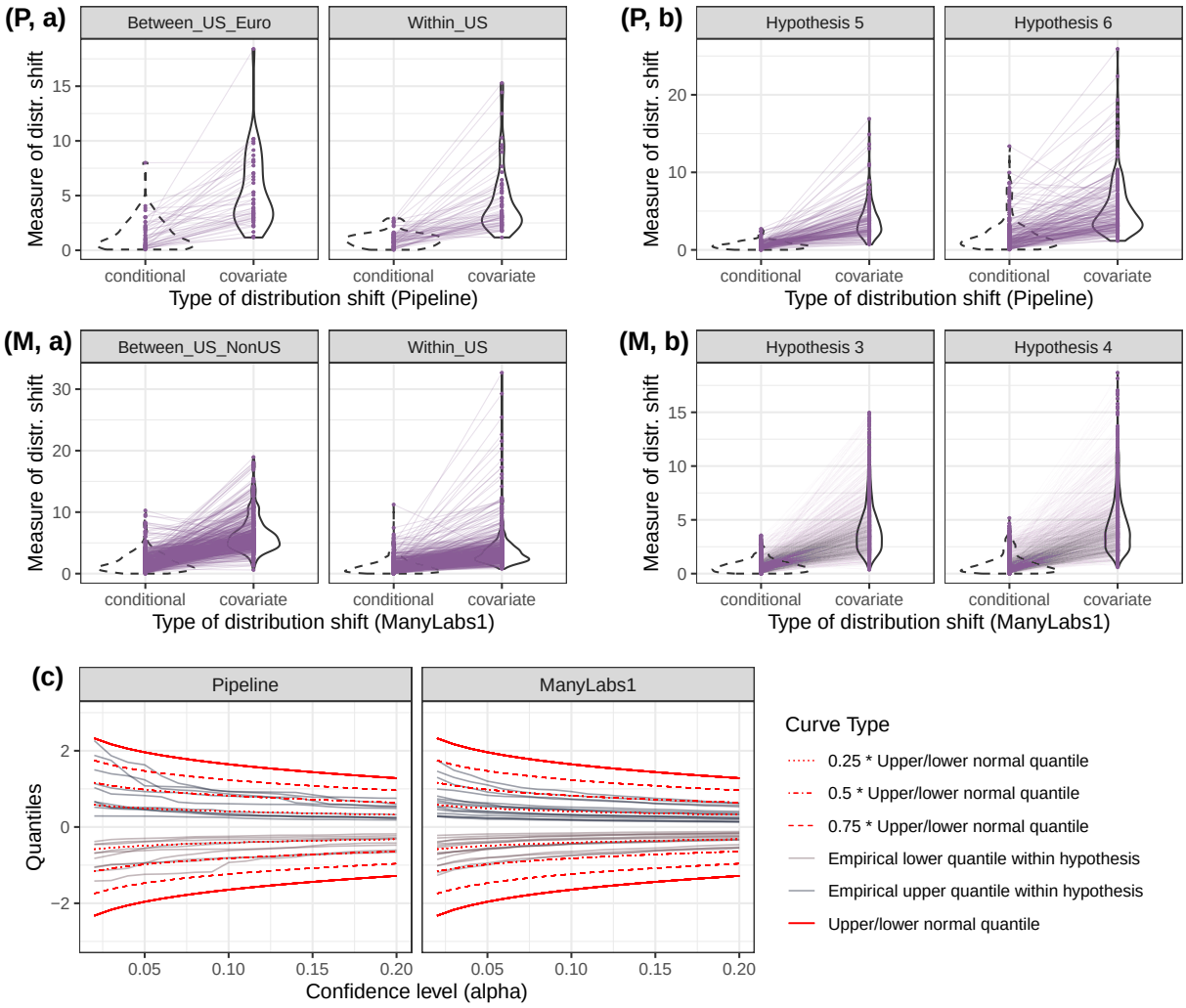

Figure 4: **Our covariate shift measures bound conditional shift measures in various contexts (pivotality)**. **[Left]**: *Conditional and covariate shift measures for site pairs between US and Europe/Non-US and site pairs within US in the Pipeline data* (P, a) *and the ManyLabs 1 data* (M, a). **[Right]**: *Conditional and covariate shift measures for all site pairs in hypotheses 5 and 6 in the Pipeline data* (P, b), *and those in hypotheses 3 and 4 in the ManyLabs 1 data* (M, b). *A few* (≤ 5) *largest values are removed for visualization.* **[Bottom]**: *Empirical quantiles of the ratios between conditional and covariate shift measures within each hypothesis in the Pipeline and ManyLabs1 datasets (grey and brown curves). The red curves are multiples of the quantiles of standard normal distribution plotted for reference.*

Then, we compute an estimate for the relative conditional shift (denoted by $\widehat{t}_{Y|X}^{j_1 \to j_2,(k)}$), and an estimate for the relative covariate shift (denoted by $\widehat{t}_{X}^{j_1 \to j_2,(k)}$). The estimation details are in Appendix B.3.

Figure 4 compares the conditional shift measure $\widehat{t}_{Y|X}^{j_1 \to j_2,(k)}$ and the covariate shift measure $\widehat{t}_{X}^{j_1 \to j_2,(k)}$ in various contexts. The left two panels (P, a) and (M, a) show site pairs (j_1, j_2) where one is in the United States (US) and the other is not in the US, as well as pairs where both sites are in the US. The right two panels (P, b) and (M, b) show site pairs within two hypotheses for each project.

In (P, a) and (M, a), we observe that the distribution shift between US-NonUS pairs tends to be larger than within-US pairs. In (P, b) and (M, b), the magnitude of distribution shifts also vary across hypotheses. Despite the variation across contexts, however, the covariate shift measure upper bounds the conditional shift

measure most of the time. In addition, when the conditional shift is larger (which is typically unobservable in a generalization task), the observable covariate shift also tends to be larger, justifying the "predictive role" of the covariate shift for the conditional shift.

Finally, panel (c) of Figure 4 provides a more quantitative illustration of the predictive role. In the figure, each curve is the $\alpha/2$ or $(1-\alpha/2)$-th empirical quantiles of the ratios $\{\widehat{t}_{Y|X}^{j_1\to j_2,(k)}/\widehat{t}_X^{j_1\to j_2,(k)}\}_{j_1\neq j_2}$ for a hypothesis k across a series of confidence levels α on the x-axis. For reference, we compare them with multiples of standard normal distribution quantiles. A few comments are in order:

- First, the absolute "bounding" relationship $|\widehat{t}_{Y|X}^{j_1\to j_2,(k)}/\widehat{t}_X^{j_1\to j_2,(k)}| \leq 1$ holds with high probability. Thus, in practice, the belief that $|\widehat{t}_{Y|X}| \leq \widehat{t}_X$ is a plausible option to establish a plausible range of the conditional shift strength. We will see reliable effect generalization based on this idea in Section 4.
- Second, if one wants to adjust the upper bound of $|\widehat{t}_{Y|X}^{j_1\to j_2,(k)}/\widehat{t}_X^{j_1\to j_2,(k)}|$ based on a desired confidence level, it is reasonable to use some multiplicative of standard normal, e.g., 0.75. Indeed, the empirical quantiles are smooth and similar to normal quantiles in general. This suggests a "smooth" and "random" nature of distribution shift, instead of being adversarial.

3.3 Theoretical Analysis: Random Distribution Shift Model

We here offer a theoretical framework to motivate the predictive role of covariate shift which justifies the empirical evidence in the last section.

We begin by modeling the data collection procedure as a two-stage sampling process. In the first stage, the underlying distribution is randomly "perturbed". With this perturbation, we aim to model unintended changes in the study population or random deviations from the experimental protocols despite efforts to keep them, etc. In the second stage, data is drawn i.i.d. from the perturbed distributions. Thus, we have three sources of uncertainty.

$$\underbrace{\widehat{\theta}_Q}_{\text{estimator on target dataset}} - \underbrace{\widehat{\theta}_P}_{\text{estimator on source dataset}} = \underbrace{\widehat{\theta}_Q - \theta(Q)}_{\text{sampling uncertainty}} + \underbrace{\theta(Q) - \theta(P)}_{\text{random shift}} + \underbrace{\theta(P) - \widehat{\theta}_P}_{\text{sampling uncertainty}}$$

Here, "sampling uncertainty" refers to the usual statistical uncertainty arising from randomly drawing observations from an underlying population P or Q, and "random shift" refers to the discrepancy between two underlying distributions P and Q due to natural "perturbations" to them. In the following, we will construct Q by randomly perturbing P. Constructing P by randomly perturbing Q or constructing both P and Q by randomly purturbing a third distribution P^0 would lead to the same asymptotics.

Our model for distribution shift includes three elements:

- We assume that the treatment distribution is invariant, since the treatment probability is fixed and chosen by the scientists for the datasets we consider here.
- There is distribution shift in observed covariates X, which we will model as random. Potentially there is shift in some unobserved effect modifiers U, which we will model as random, too.
- The outcome Y is a function of treatment indicator T, covariates X, and unobserved modifiers U. Thus, the shift in U is the driving factor for the conditional shift.

Let $Y = g(T, X, U)$, where the treatment T is independent of the modifiers (X, U) under P due to randomization. Recall that X is observed, while U is not.

Random distribution shift. The key idea of our random distribution shift model is that the original probability measure is randomly brought up and down in small pieces which, put together, leads to CLT-like behavior of the estimates with inflated variance.

To be precise, we let events $\{C_m^{(M)}\}_{m=1,\ldots,M}$ be a disjoint covering of the sample space of (X, U). We assume that these "pieces" have the same probability mass, i.e., $\mathbb{P}(C_m^{(M)}) = 1/M$ for $m = 1, \ldots, M$ and

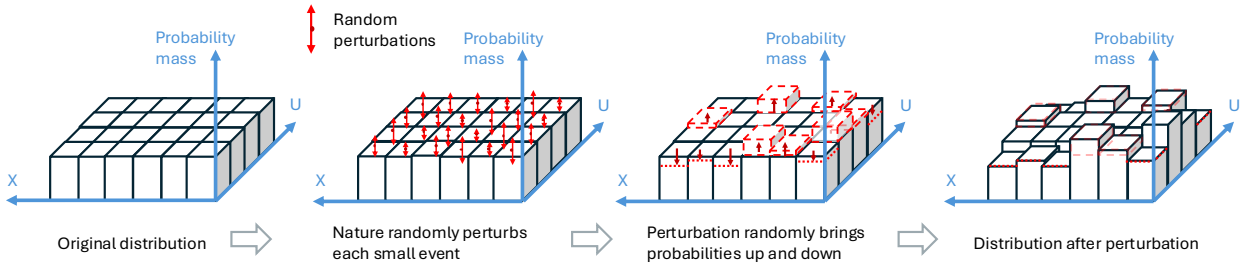

Figure 5: **Visualization of the random distribution shift model.** The original distribution is randomly perturbed to produce the distribution from which data are i.i.d. drawn. Our model assumes independent perturbation/reweighting of equal-probability small events and takes the number of small events to infinity.

that step functions on these pieces approximate square-integrable functions.[3] Later, we will take $M \to \infty$ to describe a scenario where many random factors change the probability masses of $C_m^{(M)}$ independently.

Our model describes random perturbations of P in these small event pieces. Specifically, we define the randomly re-weighted distribution Q for any event $E \subseteq \cup_{m=1}^M C_m^{(M)}$ via

$$Q(E) = \sum_{m=1}^{M} P(E \,|\, C_m^{(M)}) \cdot \frac{W_m}{\frac{1}{M}\sum_{m'=1}^{M} W_{m'}},$$

where $(W_m)_{m=1}^M$ are i.i.d. positive random variables that are bounded away from zero and have finite variance. As written above, the treatment indicator T is assumed to be independent of the modifiers (X, U) under both P and Q, and its distribution is invariant.

Figure 5 visualizes this idea, where probability masses of small events $\{C_m\}$ in the (X, U) sample space are independently perturbed by "nature". Such small, random perturbations are suitable to describe unintended but inevitable distribution shifts in such multi-site replication studies, such as unintended changes in the study population or random deviations from the experimental protocols despite efforts to keep them, etc.

Making the grid more fine-grained and taking limits $(n_Q, n_P, M \to \infty)$ we obtain a distributional CLT that describes the shift of empirical means under this two-stage sampling procedure. There are various asymptotic regimes that one could consider. Considering the asymptotic regime where $n_Q/M \to \rho \in (0, \infty)$ means sampling uncertainty and distributional uncertainty are of the same order (Jeong and Rothenhäusler, 2022). Taking $n_Q/M \to 0$ means distributional uncertainty is of larger order than sampling uncertainty (Jeong and Rothenhäusler, 2024). In the following, we focus on scenarios where sampling uncertainty and distributional uncertainty are of the same order, that is, we assume that n_Q/M and n_P/M converge to positive real numbers as we let $M \to \infty$.

Theorem 3.3 (Distributional CLT). *Let $\widehat{\mathbb{E}}_Q[\psi]$ denote the sample mean of a function $\psi(T, X, U)$ over n_Q i.i.d. draws from Q and $\widehat{\mathbb{E}}_P[\psi]$ denote the sample mean of ψ over n_P i.i.d. draws from P. Under the random distributional shift model described above, for any function $\psi(T, X, U) \in L^2(P)$, we have*

$$s_n^{-1}\left(\widehat{\mathbb{E}}_Q[\psi] - \widehat{\mathbb{E}}_P[\psi]\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where $s_n^2 = \left(\frac{1}{n_P} + \frac{1}{n_Q}\right) Var_P(\psi) + \delta_M^2 Var_P(\mathbb{E}_P[\psi|X, U])$, and $\delta_M^2 = \frac{1}{M}\frac{\mathbb{E}[W^2]}{\mathbb{E}[W]^2}$ measures the strength of perturbation. If ψ is a vector of functions, then $Var_P(\psi)$ and $Var_P(\mathbb{E}_P[\psi|X, U])$ are covariance matrices.

In Theorem 3.3, the variance term $(\frac{1}{n_P} + \frac{1}{n_Q})\mathrm{Var}_P(\psi)$ is the usual asymptotic variance one would obtain under the i.i.d. assumption that $P = Q$. In addition, random perturbations to the distributions contributes

[3] That is, for any function $\phi(X, U) \in L^2(P)$, it holds that $\mathbb{E}_P[(\phi(X, U) - \phi_M(X, U))^2] \to 0$ as $M \to \infty$, where $\phi_M = \sum_{m=1}^M 1_{C_m^{(M)}} \mathbb{E}_P[\phi(X, U)|C_m^{(M)}]$. This can be achieved relatively easily, e.g. for a continuous random variable $X \in \mathbb{R}$ one can choose C_m as intervals whose endpoints correspond to $(m-1)/M$-th quantile and the m/M-th quantile of X under P.

a factor of $\delta_{\mathrm{M}}^2 \mathrm{Var}_P(\mathbb{E}_P[\psi|X,U])$, where only the variance of $\mathbb{E}_P[\psi|X,U]$ counts because only the distribution of (X,U) is perturbed, while that of T remains invariant.

Why covariate shift often upper bounds conditional shift. In the following, we further discuss how this distributional CLT implies that covariate shift often upper bounds conditional shift.

For simplicity, we focus on deriving the generalization error for the estimators $\widehat{\theta}_P = \widehat{\mathbb{E}}_P[\phi]$ and $\widehat{\theta}_Q = \widehat{\mathbb{E}}_Q[\phi]$. A formal justification of this influence function approximation for general M-estimators can be found in Jeong and Rothenhäusler (2022). The numerator of our relative conditional shift measure (2) equals the difference-in-means estimator with $\psi = \phi - \phi_P(X)$ (ignoring the estimation of $\phi_P(X)$ for simplicity), where $\phi = \frac{T}{\pi}Y - \frac{1-T}{1-\pi}Y$ or $\phi = Y$ depending on the hypothesis. Applying the distributional CLT, for the squared relative conditional shift measure $(2)^2$, we get the estimate

$$\frac{(\widehat{\mathbb{E}}_Q[\psi] - \widehat{\mathbb{E}}_P[\psi])^2}{\widehat{\mathrm{Var}}_P(\psi)} \overset{d}{=} \left(\frac{1}{n_P} + \frac{1}{n_Q} + \delta_M^2 \frac{\mathrm{Var}_P(\mathbb{E}_P[\psi|X,U])}{\mathrm{Var}_P(\psi)}\right) Z_1 + o_P(\delta_M), \tag{5}$$

where $Z_1 \sim \chi^2(1)$. Using the distributional CLT for the covariates (taking $\psi = X_\ell$ where X_ℓ is the ℓ-th observed covariate), we obtain that standardized squared differences follows a scaled chi-square distribution:

$$\frac{(\widehat{\mathbb{E}}_Q[X_\ell] - \widehat{\mathbb{E}}_P[X_\ell])^2}{\widehat{\mathrm{Var}}_P(X_\ell)} \overset{d}{=} \left(\frac{1}{n_P} + \frac{1}{n_Q} + \delta_M^2\right) Z_1 + o_P(\delta_M). \tag{6}$$

Here, $\widehat{\mathrm{Var}}_P(X_\ell)$ is the sample variance of X_ℓ in the source data from P. Thus, up to lower order terms, equation (5) is stochastically smaller than equation (6) because $\mathrm{Var}_P(\mathbb{E}_P[\psi|X,U])/\mathrm{Var}_P(\psi) \leq 1$. In other words, the standardized conditional shift is stochastically smaller than the standardized covariate shift. This is in line with the empirical phenomenon observed in Figure 4. This also justifies replacing (3) by the stabilized version (4): this is roughly because the perturbations are homogeneous in different directions.

If we average over multiple covariates X_ℓ that are uncorrelated under P, by the distributional CLT, we can estimate the squared covariate shift measure $(4)^2$ by

$$\frac{1}{L}\sum_{\ell=1}^{L} \frac{(\widehat{\mathbb{E}}_Q[X_\ell] - \widehat{\mathbb{E}}_P[X_\ell])^2}{\widehat{\mathrm{Var}}_P(X_\ell)} \overset{d}{=} \left(\frac{1}{n_P} + \frac{1}{n_Q} + \delta_{\mathrm{M}}^2\right) \frac{Z_L}{L} + o_P(\delta_M), \tag{7}$$

where $Z_L \sim \chi^2(L)$. As $\frac{Z_L}{L} \to 1$ for $L \to \infty$, equation (7) will be close to $\frac{1}{n_P} + \frac{1}{n_Q} + \delta_{\mathrm{M}}^2$. When the covariates are correlated, one may standardize them with their empirical covariance matrix to restore (7). In our empirical studies, the covariates exhibit low correlation, hence we directly employ the formula (4).

These results motivate using a ratio of the estimated conditional shift and estimated covariate shift as a pivot to create prediction intervals. In the next section, we propose such prediction intervals and evaluate the empirical performance.

4 Effect Generalization by Exploiting the Predictive Role

In this section, we demonstrate that leveraging the predictive role of covariate shift leads to reliable generalization for target distributions. To this end, we build prediction intervals[4] for the target population estimate $\widehat{\theta}_Q$ based on our distribution shift measure and evaluate their empirical coverage.

4.1 Constructing Prediction Intervals

Before presenting the results, we begin with a high-level overview of our construction of the prediction intervals using the relationship between conditional and covariate shift measures, while we defer technical details on the estimation procedures to Appendix B.4.

[4]We again create prediction intervals for easier evaluation based on target estimates (instead of the underlying parameters).

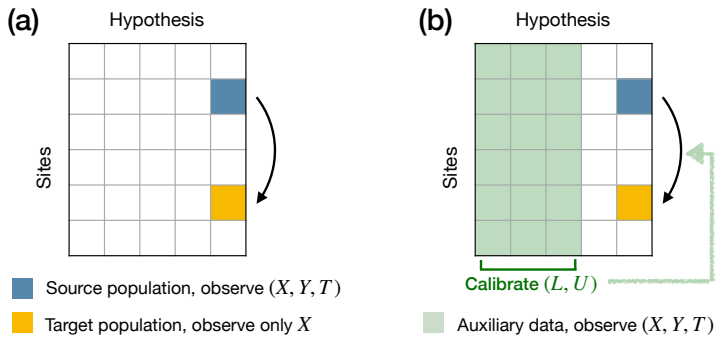

Figure 6: Generalization in two scenarios for the availability of data. **Left**: *Generalization without auxiliary data from source (blue) to target (yellow).* **Right**: *Generalization with auxiliary data (green) from the same sites for other hypotheses. The auxiliary data are used to calibrate L and U for a new generalization task.*

We consider generalization tasks where a scientist has access to full observations from the source distribution P but only the covariates X from the target distribution Q. To construct our prediction interval for the target *estimate* $\widehat{\theta}_Q$, we leverage the ratio between the covariate and conditional shift measures:

$$\widehat{r} := \widehat{t}_{Y|X} / \widehat{t}_X,$$

where $\widehat{t}_{Y|X}$ is the estimated conditional shift measure (2), and $\widehat{t}_X$ is the corresponding covariate shift measure (4). Note that one can estimate $\widehat{t}_X$ but not $\widehat{t}_{Y|X}$ in a generalization task. Suppose the distribution of $\widehat{r}$ can be characterized (e.g., using calibration approaches we will discuss below) so that one can find upper and lower bounds L and U obeying approximately

$$\mathbb{P}\Big(L \le \widehat{r} \le U\Big) \ge 1 - \alpha. \tag{8}$$

By definition, inverting the above event leads to a general form of our prediction interval for $\widehat{\theta}_Q$:

$$\widehat{C} = \Big[\widehat{\theta}_w + L \cdot \widehat{t}_X \cdot \widehat{s}_{Y|X},\ \widehat{\theta}_w + U \cdot \widehat{t}_X \cdot \widehat{s}_{Y|X}\Big], \tag{9}$$

where $\widehat{s}_{Y|X}$ is an estimate for $\mathrm{sd}_P(\phi - \phi_P)$ in (2), and $\widehat{\theta}_w$ is an estimator for $\mathbb{E}_Q[\phi_P(X)]$ in (2) which adjusts for the covariate shift. Above, all quantities in (9) except L and U can be estimated with full observations from the source distribution P and the covariate data from the target distribution Q.

We consider two ways to calibrate (L, U) under two scenarios of data availability (visualized in Figure 6):

1. **Constant calibration.** We construct prediction intervals assuming that the conditional shift measure is bounded by the covariate shift measure (i.e., using constant bounds $L = -1$ and $U = 1$). This is theoretically justified under the random distribution shift model (Section 3.3). This approach is applicable to a generalization task with no information other than covariate data from the target site.
2. **Data-adaptive calibration.** We construct prediction intervals by calibrating the relative strengths of conditional and covariate shift measures using some separate, existing data. This is applicable when some relevant auxiliary data are available (but not full observations in the target site) and we believe they inform the (relative) strengths of distribution shifts in the current generalization task.

Of course, the set of available data in the second approach can be more general; we explore other scenarios in Appendix C.1. These proposed prediction intervals are compared with three baselines:

1. `IID`. Prediction intervals under the i.i.d. assumption, i.e., $P = Q$, ignoring distribution shift.
2. `WorstCase`. Prediction intervals based on upper and lower worst-case bounds under restrictions on the distributional distance between the target distribution and the reweighted distribution, i.e., $\mathrm{KL}(Q_X \otimes P_{Y|X} \| Q) \le \rho$, where ρ is calibrated with data (see Appendix B.4 for details in both scenarios).

3. `Oracle`. Prediction intervals calibrated with true knowledge of the relative strength of covariate shift and conditional shift measures. This is the "ideal" but unrealistic version of our method.

We evaluate the generalization performance of different methods by the empirical coverage and average length of prediction intervals across all site pairs for each hypothesis.

4.2 Empirical Evaluation

4.2.1 Without Any Auxiliary Data

In the first scenario, the scientists have data from the source distribution but they do not have any information other than covariates X from the target distribution. In this setting, researchers can use our proposed approach with constant calibration.

More specifically, we consider the generalization of site j_1 to j_2 for all pairs $j_1, j_2 \in \{1, \ldots, N\}$, $j_1 \neq j_2$, for each hypothesis $k \in \{1, \ldots, K\}$ in each application. When we construct prediction intervals, we assume all data from j_1 are observed while only covariates X are observed from j_2. When we then *evaluate* the statistical performance of various generalization methods, we use the full data in site j_2 to empirically evaluate how well each estimator approximates the benchmark estimate in j_2.

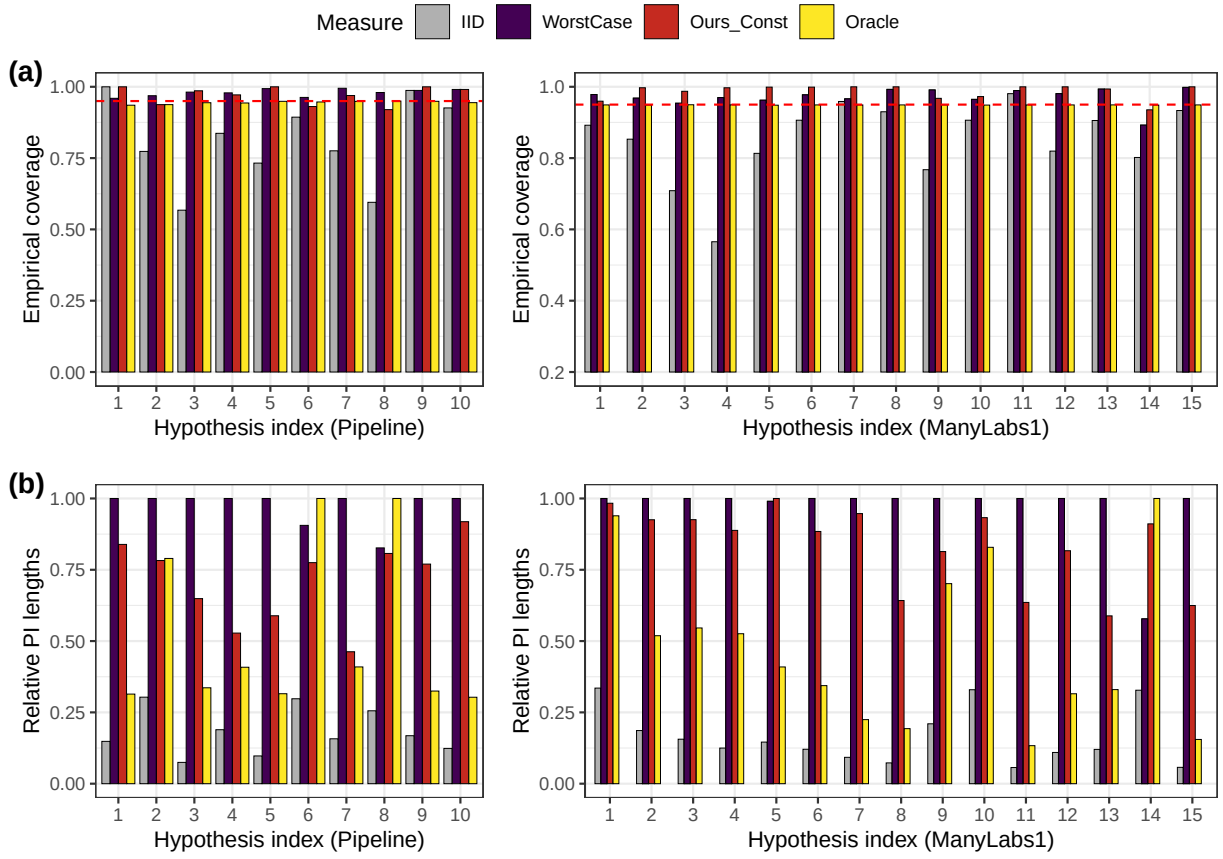

Figure 7: **Effect Generalization Without Auxiliary Data. Row (a)**: *Empirical coverage of prediction intervals via constant calibration at nominal level* $1 - \alpha = 0.95$ *and three baseline methods using the Pipeline data (left) and ManyLabs 1 data (right).* **Row (b)**: *Average length of prediction intervals for constant calibrated prediction intervals at nominal level* $1 - \alpha = 0.95$ *for the four methods, normalized by the largest average length in each study, using the Pipeline data (left) and ManyLabs 1 data (right).*

In Figure 7, we report the empirical coverage and relative lengths of prediction intervals averaged over

all pairs within each hypothesis. Across two distinct applications, our procedure (denoted as "Ours_Const" in red) achieves the target of 95% coverage in most cases (see panel (a)). `WorstCase` prediction intervals achieve the target coverage as well but are much wider than the proposed intervals (see panel (b)). Not surprisingly, intervals based on the i.i.d. assumption exhibit undercoverage.

4.2.2 With Auxiliary Data

Next, we examine how we can improve the performance of our estimator when researchers have some auxiliary data to use data-adaptive calibration for our method. We specifically consider a scenario where data from all sites exist for other hypotheses to build prediction intervals for a new hypothesis. In practice, this setting arises when there is existing data from the same set of sites on other research questions or hypotheses.

We calibrate L and U by the quantiles of site pairs from existing hypotheses, and then build a prediction interval for a new hypothesis that is only observed in one single site. To ensure stable evaluation, the ordering of the sites is randomly permuted for 10 times. Additional calibration scenarios (generalizing to new sites for existing hypotheses, and new sites for new hypotheses) in Appendix C.1 deliver similar messages.

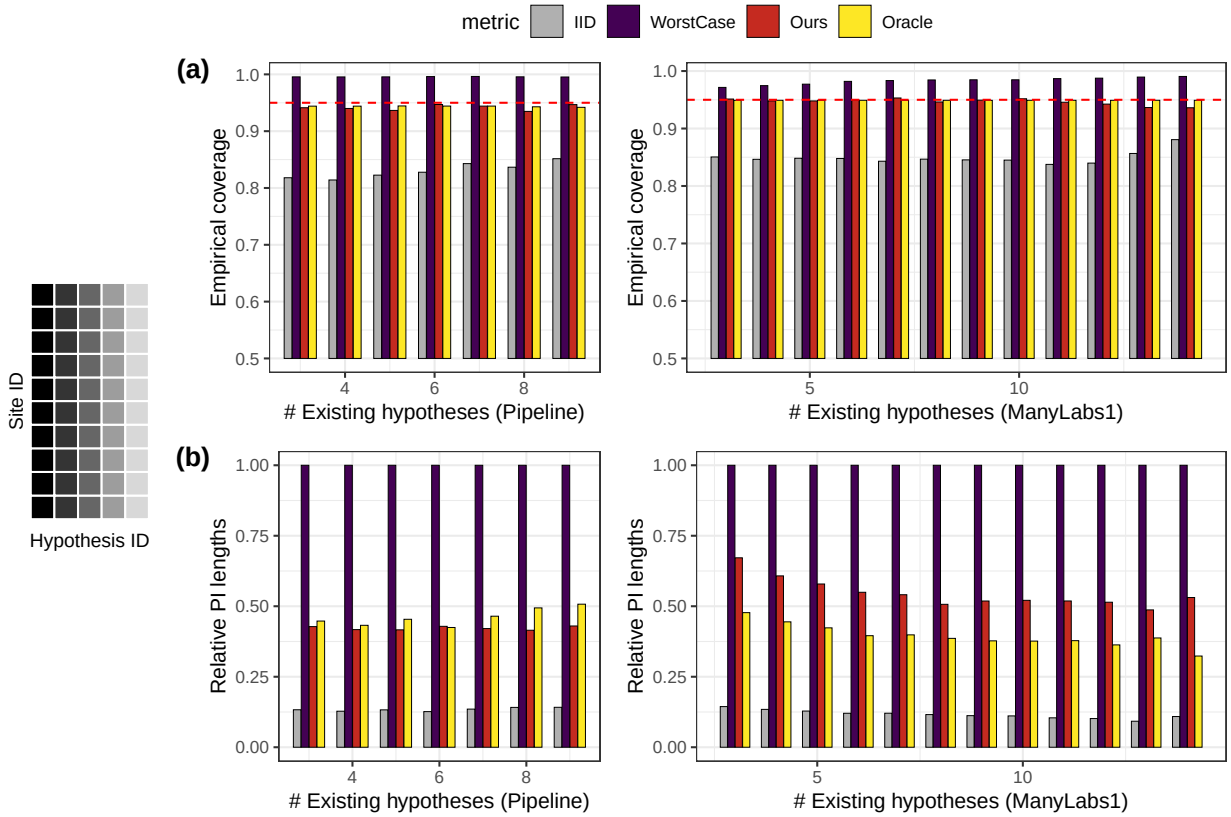

Figure 8: **Effect Generalization With Auxiliary Data:** We generalize to new studies based on distribution shift measures calibrated from the same sites in other hypotheses. **Left**: *Illustration of data collection order, where dark color means earlier.* **Row (a)**: *Average coverage of prediction intervals built with four methods over 10 random draws of study ordering, using the Pipeline data (left) and ManyLabs 1 data (right). The red dashed line is the nominal level* 0.95. **Row (b)**: *Average length of prediction intervals over 10 random draws of study ordering, normalized by the largest average length in each study, using the Pipeline data (left) and ManyLabs 1 data (right).*

In Figure 8, we report the coverage and lengths of prediction intervals. For both projects, our procedure achieves coverage close to the nominal level, with prediction intervals that are much smaller than the intervals

based on worst-case bounds, and quite close to the oracle method. As before, prediction intervals based on the i.i.d. assumption exhibit undercoverage.

5 Discussion

In this work, we offer new insights on distribution shifts when inferring parameter estimates in a new site based on data from one site and covariate data from the other one. By empirical benchmarking in large-scale replication projects, we find significant distributional shifts between sites. Moreover, approaches that only account for distribution shifts of observed covariates—thereby relying on the *explanatory* role of covariate shift—are often insufficient for explaining discrepancies between sites.

Instead of using covariates in an explanatory fashion, we propose to use covariates in a *predictive* fashion. More precisely, we suggest predicting the strength of the shift of unobserved conditional distribution based on the strength of the shift of observed covariates. We provide empirical evidence based on two large-scale replication studies and offer a theoretical justification under a random distribution shift model.

In our empirical applications, we show that our proposed prediction intervals maintain the desired coverage even in the presence of (unobservable) distributional shifts. While these intervals can sometimes be over-conservative, they offer a significant improvement over existing approaches. Our method compares favorably to worst-case approaches, which tend to be overly pessimistic and lead to excessively wide intervals.

Our empirical and theoretical findings open up several exciting future avenues for research. First, real-world scenarios may involve more complex forms of distributional change than the one studied in this work. For instance, in settings with emulated target populations, it might be reasonable to consider hybrid models where there is a combination of controlled (and potentially large) covariate shift and random shifts that arise due to inevitable deviations. Developing optimal estimation procedures for such hybrid models would be a valuable contribution. Second, the non-negligible conditional shift suggests the importance of collecting data from diverse sources to properly address the "distributional uncertainty" component in estimation. Towards this goal, the investigation of distribution shift patterns can provide insights for an important methodological challenge: determining how to prioritize data collection. For example, if there is a partial covariate shift and partial random shift, it may be beneficial to prioritize the collection of covariates most affected by the shift, rather than gathering more data across all variables equally.

Acknowledgments

We appreciate excellent research assistance by Diana Da In Lee. Egami acknowledges financial support from the National Science Foundation (SES–2318659). Rothenhäusler acknowledges financial support from the Dieter Schwarz Foundation, the Dudley Chamber fund, and the David Huntington Foundation.

References

Bansak, K. C., Paulson, E., and Rothenhäusler, D. (2024). Learning under random distributional shifts. In *International Conference on Artificial Intelligence and Statistics*, pages 3943–3951. PMLR.

Bareinboim, E. and Pearl, J. (2016). Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). Generalizing Evidence From Randomized Trials Using Inverse Probability Of Sampling Weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1193–1209.

Cai, T. T., Namkoong, H., and Yadlowsky, S. (2023). Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. *Statistical science*, 39(1):165–191.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernan, M. A. (2020). Extending Inferences from A Randomized Trial to A New Target Population. *Statistics in medicine*, 39(14):1999–2014.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694.

Deaton, A. and Cartwright, N. (2018). Understanding and Misunderstanding Randomized Controlled Trials. *Social Science & Medicine*.

Degtiar, I. and Rose, S. (2023). A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524.

Delios, A., Clemente, E. G., Wu, T., Tan, H., Wang, Y., Gordon, M., Viganola, D., Chen, Z., Dreber, A., Johannesson, M., et al. (2022). Examining the generalizability of research findings from archival data. *Proceedings of the National Academy of Sciences*, 119(30):e2120377119.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Egami, N. and Hartman, E. (2021). Covariate Selection for Generalizing Experimental Results: Application to A Large-scale Development Program in Uganda. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4):1524–1548.

Egami, N. and Hartman, E. (2023). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, 117(3):1070–1088.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 178(3):757–778.

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., and Kirchler, M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences*, 121(32):e2403490121.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Hotz, V. J., Imbens, G. W., and Mortimer, J. H. (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics*, 125(1-2):241–270.

Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9.

Hudson, R. (2023). Explicating exact versus conceptual replication. *Erkenntnis*, 88(6):2493–2514.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings Between Experimentalists and Observationalists About Causal Inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502.

Jeong, Y. and Rothenhäusler, D. (2022). Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty. *arXiv preprint arXiv:2202.11886*.

Jeong, Y. and Rothenhäusler, D. (2024). Out-of-distribution generalization under random, dense distributional shifts. *arXiv preprint arXiv:2404.18370*.

Jin, Y., Guo, K., and Rothenhäusler, D. (2023). Diagnosing the role of observable distribution shift in scientific replications. *arXiv preprint arXiv:2309.01056*.

Jin, Y. and Rothenhäusler, D. (2024). Tailored inference for finite populations: conditional validity and transfer across distributions. *Biometrika*, 111(1):215–233.

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, 9(1):103–127.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating Variation in Replicability. *Social psychology*.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.

Krefeld-Schwalb, A., Sugerman, E. R., and Johnson, E. J. (2024). Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. *Proceedings of the National Academy of Sciences*, 121(12):e2306281121.

Lu, B., Ben-Michael, E., Feller, A., and Miratrix, L. (2023). Is It Who You Are or Where You Are? Accounting for Compositional Differences in Cross-Site Treatment Effect Variation. *Journal of Educational and Behavioral Statistics*, 48(4):420–453.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363.

Madan, N., Uhlmann, E. L., Schweinsberg, M., and Tierney, W. (2016). The pipeline project.

McShane, B. B., Böckenholt, U., and Hansen, K. T. (2022). Modeling and learning from variation and covariation. *Journal of the American Statistical Association*, 117(540):1627–1630.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth Weighting? How to Think About and Use Weights in Survey Experiments. *Political Analysis*, 26(3):275–291.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Math = Male, Me= Female, Therefore Math $\neq$ Me. *Journal of personality and social psychology*, 83(1):44.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2008). *Dataset Shift in Machine Learning*. MIT Press.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66:55–67.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Stroebe, W. and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1):59–71.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386.

Tipton, E. (2013). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., and Caverly, S. (2014). Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *Journal of Research on Educational Effectiveness*, 7(1):114–135.

Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *science*, 211(4481):453–458.

A Details of datasets and data pre-processing

A.1 Pre-processing for Pipeline project

The raw datasets for the Pipeline project can be found in the OSF repository https://osf.io/q25xa/. The detailed data pre-processing script can be found in the folder `Pipeline` in the GitHub respository https://github.com/ying531/awesome-replicability-data.

We follow the data processing scripts (in the folder "SPSS Syntax files") provided in the OSF repository to compute the response variables, encode the treatment indicators, and extract the covariates including age, gender, country of birth, language, ethnicity, parent education, and family incomes. When running the analysis, we additionally process the data for each site as follows: covariates with all N/A values are excluded; otherwise, the missing observations are imputed by the site median. Since entropy balancing enforces positive weights, when running the EB-based methods, we also exclude covariates whose sample average in the target dataset falls outside the support in the source dataset.

A.2 Pre-processing for ManyLabs1 project

The raw datasets for the ManyLabs1 project an be found in the OSF repository https://osf.io/wx7ck/. The detailed data processing script can be found in the folder `ManyLabs1` in the GitHub respository https://github.com/ying531/awesome-replicability-data.

We follow the data processing scripts `Syntax.Manylabs.sps` in the OSF repository to encode the responses (`dv`) and treatment indicators (`iv`), and extract the covariates including gender, age, race, ethnicity, nationality, native language, religion, and ideology.

A.3 Reproduction code

The code for reproducing the analysis is available at https://github.com/ying531/predictive-shift. For easier reproduction, we also include analyses results (such as computed distribution shift measures and constructed KL-based bounds which can be costly to run) ready for producing the figures in the main text.

A.4 Dataset information

Table 1 lists the data indices and data collection sites for the Pipeline project from the Open Science Framework (OSF) repository. Table 2 summarizes the information for each of the 10 hypothesis studied in the Pipeline project, including the name, test statistic and formula, number of sites conducting experiments for testing this hypothesis, and total sample sizes N recruited in these sites.

Table 3 lists the data collection sites in the Manylabs1 project. Table 4 summarizes the information for each of the 15 hypotheses studied in the ManyLabs1 dataset, including the hypothesis, estimator, formula (for processed data), number of sites conducting experiments for the hypothesis, and total sample sizes N.

New Index	Raw ID	PI, Institution
1	0	Original Study data collection
2	1	Aaron Sackett, University of St. Thomas
3	2	Alexandra Mislin, American University
4	4	David Tannenbaum, University of Chicago
5	5	Daniel Storage, University of Illinois at Urbana-Champaign
6	6	Adam Hahn, University of Cologne
7	7	Nicole Legate, Illinois Institute of Technology
8	8	INSEAD Sorbonne Lab
9	9	Victoria Brescoll, Yale University
10	10	Felix Cheung, Michigan State University/University of Hong Kong
11	11	Fiery Cushman, Harvard University
12	12	Jay Van Bavel, New York University
13	13	Tatiana Sokolova, HEC Paris and University of Michigan
14	15	Jesse Graham, University of Southern California
15	16	Anne-Laure Sellier, HEC Paris
16	17	Eli Awtrey, University of Washington
17	18	Jennifer Jordan, University of Groningen
18	19	Sapna Cheryan, University of Washington
19	20	Xiaomin Sun, Beijing Normal University
20	21	Yoel Inbar, University of Toronto
21	22	Wendy Bedwell, University of South Florida
22	24	Deanna Kennedy, University of Washington Bothell
23	25	Matt Motyl, University of Illinois at Chicago
24	26	Erik Cheries, University of Massachusetts Amherst
25	27	Additional INSEAD-Sorbonne lab data for Study 1
26	141	Dan Molden, Packet 1 for Study 7
27	142	Dan Molden, Packet 2 for Study 4 and Study 8
28	311	UCI Psychology Students
29	312	UCI Business Students

Table 1: List of new index, raw site ID in the dataset, and contributing PI and site institutions in the Pipeline project dataset, taken from the Open Science Framework project repository (Madan et al., 2016).

ID	Hypothesis	Estimator	Formula	#Sites	N
1	Bigot–misanthrope	t-test	`bigot_personjudge ~ condition`	12	2861
2	Cold-hearted prosociality	Paired t-test	`tdiff ~ 1`	12	2806
3	Bad tipper	t-test	`tipper_personjudg ~ condition`	16	3658
4	Belief–act inconsistency	t-test	`beliefact_mrlblmw_rec ~ condition13`	13	3006
5	Moral inversion	t-test	`moralgood ~ condition`	14	3076
6	Moral cliff	Paired t-test	`diff ~ 1`	15	3300
7	Intuitive economics	t-test	`yz ~ condition`	15	3164
8	Burn-in-hell	Paired t-test	`tdiff ~ 1`	15	3176
9	Presumption of guilt	t-test	`companyevaluation ~ condition`	17	3806
10	Higher standard	t-test	`standard_evalu_7items ~ condition`	11	2692

Table 2: Estimator, number of sites and total sample size N for each study in the Pipeline project.

New Index	Raw Site ID	Institution, Location
1	Abington	Penn State Abington, Abington, PA
2	Brasilia	University of Brasilia, Brasilia, Brazil
3	Charles	Charles University, Prague, Czech Republic
4	Conncoll	Connecticut College, New London, CT
5	CSUN	California State University, Northridge, LA, CA
6	Help	HELP University, Malaysia
7	Ithaca	Ithaca College, Ithaca, NY
8	JMU	James Madison University, Harrisonburg, VA
9	KU	Koç University, Istanbul, Turkey
10	Laurier	Wilfrid Laurier University, Waterloo, Ontario, Canada
11	LSE	London School of Economics and Political Science, London, UK
12	Luc	Loyola University Chicago, Chicago, IL
13	McDaniel	McDaniel College, Westminster, MD
14	MSVU	Mount Saint Vincent University, Halifax, Nova Scotia, Canada
15	MTURK	Amazon Mechanical Turk (US workers only)
16	OSU	Ohio State University, Columbus, OH
17	Oxy	Occidental College, LA, CA
18	PI	Project Implicit Volunteers (US citizens/residents only)
19	PSU	Penn State University, University Park, PA
20	QCCUNY	Queens College, City University of New York, NY
21	QCCUNY2	Queens College, City University of New York, NY
22	SDSU	SDSU, San Diego, CA
23	SWPS	University of Social Sciences and Humanities Campus Sopot, Sopot, Poland
24	SWPSON	Volunteers visiting www.badania.net
25	TAMU	Texas A&M University, College Station, TX
26	TAMUC	Texas A&M University-Commerce, Commerce, TX
27	TAMUON	Texas A&M University, College Station, TX (Online participants)
28	Tilburg	Tilburg University, Tilburg, Netherlands
29	UFL	University of Florida, Gainesville, FL
30	UNIPD	University of Padua, Padua, Italy
31	UVA	University of Virginia, Charlottesville, VA
32	VCU	VCU, Richmond, VA
33	Wisc	University of Wisconsin-Madison, Madison, WI
34	WKU	Western Kentucky University, Bowling Green, KY
35	WL	Washington & Lee University, Lexington, VA
36	WPI	Worcester Polytechnic Institute, Worcester, MA

Table 3: List of new index, raw site ID in the dataset, and contributing PI and site institutions in the ManyLabs 1 dataset, taken from the original paper (Klein et al., 2014).

B Estimation details

In this section, we detail the estimation procedures for all the analyses in this paper. Appendix B.1 recalls important notations. Appendix B.2 describes the analysis for the explanatory role in Section 2.3. Appendix B.3 details the estimation for our distribution shift measures in Section 3. Finally, Appendix B.4 details our estimation and evaluation procedures for effect generalization in Section 4.

ID	Hypothesis	Estimator	Formula	#Sites	N
1	Allowedforbidden	t-test	`dv ~ iv`	36	6292
2	Anchoring1	t-test	`dv ~ iv`	36	5362
3	Anchoring2	t-test	`dv ~ iv`	36	5284
4	Anchoring3	t-test	`dv ~ iv`	36	5627
5	Anchoring4	t-test	`dv ~ iv`	36	5609
6	Contact	t-test	`dv ~ iv`	36	6336
7	Flag	t-test	`dv ~ iv`	36	6251
8	Gainloss	t-test	`dv ~ iv`	36	6271
9	Gambfal	t-test	`dv ~ iv`	36	5942
10	Iat	t-test	`dv ~ iv`	36	5851
11	Money	t-test	`dv ~ iv`	36	6333
12	Quote	t-test	`dv ~ iv`	36	6325
13	Reciprocity	t-test	`dv ~ iv`	36	6276
14	Scales	t-test	`dv ~ iv`	36	5899
15	Sunk	t-test	`dv ~ iv`	36	6330

Table 4: Estimator, number of sites and total sample size N for each study (indices and variable names in cleaned data and this paper) for ManyLabs 1 data.

B.1 Notations

We begin by revisiting some notations. A hypothesis k is replicated by sites $j \in \{1, \ldots, N_k\}$, each observing a dataset $\mathcal{D}_j^{(k)} = \{X_i^{(j,k)}, T_i^{(j,k)}, Y_i^{(j,k)}\}_{i=1}^{n_j^{(k)}}$, where X_i is the covariates, $T_i \in \{0, 1\}$ is the binary treatment, and Y_i is the outcome(s). For each hypothesis k, the estimate for site j is $\widehat{\theta}_j^{(k)} = \theta^{(k)}(\mathcal{D}_j^{(k)})$, where $\theta^{(k)}$ is the same functional that represents the analysis procedure applied to all sites (as listed in Tables 2 and 4). Here, $\widehat{\theta}_j^{(k)}$ estimates the population parameter $\theta_j^{(k)} = \theta^{(k)}(P_j^{(k)})$, where $P_j^{(k)}$ is the underlying distribution from which $\mathcal{D}_j^{(k)}$ is drawn. We assume access to a function $\phi^{(k)}(\cdot)$ such that

$$\widehat{\theta}_j^{(k)} = \frac{1}{n_j^{(k)}} \sum_{i=1}^{n_j^{(k)}} \phi^{(k)}\big(X_i^{(j,k)}, Y_i^{(j,k)}, T_i^{(j,k)}\big).$$

B.2 Estimation for the explanatory role

In this part, we detail how the prediction intervals for IID, CovShift (DR) and CovShift (EB) are constructed and evaluated in Section 2.3. The sites are denoted as $i, j \in \{1, \ldots, N\}$, where site i is the "original" site with full observations, and site j is the "target" site we want to generalize the effects to.

Estimation for IID. For any site pair (i, j) for a hypothesis k, we assume access to a consistent variance estimator $(\widehat{\sigma}_i^{(k)})^2$ for $\widehat{\theta}_i^{(k)}$, such that

$$\sqrt{n_i^{(k)}} \cdot \frac{\widehat{\theta}_i^{(k)} - \theta_i^{(k)}}{\widehat{\sigma}_i^{(k)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Note that $\widehat{\theta}_i$ and $\widehat{\sigma}_i$ can be computed using full observations $\mathcal{D}_i^{(k)}$ from the "original" site i. It is straightforward to construct these estimators for the t-tests and paired t-tests considered in this work, and we note that $\widehat{\sigma}_i^{(k)} = \widehat{\sigma}_j^{(k)} + o_P(1)$ for any $i \neq j$ if the i.i.d. assumption holds. For the `IID` method, we construct a prediction interval based on site i for $\widehat{\theta}_j^{(k)}$ via

$$\widehat{C}_{i \to j}^{\mathrm{IID},(k)} = \widehat{\theta}_i^{(k)} \pm q_{1-\alpha/2} \cdot \widehat{\sigma}_i^{(k)} \cdot \left(\sqrt{1/n_i^{(k)} + 1/n_j^{(k)}}\right), \tag{10}$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of a standard normal distribution. Under the i.i.d. assumption that $P_i^{(k)} = P_j^{(k)}$, we know that

$$\mathbb{P}\Big(\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\mathrm{IID},(k)}\Big) \to 1-\alpha.$$

For evaluation, we will use full observations from the target site. Each grey bar in (P,a) and (M,a) of Figure 3 is computed via

$$\widehat{\mathrm{Cov}}_k^{\mathrm{IID}} := \frac{1}{N_k(N_k-1)} \sum_{i=1}^{N_k} \sum_{j\neq i} \mathbb{1}\left\{\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\mathrm{IID},(k)}\right\}.$$

Thus, if the i.i.d. assumption holds, we will expect $\widehat{\mathrm{Cov}}_k^{\mathrm{IID}} \approx 1-\alpha$.

Estimation for CovShift (DR). For any site pair (i,j) in a hypothesis k, we first describe how to construct a point estimate for generalization via reweighting. We denote the estimator as $\widehat{\theta}_{i\to j}^{(k)}$ when generalizing from site i with full observations to site j with only covariate information. We will employ cross-fitting (Chernozhukov et al., 2018) to allow the use of flexible machine learning algorithms such as random forests in estimating the covariate shift weights and conditional mean functions.

First, we randomly split the data $\mathcal{D}_i^{(k)}$ and covariates in $\mathcal{D}_j^{(k)}$ into two equally-sized halves each. We use one half of data to estimate the covariate shift function $\mathrm{d}P_{j,X}^{(k)}/\,\mathrm{d}P_{i,X}^{(k)}(x)$ via $\widehat{w}(x)$, and the conditional mean function $\varphi(x) := \mathbb{E}[\phi^{(k)}(X,Y,T)\,|\,X=x]$ via $\widehat{\varphi}(x)$. These functions will be applied to the other fold of data, and construct the reweighted estimator

$$\widehat{\theta}_{i\to j}^{(k)} = \frac{1}{n_i^{(k)}} \sum_{\ell} \widehat{w}(X_\ell^{(i,k)}) \cdot \left\{\phi^{(k)}(X_\ell^{(i,k)}, Y_\ell^{(i,k)}, T_\ell^{(i,k)}) - \widehat{\varphi}(X_\ell^{(i,k)})\right\} + \frac{1}{n_j^{(k)}} \sum_{\ell=1}^{n_j^{(k)}} \widehat{\varphi}(X_\ell^{(j,k)}). \tag{11}$$

Following Jin and Rothenhäusler (2024), if the covariate shift condition holds, one can show that for the t-test and paired t-test considered in this work, as long as $\widehat{w}$ and $\widehat{\varphi}$ converge to the true covariate shift weight function and the true conditional mean function with a rate of $o_P((n_i^{(k)})^{-1/4})$, it holds that

$$\frac{\widehat{\theta}_j^{(k)} - \widehat{\theta}_{i\to j}^{(k)}}{\widehat{\sigma}_{i\to j}^{(k),\mathrm{CovShift}}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\widehat{\sigma}_{i\to j}^{(k),\mathrm{CovShift}}$ is any consistent estimator for $\sigma_{i\to j}^{(k),\mathrm{CovShift}}$, and

$$(\sigma_{i\to j}^{(k),\mathrm{CovShift}})^2 = \frac{\mathbb{E}_i^{(k)}[w(X)^2 \cdot (\phi^{(k)}(X,Y,T) - \varphi^{(k)}(X))^2]}{n_i^{(k)}} + \frac{\mathbb{E}_i^{(k)}[w(X) \cdot (\phi^{(k)}(X,Y,T) - \varphi^{(k)}(X))^2]}{n_j^{(k)}}.$$

As such, we construct the prediction interval for CovShift (DR) via

$$\widehat{C}_{i\to j}^{\mathrm{CovShift},(k)} = \widehat{\theta}_{i\to j}^{(k)} \pm q_{1-\alpha/2} \cdot \widehat{\sigma}_{i\to j}^{(k),\mathrm{CovShift}},$$

where $\widehat{\sigma}_{i\to j}^{(k),\mathrm{CovShift}}$ is constructed by plugging in $\widehat{w}$ and $\widehat{\varphi}$ into the definition of $\sigma_{i\to j}^{(k),\mathrm{CovShift}}$. Based on the arguments above, assuming covariate shift, under standard assumptions above, we would have

$$\mathbb{P}\Big(\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\mathrm{CovShift},(k)}\Big) \to 1-\alpha.$$

For evaluation, we will use full observations from the target site. Each green bar in (P,a) and (M,a) of Figure 3 is computed via

$$\widehat{\mathrm{Cov}}_k^{\mathrm{CovShift}} := \frac{1}{N_k(N_k-1)} \sum_{i=1}^{N_k} \sum_{j\neq i} \mathbb{1}\left\{\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\mathrm{CovShift},(k)}\right\}.$$

If the covariate shift assumption holds, we expect $\widehat{\mathrm{Cov}}_k^{\mathrm{CovShift}} \approx 1-\alpha$ under standard regularity conditions.

Estimation for CovShift (EB). The idea for constructing the point estimate for CovShift (EB) is similar to CovShift (DR), with the only exception that we obtain the weights $\widehat{w}_\ell^{(i,k)}$ are obtained by entropy balancing (Hainmueller, 2012) following the procedure in Jin et al. (2023), while $\widehat{\sigma}_{i\to j}^{(k),\text{CovShift}}$ is obtained in the same way as in CovShift (DR). We then construct the point estimate

$$\widehat{\theta}_{i\to j}^{(k)} = \frac{1}{n_i^{(k)}} \sum_\ell \widehat{w}_\ell^{(i,k)} \cdot \phi^{(k)}(X_\ell^{(i,k)}, Y_\ell^{(i,k)}, T_\ell^{(i,k)}) \tag{12}$$

and prediction interval

$$\widehat{C}_{i\to j}^{\text{CovShift},(k)} = \widehat{\theta}_{i\to j}^{(k)} \pm q_{1-\alpha/2} \cdot \widehat{\sigma}_{i\to j}^{(k),\text{CovShift}}.$$

Following Jin et al. (2023), assuming covariate shift, if the weight is a logistic function of the covariates or if $\varphi(x)$ is a linear function of the covariates, we would have

$$\mathbb{P}\Big(\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\text{CovShift},(k)}\Big) \to 1-\alpha.$$

For evaluation, we will use full observations from the target site. For evaluation, each purple bar in (P,a) and (M,a) of Figure 3 is computed via

$$\widehat{\text{Cov}}_k^{\text{CovShift}} := \frac{1}{N_k(N_k-1)} \sum_{i=1}^{N_k} \sum_{j\neq i} \mathbb{1}\left\{\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\text{CovShift},(k)}\right\}$$

using the prediction intervals for CovShift (EB). Thus, if the covariate shift assumption holds, we will expect $\widehat{\text{Cov}}_k^{\text{CovShift}} \approx 1-\alpha$ under the stated linear assumptions which are standard in the balancing literature. We note that CovShift (EB) is more stable than CovShift (DR) for small-to-moderate sample sizes, as is the case for the datasets analyzed in this work.

B.3 Estimation for distribution shift measures

We then proceed to detail the estimation procedure for our new distribution shift measures. To begin with, we note the following decomposition by Jin et al. (2023), which measures the contributions of distribution shifts (on the super-population level) to effect discrepancy:

$$\theta(Q) - \theta(P) = \underbrace{\theta(Q) - \theta(Q_X \times P_{Y|X})}_{\text{Contribution of conditional shift}} + \underbrace{\theta(Q_X \times P_{Y|X}) - \theta(P)}_{\text{Contribution of covariate shift}} \tag{13}$$

where $\theta(\cdot)$ is the functional for the parameter of interest, P is the source distribution, Q is the target distribution, and $Q_X \times P_{Y|X}$ is the reweighted distribution. Note that the contribution of conditional shift will be zero under the covariate shift assumption (Definition 3.1). In multi-site replication studies, for generalizing estimates for a hypothesis k from site i to site j, we will take $\theta = \theta^{(k)}$, $P = P_i^{(k)}$, and $Q = P_j^{(k)}$.

Computing the conditional shift measure. Following (13), we recall our definitions of the population-level conditional shift measure (for generalizing from P to Q) in Section 3.1, denoted as

$$t_{Y|X} := \frac{\Delta_{Y|X}}{s_{Y|X}}, \quad \Delta_{Y|X} = \theta(Q) - \theta(Q_X \times P_{Y|X}), \quad s_{Y|X}^2 = \text{Var}_P\left(\phi(X,Y,T) - \mathbb{E}_P[\phi(X,Y,T)\,|\,X]\right),$$

where the contributions of the conditional shift is rescaled by the standard deviation of its influence function to ensure scale invariance.

Following the notations in the preceding subsection, we compute the conditional shift measure from site i to site j in hypothesis k via the following formula:

$$\widehat{t}_{Y|X}^{i\to j,(k)} = \frac{\widehat{\Delta}_{Y|X}^{i\to j,(k)}}{\widehat{s}_{Y|X}^{i\to j,(k)}} := \frac{\widehat{\theta}_j^{(k)} - \widehat{\theta}_{i\to j}^{(k)}}{\widehat{s}_{Y|X}^{i\to j,(k)}} \tag{14}$$

where $\widehat{\theta}_j^{(k)}$ is the target estimator for $\theta(Q)$, $\widehat{\theta}_{i\to j}^{(k)}$ is the doubly robust estimator (11) or the entropy balancing estimator (12) in the previous part, so that $\widehat{\Delta}_{Y|X}$ is an estimator for the contribution of conditional shift. In addition, $\widehat{s}_{Y|X}^{i\to j,(k)}$ is a consistent estimator for $\mathrm{Var}_P\left(\phi(X,Y,T)-\mathbb{E}_P[\phi(X,Y,T)\,|\,X]\right)^{1/2}$, which we detail in Appendix E.2 and introduce its fast convergence properties.

Computing the covariate shift measure. Finally, we compute the "stabilizes" covariate shift measure as mentioned in the main text. Namely, supposing there are L covariates $\{X_\ell\}_{\ell=1}^L$, we compute

$$\widehat{t}_X^{i\to j,(k)} := \sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\Big(\frac{\widehat{\mathbb{E}}_Q[X_\ell]-\widehat{\mathbb{E}}_P[X_\ell]}{\widehat{\sigma}_P(X_\ell)}\Big)^2}, \tag{15}$$

where $\widehat{\sigma}_P(X_\ell)$ is the empirical standard deviation of X_ℓ in the source dataset. Note that $\widehat{t}_X^{i\to j,(k)}$ is pivotal as $n_i^{(k)}, n_j^{(k)}\to\infty$ under the i.i.d. assumption.

Computing the ratios. After computing the two measures $\widehat{t}_{Y|X}^{i\to j,(k)}$ and $\widehat{t}_X^{i\to j,(k)}$, we simply measure their relative strengths by the ratio

$$\widehat{r}_{i\to j}^{(k)} = \widehat{t}_{Y|X}^{i\to j,(k)}/\widehat{t}_X^{i\to j,(k)}.$$

Alternative definitions of distribution shift measures will be explored in Appendix D, yet we find they either (i) are scale-dependent (hence interpretation is sensitive the definition of the parameter functional $\theta(\cdot)$), or (ii) lead to unstable performance in estimation and effect generalization.

Idea for effect generalization based on distribution shift measures. Finally, we recall the high-level idea of effect generalization based on our distribution shift measures. If the distribution of the ratio $\widehat{r}^{i\to j,(k)}$ (which depends on both sampling uncertainty and distribution shifts) can be characterized, so that one can find upper and lower bounds L and U (either by asymptotic distribution or data-adaptive calibration) such that (approximately)

$$\mathbb{P}\Big(L\le \widehat{r}_{i\to j}^{(k)}\le U\Big)\ge 1-\alpha, \tag{16}$$

then, inverting this fact would give a prediction interval for $\widehat{\theta}_j^{(k)}$, which is

$$\widehat{C}_{i\to j}^{(k)} = \Big[\widehat{\theta}_{i\to j}^{(k)} + L\cdot\widehat{t}_X^{i\to j,(k)}\cdot\widehat{s}_X^{i\to j,(k)},\ \widehat{\theta}_{i\to j}^{(k)} + U\cdot\widehat{t}_X^{i\to j,(k)}\cdot\widehat{s}_X^{i\to j,(k)}\Big]. \tag{17}$$

Above, except for L and U, all quantities can be estimated with full observations from site i and covariates from site j. Next, we will detail how L and U are calibrated in Section 4.

B.4 Estimation for effect generalization

In this part, we detail our estimation and evaluation procedures for effect generalization in Section 4. We first introduce the IID method and the Oracle method evaluated in both Figure 7 and Figure 8. Then, we introduce WorstCase and Ours methods for constant calibration and adaptive calibration in the two figures, respectively.

IID method. With the i.i.d. assumption, we construct prediction intervals as (10) for generalizing from site i to site j for hypothesis k. That is, we use no covariate information in the sites, and the empirical coverage of the IID method is mainly plotted for reference. For coverage and lengths, we average over all site pairs for a given hypothesis in all scenarios.

Oracle method. This method uses all site pairs to calibrate the range of $\widehat{r}^{i\to j,(k)}$, namely, we compute

$$L^{\mathrm{Orc},(k)} := \mathrm{Quantile}\Big(\alpha/2;\, \{\widehat{r}^{(k)}_{i\to j}\}_{i\neq j}\Big), \quad U^{\mathrm{Orc},(k)} := \mathrm{Quantile}\Big(1-\alpha/2;\, \{\widehat{r}^{(k)}_{i\to j}\}_{i\neq j}\Big)$$

for the bounds L and U in (16). As its name suggests, it is the ideal prediction interval when we have perfect knowledge of how distribution shifts between all sites for a hypothesis. Note that this approach uses much more information than available in a real generalization task, and is hence evaluated just for reference. For coverage and lengths, we average over all site pairs for a given hypothesis in all scenarios.

B.4.1 Constant calibration

Our method (constant calibration). We take constants $L=-1$ and $U=1$ in (16), i.e., we believe that the conditional shift is upper bounded by the covariate shift. This leads to the prediction interval

$$\widehat{C}^{\mathrm{Ours},(k)}_{i\to j} = \Big[\widehat{\theta}^{(k)}_{i\to j} - \widehat{t}^{\,i\to j,(k)}_X \cdot \widehat{s}^{\,i\to j,(k)}_X,\ \widehat{\theta}^{(k)}_{i\to j} + \widehat{t}^{\,i\to j,(k)}_X \cdot \widehat{s}^{\,i\to j,(k)}_X\Big],$$

which is computable in a real generalization task with $\mathcal{D}^{(k)}_i$ and covariates in $\mathcal{D}^{(k)}_j$. The barplots in Figure 8 show the empirical coverage

$$\frac{1}{N_k(N_k-1)}\sum_{i\neq j}\mathbb{1}\Big\{\widehat{\theta}^{(k)}_j \in \widehat{C}^{\mathrm{Ours},(k)}_{i\to j}\Big\}$$

and average lengths

$$\frac{1}{N_k(N_k-1)}\sum_{i\neq j}\Big|\widehat{C}^{\mathrm{Ours},(k)}_{i\to j}\Big|$$

after normalization by the largest average length for each hypothesis k.

Worst-case method. We also evaluate the performance of worst-case bounds on the conditional shift, calibrated with data at hand. These worst-case bounds estimate the range of target parameters under the constraint that the unknown conditional shift is bounded in a KL-divergence ball.

Before we introduce our approach, we first remark two aspects about this approach:

1. Rigorously speaking, this is not a feasible generalization approach since we need full observations from all sites (especially the outcomes from the target site) to calibrate the KL bound $\widehat{\mathrm{KL}}^{(k)}_{\mathrm{upp}}$, which is typically not available in a real generalization task. As such, we mainly use it for reference.
2. There are several approximations in this approach, since the estimation uncertainty in $\widehat{\mathrm{KL}}^{(k)}_{\mathrm{upp}}$ is not accounted for, and it usually needs to account for larger uncertainty to cover the actual estimator than the underlying parameter. Thus, the intervals we obtain here can be viewed as underestimating the actual uncertainty, and a rigorous approach would construct even wider intervals.

Specifically, let P be the source distribution and Q be the target distribution. The strength of conditional shift can be characterized by the KL divergence between the reweighted distribution $Q_X \times P_{Y|X}$ and the target distribution Q, i.e.,

$$\begin{aligned}\mathrm{KL}\big(Q \,\|\, Q_X \times P_{Y|X}\big) &= \mathbb{E}_{Q_X\times P_{Y|X}}\left[\frac{\mathrm{d}Q}{\mathrm{d}(Q_X\times P_{Y|X})}(X,Y)\cdot\log\frac{\mathrm{d}Q}{\mathrm{d}(Q_X\times P_{Y|X})}(X,Y)\right]\\ &= \mathbb{E}_{Q_X\times P_{Y|X}}\left[\frac{\mathrm{d}Q_{Y|X}}{\mathrm{d}P_{Y|X}}(X,Y)\cdot\log\frac{\mathrm{d}Q_{Y|X}}{\mathrm{d}P_{Y|X}}(X,Y)\right].\end{aligned}$$

To estimate this quantity, we first use a classification model to estimate the joint density ratio $\mathrm{d}Q_{X,Y}/\mathrm{d}P_{X,Y}(x,y)$ via $\widehat{w}_{X,Y}(\cdot)$, and then the covariate density ratio $\mathrm{d}Q_X/\mathrm{d}P_X(x)$ via $\widehat{w}_X(\cdot)$. Then, we estimate the conditional

density ratio $\mathrm{d}Q_{Y|X}/\,\mathrm{d}P_{Y|X}(x,y)$ via $\widehat{w}_{X,Y}(x,y)/\widehat{w}_X(x)$, and plug in the definition to obtain an estimator for the KL-divergence, denoted as $\widehat{\mathrm{KL}}_{i\to j}^{(k)}$ when taking $P = P_i^{(k)}$ and $Q = P_j^{(k)}$.

After obtaining $\widehat{\mathrm{KL}}_{i\to j}^{(k)}$ for all pairs of studies, we calibrate an upper bound for the conditional KL-divergence for any given hypothesis k via

$$\widehat{\mathrm{KL}}_{\mathrm{upp}}^{(k)} := \mathrm{Quantile}\Big(0.99;\, \{\widehat{\mathrm{KL}}_{i\to j}^{(k)}\}_{i\neq j}\Big),$$

where we take the 0.99 quantile to avoid outliers. Then, we compute upper and lower bounds for the parameters $\theta(P_j^{(k)})$ by solving the following optimization program:

$$\begin{aligned} &\text{Maximize/minimize} \quad \theta(\bar{Q}) \\ &\text{Subject to} \quad \mathrm{KL}\big(\bar{Q} \,\|\, Q_X \times P_{Y|X}\big) \le \widehat{\mathrm{KL}}_{\mathrm{upp}}^{(k)}. \end{aligned}$$

Algorithms for solving the above program with data are standard in the literature; see, e.g., Hu and Hong (2013). We then use the maximized and minimized objective as upper and lower bounds for the target estimator, giving rise to the prediction interval

$$\widehat{C}_{i\to j}^{\mathrm{KL},(k)} := \Big[\widehat{U}_{i\to j}^{\mathrm{KL},(k)},\, \widehat{U}_{i\to j}^{\mathrm{KL},(k)}\Big].$$

The barplots in Figure 8 show the empirical coverage

$$\frac{1}{N_k(N_k-1)} \sum_{i\neq j} \mathbb{1}\Big\{\widehat{\theta}_j^{(k)} \in \widehat{C}_{i\to j}^{\mathrm{KL},(k)}\Big\}$$

and average lengths

$$\frac{1}{N_k(N_k-1)} \sum_{i\neq j} \Big|\widehat{C}_{i\to j}^{\mathrm{KL},(k)}\Big|$$

after normalization for each hypothesis k.

B.4.2 Data-adaptive calibration

Data-adaptive calibration uses separate datasets, which we assume to be available at a generalization task, to calibrate the strength of distribution shift. We will follow the notations in the preceding part.

Ours (data-adaptive calibration). In Figure 8, we assume that data for hypothesis $k_1, \ldots, k_t$ are available when we want to generalize between sites for a new hypothesis k_{t+1}. Thus, we calibrate the lower and upper bounds in (16) at step t by

$$L^{\mathrm{Ours},(t)} := \mathrm{Quantile}\Big(\alpha/2;\, \{\widehat{r}_{i\to j}^{(k_s)}\}_{i\neq j, s\le t}\Big), \quad U^{\mathrm{Ours},(t)} := \mathrm{Quantile}\Big(1-\alpha/2;\, \{\widehat{r}_{i\to j}^{(k_s)}\}_{i\neq j, s\le t}\Big). \tag{18}$$

The idea is that if the distribution of $\widehat{r}_{i\to j}^{(k)}$ is "pivotal" across hypothesis, using data for other hypotheses (other outcomes) to calibrate new hypotheses will lead to reliable coverage. We then construct prediction intervals for sites in new hypotheses k_s, $s > t$ by

$$\widehat{C}_{i\to j}^{(k_s)} = \Big[\widehat{\theta}_{i\to j}^{(k_s)} + L^{\mathrm{Ours},(t)} \cdot \widehat{t}_X^{\,i\to j,(k_s)} \cdot \widehat{s}_X^{\,i\to j,(k_s)},\; \widehat{\theta}_{i\to j}^{(k_s)} + U^{\mathrm{Ours},(t)} \cdot \widehat{t}_X^{\,i\to j,(k_s)} \cdot \widehat{s}_X^{\,i\to j,(k_s)}\Big].$$

The coverage and lengths of them are similarly evaluated. We also randomly order the hypotheses $(k_1, \ldots, k_t)$ and evaluate for ten times.

Worst-case method. Similar to the previous worst-case method, we will calibrate an upper bound of conditional shift and compute prediction intervals. Here, the upper bound will be calibrated with the observed data. Specifically, given all sites for hypotheses $\{k_1, \ldots, k_t\}$, we compute individual KL divergences following Section B.4.1, and then compute

$$\widehat{\mathrm{KL}}_{\mathrm{upp}}^{(t)} := \mathrm{Quantile}\Big(0.99; \{\widehat{\mathrm{KL}}_{i\to j}^{(k_s)}\}_{i\neq j, s\leq t}\Big). \tag{19}$$

For each future site pair (i, j) for hypothesis k_s, $s > t$, we solve an empirical version of

$$\begin{aligned} \text{Maximize/minimize} \quad & \theta(\bar{Q}) \\ \text{Subject to} \quad & \mathrm{KL}\big(\bar{Q} \,\|\, P_{j,X}^{(k_s)} \times P_{i,Y|X}^{(k_s)}\big) \leq \widehat{\mathrm{KL}}_{\mathrm{upp}}^{(t)}, \end{aligned}$$

and use the obtained maximized/minimized objectives as the upper/lower bounds. Note that this time, all quantities are computable in a real generalization task.

C Additional empirical results

C.1 Other calibration scenarios

In this part, we present additional calibration scenarios omitted in Section 4 in the main text, where we use certain observed data to calibrate the relative strength of covariate and conditional shifts, and construct prediction intervals in future generalization tasks. We omit the detailed procedures as they follow exactly the same ideas as Appendix B.4.2, except for the construction of the bounds L and U in (16).

Calibration with other sites. The second scenario is to calibrate the measures with existing sites involving all hypotheses for new sites. We randomly order the sites with $(j_1, \ldots, j_N)$ as a permutation of $(1, \ldots, N)$. Then, at each step $t \in \{1, \ldots, N-1\}$, we assume data from sites $\{j_1, \ldots, j_t\}$ for all the hypotheses are observed, and use the empirical quantiles of $\{\widehat{r}_{i\to j}^{(k)}\}_{i,j\in\{j_1,\ldots,j_t\},k=1,\ldots,K}$ as L and U in the construction of prediction intervals (9). Finally, for each pair of sites $j_1, j_2 \in \{j_{t+1}, \ldots, j_{29}\}$, we consider the task of generalization from fully observed data in site j_1 for hypothesis k to partially observed site j_2 for all hypotheses $j \in \{1, \ldots, 10\}$, using the aforementioned quantiles to construct prediction intervals following (9). On the other hand, the KL-divergence bound for `WorstCase` is also calibrated with these existing pairs in a way that is similar to (19).

The empirical coverage and PI lengths calibrated with other sites are reported in Figure 9. Again, the `WorstCase` method exhibits overcoverage and very wide intervals, while our method achieves valid coverage while being close to the `Oracle` method.

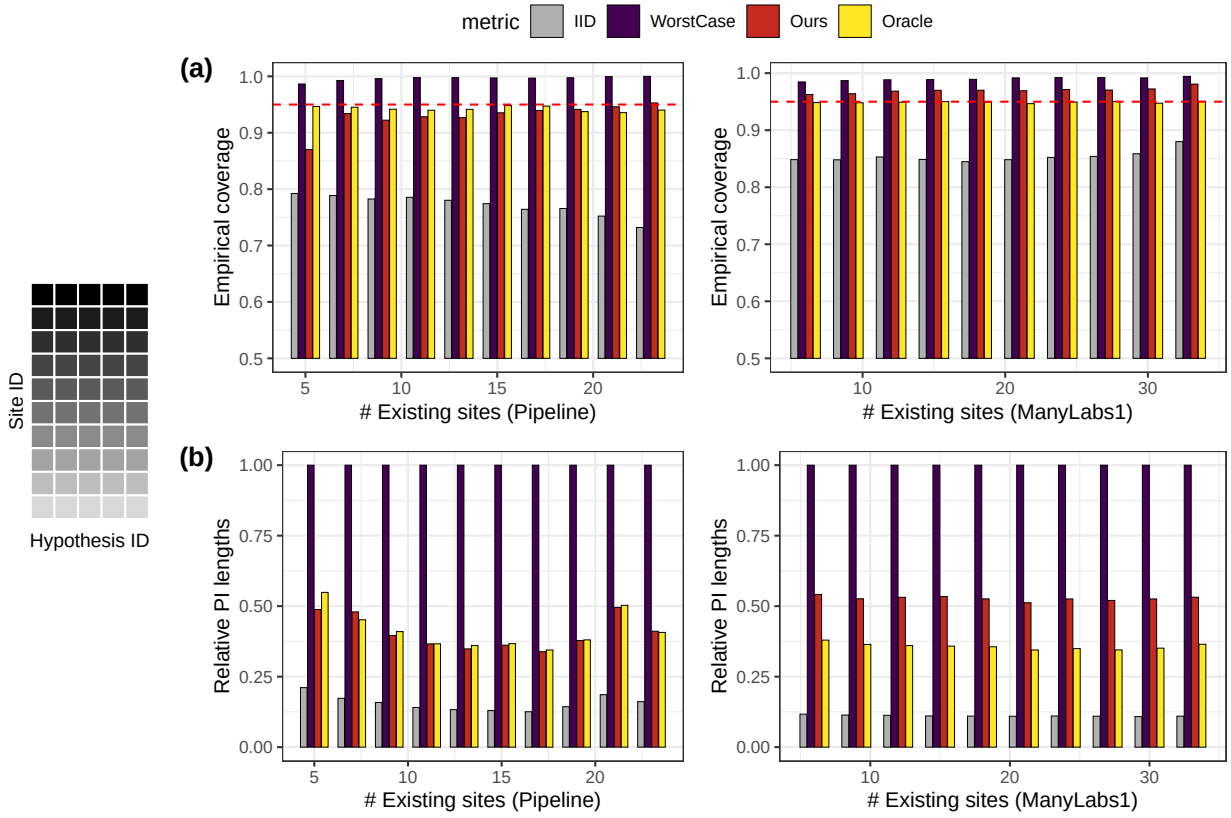

Figure 9: *Generalization in new studies based on distribution shift measures from other sites.* **Left**: *Illustration of data collection order, where dark color means earlier.* **Row (a)**: *Average coverage of prediction intervals using the Pipeline data (left) and ManyLabs 1 data (right).* **Row (b)**: *Average length of prediction intervals using the Pipeline data (left) and ManyLabs 1 data (right). Details are otherwise as Figure 8.*

Calibration with other sites and other hypotheses. The final scenario is the most challenging, where for a new generalization task, only data from other sites for other hypotheses are available. Specifically, we randomly order the sites by $(j_1, \ldots, j_N)$ and hypotheses by $(k_1, \ldots, k_{10})$. Then, at each step $t \in \{1, \ldots, 9\}$, data for studies $\{k_1, \ldots, k_K\}$ are available in sites $\{j_1, \ldots, j_{3t}\}$, we use the empirical quantiles of $\{\widehat{r}_{i \to j}^{(k)}\}_{i,j \in \{j_1,\ldots,j_{3t}\}, k \in \{k_1,\ldots,k_K\}}$ as L and U in the construction of prediction intervals (9). Finally, for each pair of sites $j_1, j_2 \in \{j_{3t+1}, \ldots, j_N\}$, we consider generalization from site j_1 to site j_2 for each hypothesis $k \in \{k_{t+1}, \ldots, k_K\}$, using the aforementioned quantiles to construct prediction intervals following (9). The KL-divergence bound for `WorstCase` is also calibrated with these existing pairs similar to (19).

The empirical coverage and length of PIs are reported in Figure 10. Similar to the observations in other scenarios, `WorstCase` is much more conservative, while our method achieves valid coverage with prediction interval lengths close to `Oracle`. This scenario is the most challenging among all, since the sites and hypotheses are entirely disjoint between existing data and new generalization tasks.

C.2 Relative strengths of distribution shift measures

In this part, we report additional results for the relative strengths of distribution shift measures in both projects, which complement Figure 4 in the main text. In particular, Figures 11 and 12 plots distribution shift measures in each hypothesis of the Pipeline project, computed with entropy balancing and the doubly robust estimator, respectively. Figures 13 and 14 plot those for the ManyLabs1 project.

Consistent with Figure 4, we see that the covariate shift upper bounds the conditional shift most of the

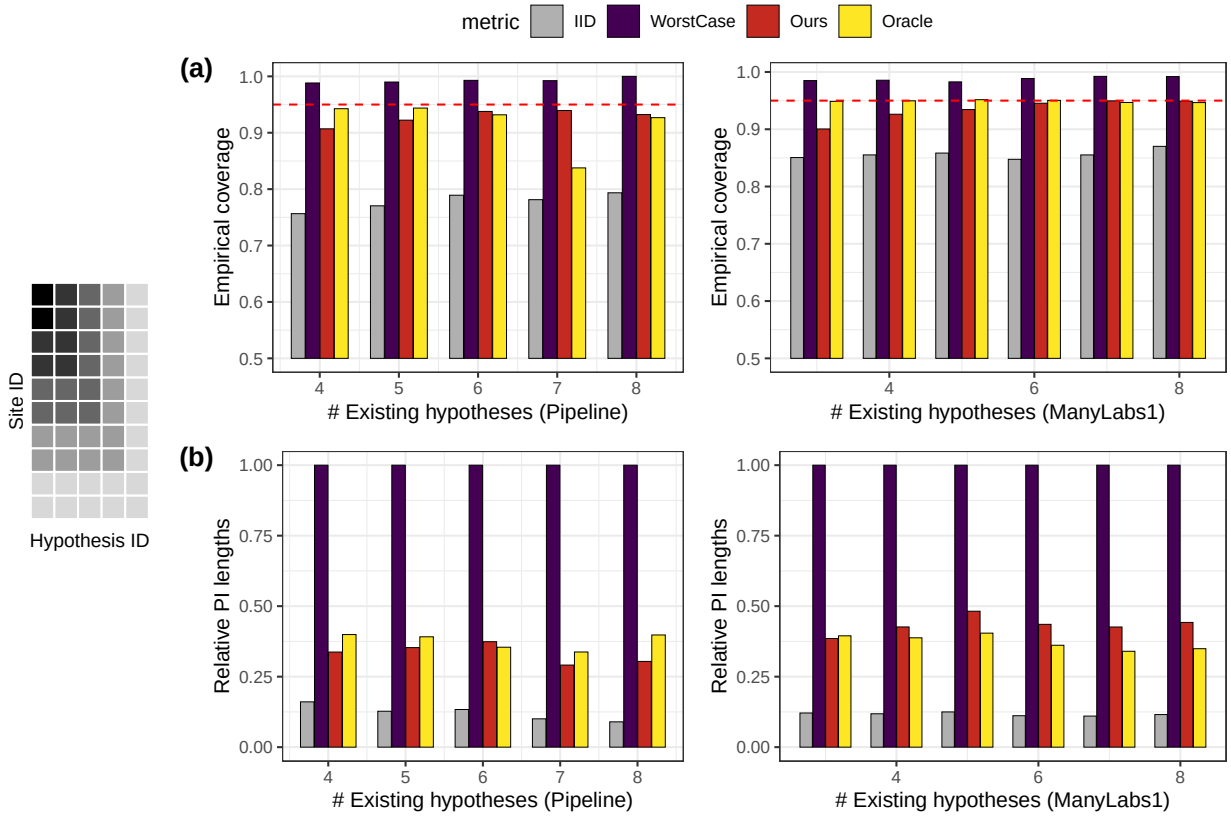

Figure 10: *Generalization in new studies based on distribution shift measures from other sites and other hypotheses for new sites and new hypotheses.* **Left**: *Illustration of data collection order, where dark color means earlier.* **Row (a)**: *Average coverage of prediction intervals using the Pipeline data (left) and ManyLabs 1 data (right).* **Row (b)**: *Average length of prediction intervals using the Pipeline data (left) and ManyLabs 1 data (right). Details are otherwise as Figure 8.*

time, but the balancing method tends to produce more stable estimates with small-to-moderate sample sizes.

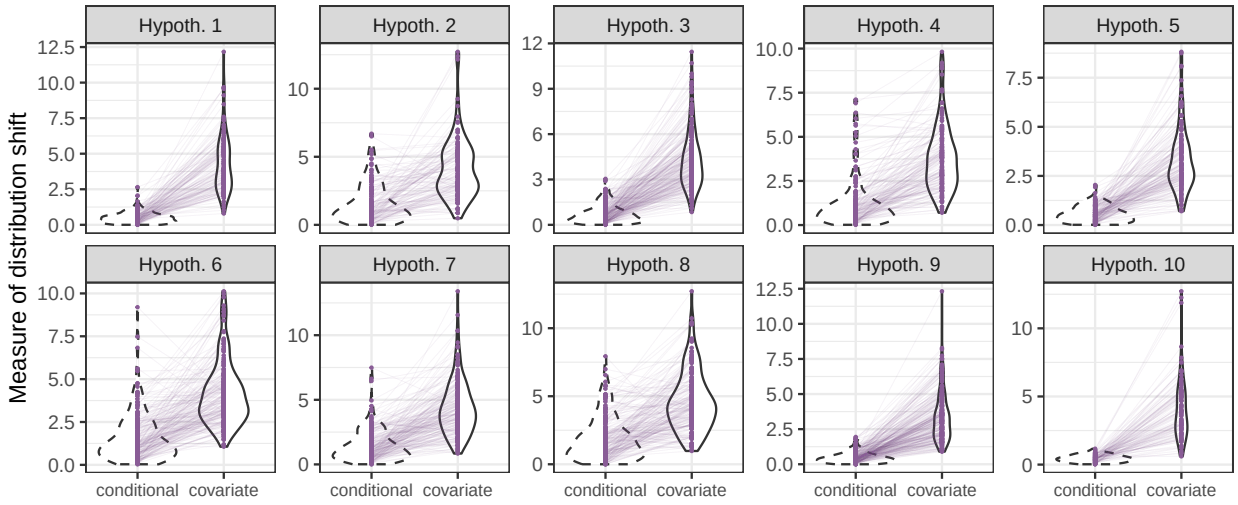

Figure 11: Distribution shift measures between all site pairs in each hypothesis in the Pipeline project, where the covariate shift adjustment uses entropy balancing.

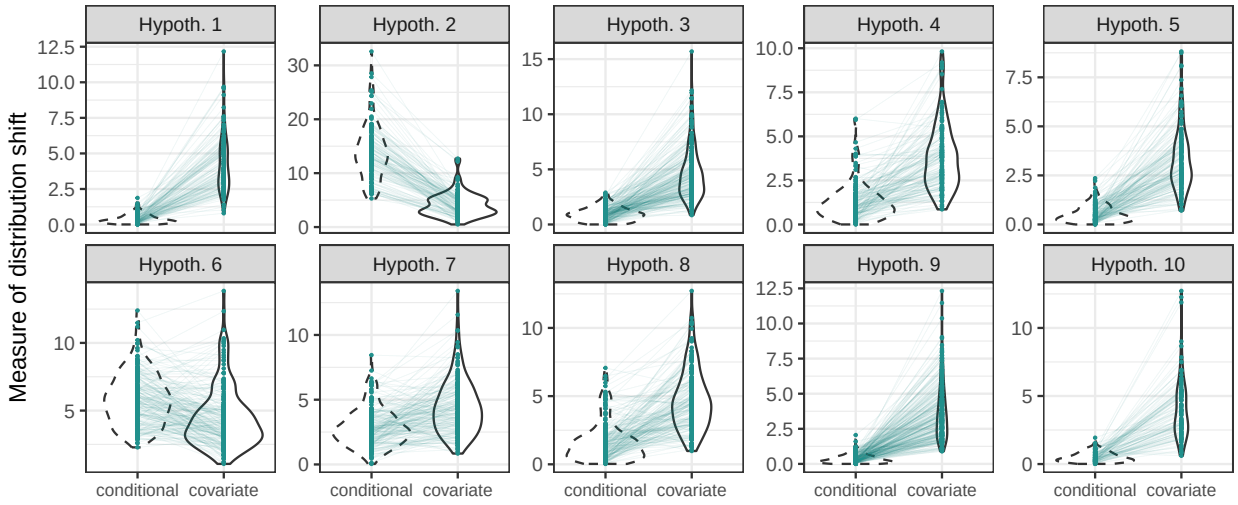

Figure 12: Distribution shift measures between all site pairs in each hypothesis in the Pipeline project, where the covariate shift adjustment uses the doubly robust estimator.

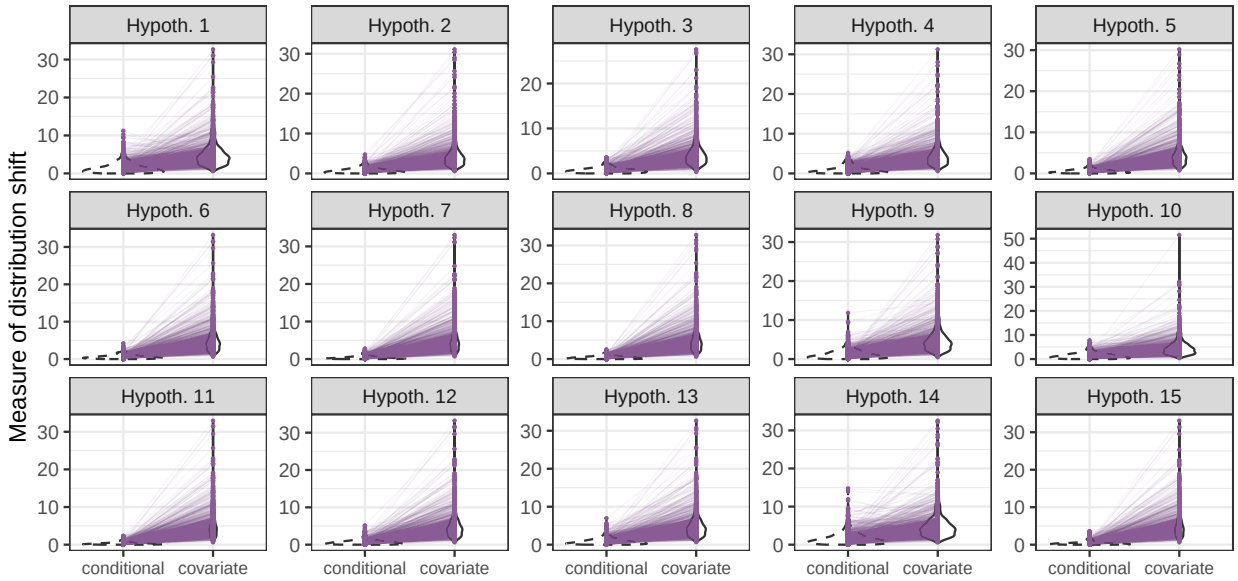

Figure 13: Distribution shift measures between all site pairs in each hypothesis in the ManyLabs1 project, where the covariate shift adjustment uses entropy balancing.

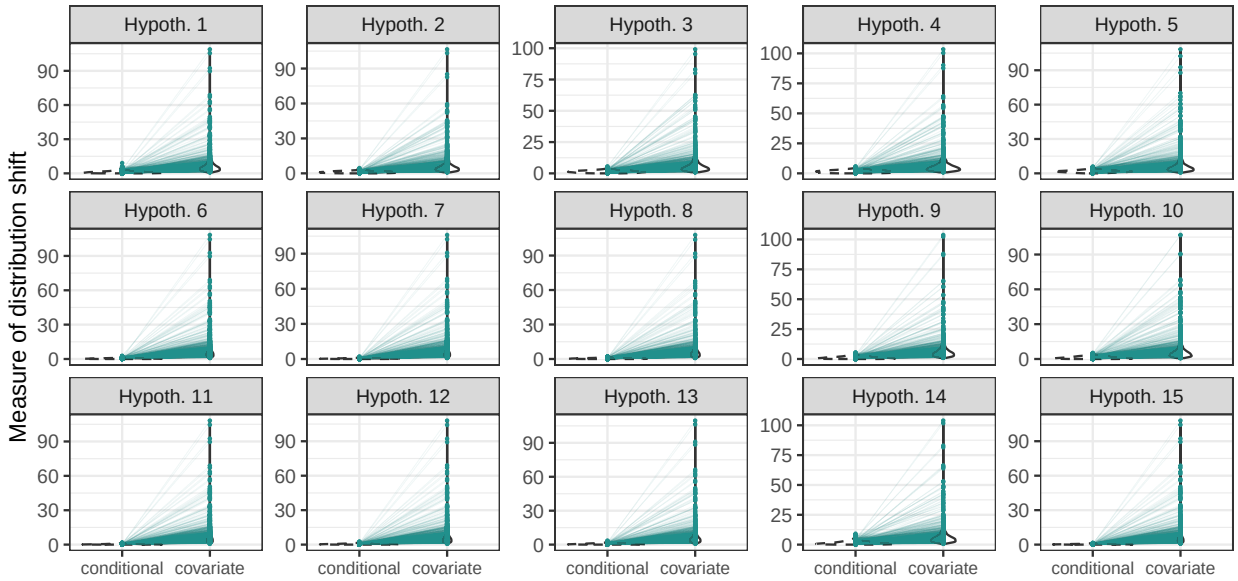

Figure 14: Distribution shift measures between all site pairs in each hypothesis in the ManyLabs1 project, where the covariate shift adjustment uses the doubly robust estimator.

D Exploring alternative distribution shift measures

This selection collects our results on exploring alternative distribution shift measures. Here we exclusively focus on entropy-balancing-based estimation for stability and conciseness, while doubly-robust estimation exhibits similar patterns. Also, we only present results for the Pipeline project for conciseness.

We introduce two sets of alternative distribution shift measures. By comparing our measures in the main text with them, we demonstrate the importance of (i) re-scaling by standard deviation to ensure scale invariance, and (ii) stabilizing the covariate shift measure in our definitions in Section 3.1.

D.1 Alternative distribution shift measures

Following the notations in Section 3.1, we consider generalizing from a site with distribution P to a site with distribution Q, and the parameter of interest has an influence function ϕ. Recall that $\phi_P(x) := \mathbb{E}_P[\phi \mid X = x]$.

Marginal shift measures. The first set of distribution shift measures follow Jin et al. (2023), which are not rescaled by the standard deviation. Namely,

$$\begin{aligned} \text{absolute of conditional shift} &= \mathbb{E}_Q[\phi(T,Y) - \phi_P(X)], \\ \text{absolute covariate shift} &= \mathbb{E}_Q[\phi_P(X)] - \mathbb{E}_P[\phi_P(X)]. \end{aligned}$$

These quantities describe the contributions of various types of distribution shifts to the discrepancy between effect estimates from two studies (sites) in Jin et al. (2023). Their estimation is already included in Appendix B.3, following which we denote the estimators as $\widehat{\Delta}_{Y|X}$ and $\widehat{\Delta}_X$, respectively. Namely,

$$\widehat{\Delta}_{Y|X}^{i\to j,(k)} := \widehat{\theta}_j^{(k)} - \widehat{\theta}_{i\to j}^{(k)}, \tag{20}$$

$$\widehat{\Delta}_{X}^{i\to j,(k)} := \widehat{\theta}_{i\to j}^{(k)} - \widehat{\theta}_i^{(k)}, \tag{21}$$

where $\widehat{\theta}_{i\to j}^{(k)}$ is the reweighted estimator using full observations from site i and covariates from site j.

These unscaled measures may lack interpretability in certain cases. For one thing, the magnitude of these quantities depends on how sensitive the function ϕ is to shifts in the probability space: for instance, if $\phi(X)$ is highly heterogeneous, then even small changes in the distribution of X would lead to large values of absolute covariate shift. While this is meaningful for diagnosing how the effect discrepancy relies on the distribution shifts and guiding future data collection efforts as in Jin et al. (2023), this might be undesirable when we are interested in *understanding the distribution shift itself*.

We will see later that with marginal shift measures, the conditional shift is usually much larger than the covariate shift measure, which is consistent with the (somewhat pessimistic) findings in Jin et al. (2023) and a similar work of Cai et al. (2023). This is mainly due to the fact that $\mathrm{sd}(\phi - \phi_P(X))$ is much larger than $\mathrm{sd}(\phi_P(X))$, i.e., the explanatory power of X for the parameter is low. However, this hides the fact that the strength of perturbation to the probability space is indeed the other way.

Relative shift measures. The second set of distribution shift measures follow Section 3.1, but we adopt the relative conditional shift instead of the stabilized one. Thus, we call them relative shift measures. The estimation of the relative conditional shift is straightforward; following Appendix B.3, we use

$$\widehat{\Delta}_{\mathrm{rel},X}^{i\to j,(k)} = \frac{\widehat{\Delta}_X^{i\to j,(k)}}{\widehat{s}_{i,X}} = \frac{\widehat{\theta}_{i\to j}^{(k)} - \widehat{\theta}_i^{(k)}}{\widehat{s}_X^{i\to j,(k)}}, \tag{22}$$

where $\widehat{\theta}_{i\to j}^{(k)}$ is the reweighted estimator using entropy balancing or doubly robust estimator, and $\widehat{s}_X^{i\to j,(k)}$ is a consistent estimator for $\mathrm{sd}_P(\phi_P(X))$ which can be obtained following the estimation of $\mathrm{sd}_P(\phi - \phi_P(X))$.

The issue with $\widehat{\Delta}_{\mathrm{rel},X}^{i\to j,(k)}$ is that the quantity $\widehat{s}_X^{i\to j,(k)}$ can be extremely small in some cases when the explanatory power of X for ϕ is low, as typical in the datasets we study here. Thus, even if we also observe a bounded role of relative covariate shift for relative conditional shift, the estimation is so unstable that it is not appropriate to be used in generalization tasks.

Summary. We summarize the three sets of distribution shift measures in Table 5 for the ease of reference. We also include the notations for their ratios to be used in the next two subsections.

Name	Conditional shift measure	Covariate Shift measure	Shift ratio
Stabilized	$\widehat{t}_{Y\|X}$, (14)	$\widehat{t}_{Y\|X}$, (15)	$\widehat{r}^{\text{stab}}$
Relative	$\widehat{t}_{Y\|X}$, (14)	$\widehat{\Delta}_{\text{rel},X}$, (22)	$\widehat{r}^{\text{rel}}$
Marginal	$\widehat{\Delta}_{Y\|X}$, (20)	$\widehat{\Delta}_X$, (21)	$\widehat{r}^{\text{mgn}}$

Table 5: Summary of notations and estimations of distribution shift measures.

D.2 The importance of rescaling for the predictive role

In this part, we demonstrate that rescaling is important for revealing the predictive role of covariate shift for the unknown conditional shift.

Figure 15 plots the distribution (violin plots) and pairwise relations (connected segments) of each pair of distribution shift measures in Table 5 across all pairs of sites for Hypothesis 5 in the Pipeline project:

- First, in the left panel, the relationship between the marginal measures $\widehat{\Delta}_{Y|X}$ and $\widehat{\Delta}_X$ in each pair is somewhat arbitrary. This means knowledge of $\widehat{\Delta}_X$ does not necessarily allow to control $\widehat{\Delta}_{Y|X}$.
- The middle panel of Figure 15 shows that the relative measure of covariate shift $\widehat{\Delta}_{\text{rel},X}$ bounds the conditional shift measure $\widehat{t}_{Y|X}$ most of the time. This reveals the importance of normalization with standard deviation for interpretability. Without being scale-invariant, the marginal measures fail to reveal the predictive role since the conditional "sensitivity", quantified by $\text{sd}(\phi - \phi_P(X))$, is larger than $\text{sd}(\phi_P(X))$. However, estimated values of $\widehat{\Delta}_{\text{rel},X}$ can be extremely large, since $\text{sd}(\phi_P(X))$ and its estimated value can be tiny when the explanatory power of the covariates is low. This is also not desirable in practice as it will cause instability in downstream tasks such as effect generalization; we will explore this in the next part.
- Finally, the right panel of shows the stabilized measures introduced in the main text. They reveal the predictive role of covariate shift due to scale invariance; in addition, they are more stable than the relative measures, and the bounding role is tighter due to fewer extreme estimated values.

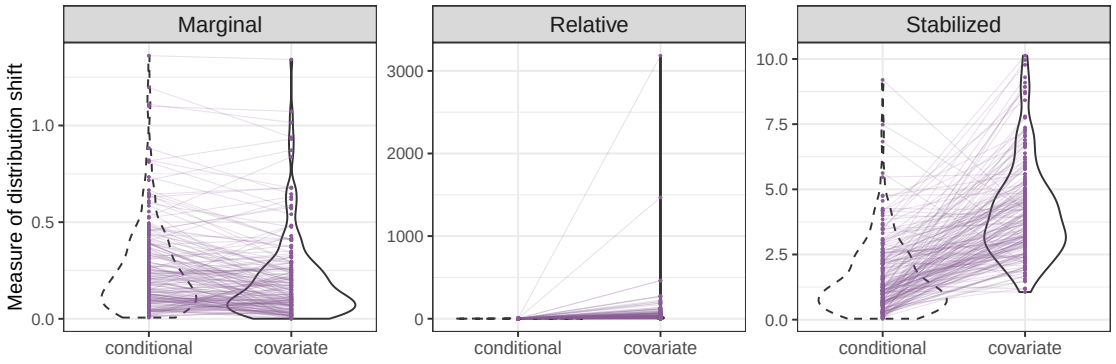

Figure 15: *(Relative) magnitude of measures of conditional shift (dashed) and covariate shift (solid) across all site pairs in hypothesis 5 of the Pipeline project, analyzed with entropy balancing.* **Left**: *Marginal measures* $\widehat{\Delta}_{Y|X}$ *and* $\widehat{\Delta}_X$. **Middle**: *Relative measures* $\widehat{t}_{Y|X}$ *and* $\widehat{\Delta}_{\text{rel},X}$. **Right**: *Stabilized measures* $\widehat{t}_{Y|X}$ *and* $\widehat{t}_X$.

With a similar goal as panel (c) of Figure 4 in the main text, we explore the stability of the three sets of distribution shift ratios by their within-hypothesis quantiles. If quantiles of the ratios are stable across hypotheses, then the ratio is "pivotal" and generalizable, meaning that external knowledge of the magnitude

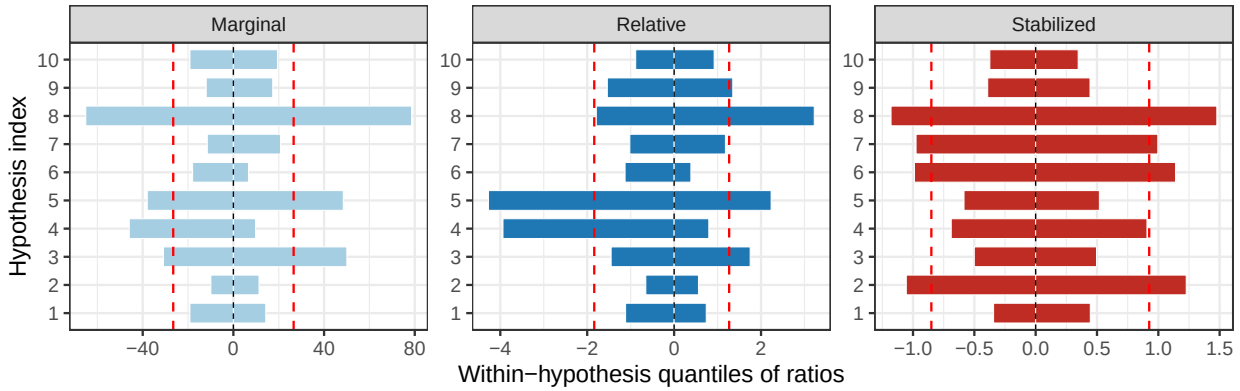

Figure 16: *Lower and upper within-hypothesis quantiles of ratios which, once known, lead to exact* 95% *empirical coverage of the prediction intervals for the Pipeline dataset. The left ends of the bar plot are the lower quantiles; the right ends are the upper quantiles. The red dashed lines are the global quantiles over all studies. Ideally, the quantiles should be invariant across studies for meaningful empirical calibration.* **Left**: *quantiles of* $\widehat{r}^{\text{mgn}}$ *(marginal);* **Middle**: *quantiles of* $\widehat{r}^{\text{rel}}$ *(relative);* **Right**: *quantiles of* $\widehat{r}^{\text{stab}}$ *(stabilized).*

of distribution shift ratios from other data sources may be useful for the data at hand. From Figure 4, we see that the quantiles of the marginal ratios and ralative ratios are quite variable. In contrast, the within-hypothesis quantiles of the stabilized ratio are more "pivotal"; they are stable across hypotheses and also close to the global quantile. We will see next the implications of such stability for generalization.

D.3 The importance of stability for generalization

We evaluate generalization tasks similar to Section 4 with the three sets of distribution shift measures. Again, similar to the ideas of (9), we construct prediction intervals for the target site estimator $\widehat{\theta}_j^{(k)}$ by calibrating lower and upper bounds for the ratios between each suite of distribution shifts. The detailed estimation procedures follow Appendix B.4.

Specifically, we aim to find lower and upper bounds for the ratios, such that (approximately)

$$\begin{aligned} &\mathbb{P}\Big(L^{\text{mgn}} \le \widehat{r}_{i\to j}^{\text{mgn},(k)} \le U^{\text{mgn}}\Big) \ge 1-\alpha, \quad \widehat{r}_{i\to j}^{\text{mgn},(k)} = \widehat{\Delta}_{Y|X}^{i\to j,(k)} / \widehat{\Delta}_{X}^{i\to j,(k)}, \\ &\mathbb{P}\Big(L^{\text{rel}} \le \widehat{r}_{i\to j}^{\text{rel},(k)} \le U^{\text{rel}}\Big) \ge 1-\alpha, \qquad \widehat{r}_{i\to j}^{\text{rel},(k)} = \widehat{t}_{Y|X}^{i\to j,(k)} / \widehat{\Delta}_{\text{rel},X}^{i\to j,(k)}, \end{aligned} \tag{23}$$

and the bounds for the ratio between stabilized measures in the main text follow the idea of (8).

Inverting the events in (23) and by the definition of the measures, we set the prediction intervals

$$\begin{aligned} \widehat{C}_{i\to j}^{\text{mgn},(k)} &:= \big[\ \widehat{\theta}_{i\to j}^{(k)} + \widehat{\Delta}_{X}^{i\to j,(k)} \cdot L^{\text{mgn}},\ \widehat{\theta}_{i\to j}^{(k)} + \widehat{\Delta}_{X}^{i\to j,(k)} \cdot U^{\text{mgn}}\ \big] \\ \widehat{C}_{i\to j}^{\text{rel},(k)} &:= \big[\ \widehat{\theta}_{i\to j}^{(k)} + \widehat{\Delta}_{\text{rel},X}^{i\to j,(k)} \cdot \widehat{s}_{Y|X}^{i\to j,(k)} \cdot L^{\text{rel}},\ \widehat{\theta}_{i\to j}^{(k)} + \widehat{\Delta}_{\text{rel},X}^{i\to j,(k)} \cdot \widehat{s}_{Y|X}^{i\to j,(k)} \cdot U^{\text{rel}}\ \big], \end{aligned}$$

and recall that $\widehat{C}_{i\to j}^{(k)}$ is the prediction interval (9) based on our shift measures in the main text. We then evaluate the empirical coverage and average length of these prediction intervals.

Oracle calibration. For reference, we evaluate the `Oracle` method in the main text for the three sets of shift measures, in order to show their performance in the most ideal case where the distribution of their ratios is perfectly known. Here, the L and U values in (23) are the empirical quantiles of the shift ratios between all site pairs within each hypothesis. The empirical coverage and average length of prediction intervals within each hypothesis are in Figure 17. All three sets of measures lead to perfect 0.95 coverage as expected. However, the prediction intervals by the stabilized measures several folds shorter (the y-axis is log-scaled for easier visualization), showing the importance of estimation stability.

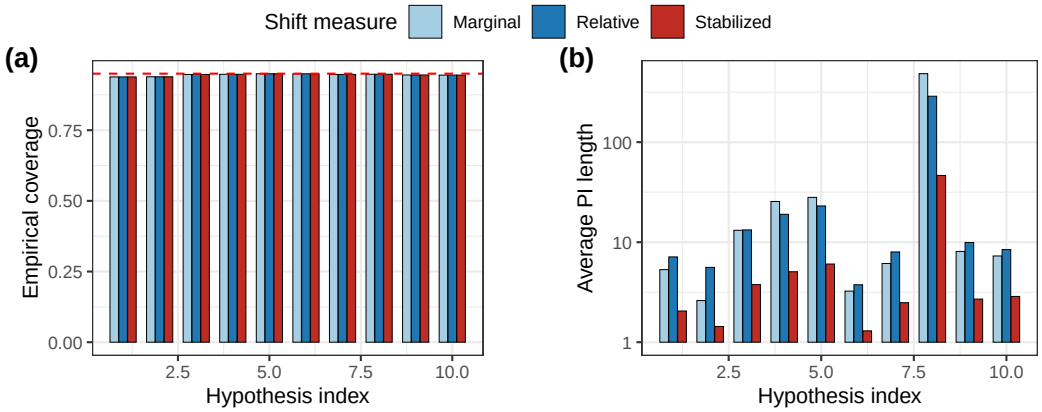

Figure 17: **Left**: *Empirical coverage of oracle calibrated prediction intervals at nominal level* $1-\alpha = 0.95$. *The coverage is ensured to be* 95% *since full observations are used.* **Right**: *Average length of prediction intervals for in-study calibrated prediction intervals at nominal level* $1-\alpha = 0.95$ *based on three measures. The y-axis on the right is log-scaled for visualization.*

Constant calibration. Similar to Section 4, here we simply take all three lower quantiles to be -1, and all three upper quantiles to be 1, with the belief that the covariate shift measure upper bounds the conditional shift measure. The hypothesis-wise coverage and average length of constant-calibrated prediction intervals are in Figure 18. It is not surprising that assuming $|\widehat{\Delta}_{Y|X}| \leq |\widehat{\Delta}_X|$ leads to poor coverage (marginal). Assuming that the conditional shift measure is bounded by the covariate shift measure leads to satisfying coverage for both the relative and stabilized measure. However, the stabilized measures lead to much shorter prediction intervals and slightly better coverage again due to stability.

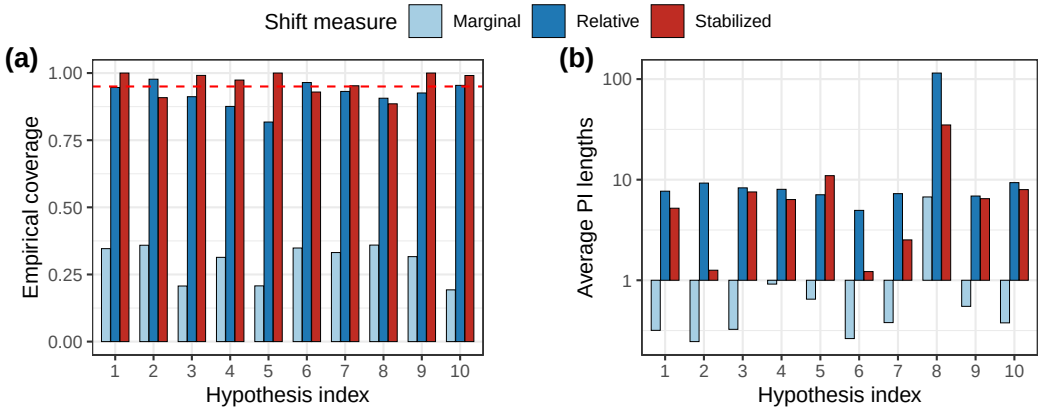

Figure 18: **Left**: *Empirical coverage of constant calibrated prediction intervals at nominal level* $1-\alpha = 0.95$. **Right**: *Average length of prediction intervals for constant calibrated prediction intervals at nominal level* $1-\alpha = 0.95$ *based on three measures. The y-axis on the right is log-scaled for visualization.*

Data-adaptive calibration. Finally, we consider the data-adaptive calibration scenario where full observations from other sites/hypotheses are available, which are used to compute the quantiles for a new generalization task. The specific methods are the same as Section 4 and Appendix C.1, with detailed procedures following Appendix B.4.

The first scenario is the same as Section 4 in the main text, where data for some other hypotheses in all sites are available, and the new generalization task involves new hypotheses. Figure 19 illustrates the

order of data collection, as well as the coverage and length of prediction intervals, averaged over 10 random draws of hypothesis ordering. We see that all three measures lead to satisfactory coverage, meaning that *the distribution shift measures tend to be "generalizable" across hypotheses/estimators/outcomes*. Yet, the stabilized measures still yield the shortest prediction intervals.

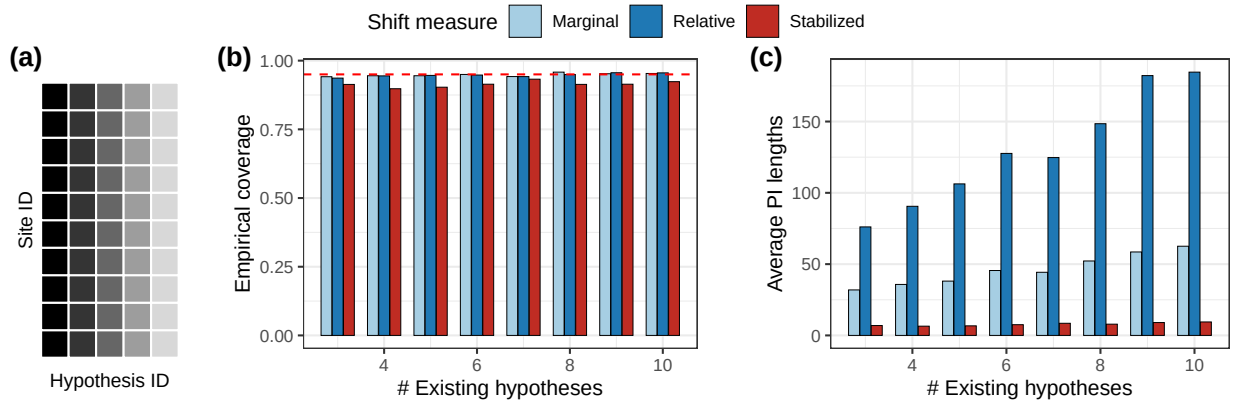

Figure 19: *Generalization based on distribution shift measures calibrated with data for other hypotheses in the same sites.* **Left**: *Illustration of data collection order, where dark color means earlier.* **Middle**: *Average coverage (bars) of prediction intervals over 10 random draws of study ordering. The red dashed line is the nominal level* 0.95. **Right**: *Average length of prediction intervals based on three sets of shift measures.*

The second scenario is to calibrate the measures with existing sites involving all hypotheses for new sites, same as "calibration with other sites" in Appendix C.1. Figure 20 presents the corresponding results. We see that all measures lead to satisfactory coverage, while the stabilized measures lead to much shorter prediction intervals.

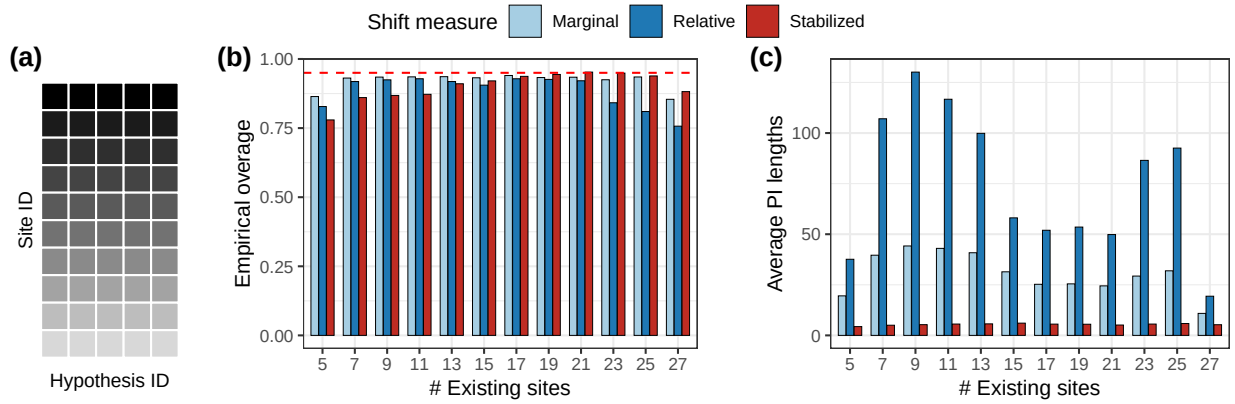

Figure 20: *Data collection order, average coverage and length of prediction intervals for generalization in new sites based on data from the same studies in other sites. Details are otherwise the same as in Figure 19.*

Finally, we use data from other sites for other hypotheses to calibrate the upper and lower bounds for the distribution shift measures, which is the same as "Calibration with other sites and other hypothesis" in Appendix C.1. Figure 21 presents the results for the third scenario. Due to the limited samples, we observe slight undercoverage when only two hypotheses and sites are available. Similar to other scenarios, the stabilized measure leads to much shorter prediction intervals than the other two.

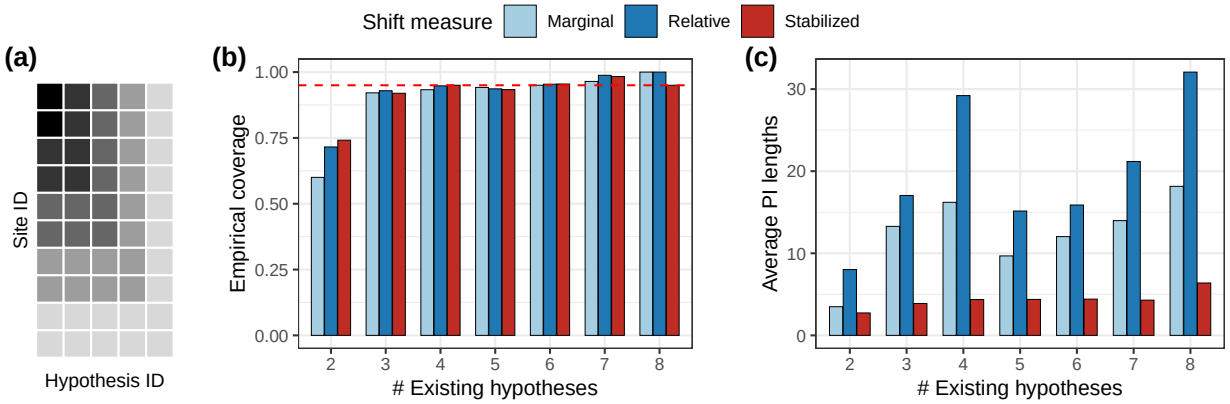

Figure 21: *Data collection order, average coverage and length of prediction intervals for generalization between new sites in new studies based on data for other studies from other sites. Details are otherwise as in Figure 19.*

E Technical details and proofs

E.1 Proof of distributional CLT

Proof. Let $n_Q(M)$ and $n_P(M)$ be sequences of natural numbers such that $n_Q(M)/M$ and $n_P(M)/M$ converge to positive real numbers. In the following, for simplicity, we will suppress the dependence of n_Q and n_P on M. We will first show the result for bounded, one-dimensional ϕ with $\mathbb{E}_P[\phi] = 0$. Define $\sigma_M^2 = \frac{1}{M}\sum_{m=1}^M \mathbb{E}_P[\phi|C_m^{(M)}]^2 - \mathbb{E}_P[\phi]^2$. As $M \to \infty$, by assumption we have $\sigma_M^2 \to \mathrm{Var}_P(\mathbb{E}_P[\phi|X,U]) =: \sigma^2$. If $\mathrm{Var}_P(\mathbb{E}_P[\phi|X,U]) = 0$, then $\mathbb{E}_P[\phi] = \mathbb{E}_Q[\phi]$, thus only uncertainty due to i.i.d. sampling remains and the statement of the theorem is trivial. Thus, in the following we will assume $\mathrm{Var}_P(\mathbb{E}_P[\phi|X,U]) > 0$. We can use $\mathbb{E}_P[\phi] = \frac{1}{M}\sum_{m=1}^M \mathbb{E}_P[\phi|C_m^{(M)}]$ to obtain

$$
\begin{aligned}
&\frac{\sqrt{M}(\mathbb{E}_Q[\phi] - \mathbb{E}_P[\phi])}{\sigma_M \mathrm{sd}(W)/E[W]} \\
&= \frac{\frac{M^{-1/2}\sum_{m=1}^M (W_m - 1/M\sum_{m'} W_{m'})\mathbb{E}_P[\phi|C_m^{(M)}]}{1/M\sum_m W_m}}{\sigma_M \mathrm{sd}(W)/E[W]} \\
&= \frac{M^{-1/2}\frac{\sum_{m=1}^M (W_m - 1/M\sum_{m'} W_{m'})\mathbb{E}_P[\phi|C_m^{(M)}]}{E[W]}}{\sigma_M \mathrm{sd}(W)/E[W]} + o_P(1/\sqrt{M}) \\
&= \frac{M^{-1/2}\sum_{m=1}^M (W_m - \mathbb{E}[W])\mathbb{E}_P[\phi|C_m^{(M)}]}{\sigma_M \mathrm{sd}(W)} + o_P(1/\sqrt{M}).
\end{aligned}
$$

We will now check Lindeberg's condition with $v_M = \sum_{m=1}^M \mathbb{E}_P[\phi|C_m^{(M)}]^2 - \mathbb{E}_P[\phi]^2$. By assumption $v_M/M \to \sigma^2 > 0$. Furthermore, by assumption $|\phi|_\infty \leq B$ for some constant $B > 0$. Let $\epsilon > 0$. Then

$$
\begin{aligned}
&\limsup_{M\to\infty} \frac{1}{v_M}\mathbb{E}[\sum_{m=1}^M (W_m - \mathbb{E}[W])^2(\mathbb{E}_P[\phi|C_m^{(M)}] - \mathbb{E}_P[\phi])^2 1_{|W_m - \mathbb{E}[W]||\mathbb{E}_P[\phi|C_m^{(M)}] - \mathbb{E}_P[\phi]| \geq \epsilon\sqrt{v_M}}] \\
&\leq \limsup_{M\to\infty} \frac{1}{M\sigma^2}\mathbb{E}[\sum_{m=1}^M (W_m - \mathbb{E}[W])^2 4B^2 1_{|W_m - \mathbb{E}[W]| \geq \sqrt{M}\sigma\epsilon/(4B)}] \\
&\leq \limsup_{M\to\infty} \frac{4B^2}{\sigma^2}\mathbb{E}[(W - \mathbb{E}[W])^2 1_{|W - \mathbb{E}[W]| \geq \sqrt{M}\sigma\epsilon/(4B)}]
\end{aligned}
$$

By dominated convergence, this term is zero. Thus, by Lindeberg's CLT,

$$\frac{\sqrt{M}(\mathbb{E}_Q[\phi] - \mathbb{E}_P[\phi])}{\sigma_M \mathrm{sd}(W)/E[W]} \xrightarrow{d} \mathcal{N}(0,1).$$

By Slutsky, we get the distributional CLT

$$\frac{\sqrt{M}(\mathbb{E}_P[\phi] - \mathbb{E}_Q[\phi])}{\mathrm{sd}_P(\mathbb{E}_P[\phi|X,U])\mathrm{sd}(W)/E[W]} \xrightarrow{d} \mathcal{N}(0,1).$$

We will now combine this result with uncertainty due to i.i.d. sampling. Recall that we consider the case where sampling uncertainty and distributional uncertainty are of the same order, i.e. n_Q/M and n_P/M converge to some positive constants.

$$\begin{aligned}
&\widehat{\mathbb{E}}_P[\phi(T,D)] - \widehat{\mathbb{E}}_Q[\phi(T,D)] \\
&= \underbrace{\widehat{\mathbb{E}}_P[\phi(T,D)] - \mathbb{E}_P[\phi(T,D)]}_{\text{use standard CLT for i.i.d. data}} - \underbrace{(\widehat{\mathbb{E}}_Q[\phi(T,D)] - \mathbb{E}_Q[\phi(T,D)])}_{\text{use Berry-Esseen}} + \underbrace{\mathbb{E}_P[\phi(T,D)] - \mathbb{E}_Q[\phi(T,D)]}_{\text{use distributional CLT}}
\end{aligned}$$

We can apply a standard CLT to the first term, since P is fixed and the sample mean $\widehat{\mathbb{E}}_P[\phi(T,D)]$ is independent of the remaining terms. For the remaining terms, one complication is that the distribution Q is not fixed, but shifts randomly.

Recall that for now we focus on bounded ϕ, which implies bounded third moments. We will now use Berry–Esseen. For any $x \in \mathbb{R}$, conditionally on the random shift $(W_m)_{m=1,\dots,M}$,

$$\sup_x \left| P\left(\frac{\sqrt{n_Q}}{\mathrm{sd}_Q(\phi)} (\widehat{\mathbb{E}}_Q[\phi(T,D)] - \mathbb{E}_Q[\phi(T,D)]) \le x \middle| (W_m)_{m=1,\dots,M} \right) - \Phi(x) \right| \le \frac{7.59 \mathbb{E}_Q[|\phi|^3]}{\mathrm{sd}_Q(\phi)^3 \sqrt{n_Q}} \tag{24}$$

By assumption ϕ is bounded, and by the distributional CLT above $\mathrm{sd}_Q(\phi) \xrightarrow{P} \mathrm{sd}_P(\phi) > 0$ for $M \to \infty$. Thus, conditionally on the random shift $(W_m)_{m=1,\dots,M}$,

$$\sup_x \left| P\left(\frac{\sqrt{n_Q}}{\mathrm{sd}_P(\phi)} (\widehat{\mathbb{E}}_Q[\phi(T,D)] - \mathbb{E}_Q[\phi(T,D)]) \le x \middle| (W_m)_{m=1,\dots,M} \right) - \Phi(x) \right| \to 0$$

Define

$$\begin{aligned}
Z &:= \mathbb{E}_P[\phi(T,D)] - \mathbb{E}_Q[\phi(T,D)], \\
Z' &:= \widehat{\mathbb{E}}_Q[\phi(T,D)] - \mathbb{E}_Q[\phi(T,D)].
\end{aligned}$$

Let $\delta_M^2 = \frac{1}{M}\frac{\mathrm{Var}(W)}{E[W]^2}$. By assumption, $n_Q/M \to \rho$ for some constant $\rho > 0$. Thus $n_Q \delta_K^2 \to \rho \mathrm{Var}(W)/E[W]^2$. As $M \to \infty$, for any $x \in \mathbb{R}$

$$\begin{aligned}
&P\left(\left(\frac{\mathrm{Var}_P(\phi)}{n_Q} + \sigma^2 \delta_M^2 \right)^{-1/2} (Z + Z') \ge x \right) \\
&= E\left[P\left(\frac{\sqrt{n_Q}}{\mathrm{sd}_P(\phi)} Z' \ge \frac{\sqrt{n_Q}}{\mathrm{sd}_P(\phi)} \left(\frac{\mathrm{Var}_P(\phi)}{n_Q} + \sigma^2 \delta_M^2 \right)^{1/2} x - \frac{\sqrt{n_Q}}{\mathrm{sd}_P(\phi)} Z \middle| (W_m)_{m=1,\dots,M} \right) \right] \\
&\overset{\text{Berry-Esseen}}{=} E\left[1 - \Phi\left(\left(1 + \rho \frac{\sigma^2 \mathrm{Var}(W)}{\mathrm{Var}_P(\phi) E[W]^2} \right)^{1/2} x - \frac{\sqrt{n_Q}}{\mathrm{sd}_P(\phi)} Z \right) \right] + o(1)
\end{aligned}$$

In the third line, we used that $\mathrm{sd}_Q(\phi) \xrightarrow{P} \mathrm{sd}_P(\phi)$. Here, Φ is the cdf of a standard Gaussian random variable. Using weak convergence of $\sqrt{M}Z \xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathrm{Var}(W)/E[W]^2)$, and using that Φ is a continuous bounded

function we get

$$E\left[1-\Phi\left(\left(1+\rho\frac{\sigma^2\mathrm{Var}(W)}{\mathrm{Var}_P(\phi)E[W]^2}\right)^{1/2}x-\frac{\sqrt{n_Q}}{\mathrm{sd}_P(\phi)}Z\right)\right]$$
$$\to E\left[1-\Phi\left(\left(1+\rho\frac{\sigma^2\mathrm{Var}(W)}{\mathrm{Var}_P(\phi)E[W]^2}\right)^{1/2}x-\sqrt{\rho}\cdot\frac{\sigma\mathrm{sd}(W)}{\mathrm{sd}_P(\phi)E[W]}G\right)\right],$$

where G is a standard Gaussian random variable. With constant $L=\sqrt{\rho}\sigma\mathrm{sd}(W)/(\mathrm{sd}_P(\phi)E[W])$ we can re-write this as

$$E[1-\Phi(\sqrt{1+L^2}x-LG)]=P(G'\geq\sqrt{1+L^2}x-LG)=P(\frac{G'+LG}{\sqrt{1+L^2}}\geq x)=1-\Phi(x),$$

where G' is a standard Gaussian random variable, independent of G. To summarize,

$$P\left(\left(\frac{\mathrm{Var}_P(\phi)}{n_Q}+\sigma^2\delta_M^2\right)^{-1/2}(Z+Z')\geq x\right)\to 1-\Phi(x).$$

Since the data from P is independent of the perturbation and the data from Q,

$$\left(\mathrm{Var}_P(\phi)\left(\frac{1}{n_P}+\frac{1}{n_Q}\right)+\sigma^2\delta_M^2\right)^{-1/2}\left(\widehat{\mathbb{E}}_P[\phi(T,D)]-\widehat{\mathbb{E}}_Q[\phi(T,D)]\right)\xrightarrow{d}\mathcal{N}(0,1).\tag{25}$$

We will now extend this result from bounded functions ϕ to square-integrable functions $\phi\in L^2(P)$. Define the bounded function

$$\phi_b=\phi 1_{|\phi|\leq b}+1_{|\phi|\geq b}\mathbb{E}_P\left[\phi\middle||\phi|\geq b\right].$$

We have $\mathbb{E}_P[\phi]=\mathbb{E}_P[\phi_b]$. Applying Chebychev conditionally on the random perturbation,

$$P\left(|\widehat{\mathbb{E}}_Q[\phi_b-\phi]-\widehat{\mathbb{E}}_Q[\phi_b-\phi]|\geq\epsilon\middle|(W_m)_{m=1,\ldots,M}\right)\leq\frac{\mathrm{Var}_Q(\phi_b-\phi)}{\epsilon^2}\left(\frac{1}{n_P}+\frac{1}{n_Q}\right)$$

Take expectations over the random perturbation $(W_m)_{m=1,\ldots,M}$ we obtain that for all $\epsilon>0$,

$$P(|\widehat{\mathbb{E}}_Q[\phi_b-\phi]-\mathbb{E}_Q[\phi_b-\phi]|\geq\epsilon)\leq\frac{\mathbb{E}_P[(\phi-\phi_b)^2]}{\epsilon^2}\left(\frac{1}{n_P}+\frac{1}{n_Q}\right)$$

Similarly, since $W_m\geq w_0$,

$$|\mathbb{E}_Q[\phi-\phi_b]|=|\sum_{m=1}^M\frac{W_m}{\sum_{m'}W_{m'}}\mathbb{E}_P[\phi-\phi_b|C_m]|\leq\frac{1}{Mw_0}|\sum_{m=1}^M W_m\mathbb{E}_P[\phi-\phi_b|C_m]|.$$

Applying Chebychev and using that $\mathbb{E}_P[\phi-\phi_b]=0$, we get

$$P(|\mathbb{E}_Q[\phi-\phi_b]|\geq\epsilon)\leq\frac{\sum_{m=1}^M(\mathbb{E}_P[\phi-\phi_b|C_m])^2}{w_0^2\epsilon^2M^2}\leq\frac{\mathrm{Var}_P(\phi-\phi_b)}{w_0^2\epsilon^2M}.$$

In the last equation, we used Jensen's inequality. Combining the two applications of Chebychev,

$$\begin{aligned}P(|\widehat{\mathbb{E}}_Q[\phi-\phi_b]|\geq\epsilon)&\leq P(|\mathbb{E}_Q[\phi-\phi_b]|\geq\epsilon/2)+P(|\widehat{\mathbb{E}}_Q[\phi-\phi_b]-\mathbb{E}_Q[\phi-\phi_b]|\geq\epsilon/2)\\&\leq\frac{4}{\epsilon^2}\left(\frac{1}{n_Q}+\frac{1}{n_P}+\frac{1}{w_0^2M}\right)\mathrm{Var}_P(\phi-\phi_b)\end{aligned}$$

Since by assumption $n_P(M) \sim n_Q(M) \sim M$, for any $\epsilon' > 0$, $\epsilon > 0$, we can choose a bounded function ϕ_b such that $P(\sqrt{M}|\widehat{E}_Q[\phi - \phi_b] - E_P[\phi - \phi_b]| \geq \epsilon) \leq \epsilon'$ for $M \to \infty$. Let

$$\sigma^2_{b,M} = \left(\mathrm{Var}_P(\phi_b) \left(\frac{1}{n_Q} + \frac{1}{n_P} \right) + \mathrm{Var}_P(\mathbb{E}_P[\phi_b|X,U])\delta^2_M \right).$$

Then, for any $\epsilon' > 0$ there exists a $\epsilon > 0$ such that as $M \to \infty$, we have

$$P\left(\sigma^{-1}_{b,M} |\widehat{E}_Q[\phi - \phi_b] - E_P[\phi - \phi_b]| \geq \epsilon \right) \leq \epsilon'. \tag{26}$$

For any $\delta > 0$, for $b > 0$ large enough

$$\mathrm{Var}_P(\phi) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) + \mathrm{Var}_P(\mathbb{E}_P[\phi|X,U])\delta^2_M \leq (1+\delta)^2 \sigma^2_{b,M}. \tag{27}$$

Then, for $M \to \infty$,

$$\begin{aligned}
&\limsup_{M\to\infty} P\left(\left(\mathrm{Var}_P(\phi) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) + \mathrm{Var}_P(\mathbb{E}_P[\phi|X,U])\delta^2_M \right)^{-1/2} \left(\widehat{\mathbb{E}}_P[\phi] - \widehat{\mathbb{E}}_Q[\phi] \right) \leq x \right) \\
&\leq \limsup_{M\to\infty} P\left(\sigma^{-1}_{b,M} \left(\widehat{\mathbb{E}}_P[\phi] - \widehat{\mathbb{E}}_Q[\phi] \right) \leq \max(x(1+\delta), x) \right) \\
&\leq \limsup_{M\to\infty} P\left(\sigma^{-1}_{b,M} \left(\widehat{\mathbb{E}}_P[\phi_b] - \widehat{\mathbb{E}}_Q[\phi_b] \right) \leq \max(x(1+\delta), x) + \epsilon \right) \\
&+ \limsup_{M\to\infty} P\left(\sigma^{-1}_{b,M} \left| \widehat{\mathbb{E}}_P[\phi - \phi_b] - \widehat{\mathbb{E}}_Q[\phi - \phi_b] \right| \geq \epsilon \right) \\
&\leq \Phi(\max(x(1+\delta), x) + \epsilon) + \epsilon'.
\end{aligned}$$

In the last line, we used equation (25) and equation (26). Since $\delta > 0$, $\epsilon' > 0$ and $\epsilon > 0$ can be chosen arbitrary small,

$$\limsup_{M\to\infty} P\left(\left(\mathrm{Var}_P(\phi) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) + \mathrm{Var}_P(\mathbb{E}_P[\phi|X,U])\delta^2_M \right)^{-1/2} \left(\widehat{\mathbb{E}}_P[\phi] - \widehat{\mathbb{E}}_Q[\phi] \right) \leq x \right) \leq \Phi(x).$$

With an analogous argument,

$$\liminf_{M\to\infty} P\left(\left(\mathrm{Var}_P(\phi) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) + \mathrm{Var}_P(\mathbb{E}_P[\phi|X,U])\delta^2_M \right)^{-1/2} \left(\widehat{\mathbb{E}}_P[\phi] - \widehat{\mathbb{E}}_Q[\phi] \right) \leq x \right) \geq \Phi(x).$$

Thus, as $M \to \infty$,

$$\left(\mathrm{Var}_P(\phi) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) + \mathrm{Var}_P(\mathbb{E}_P[\phi|X,U])\delta^2_M \right)^{-1/2} \left(\widehat{\mathbb{E}}_P[\phi] - \widehat{\mathbb{E}}_Q[\phi] \right) \xrightarrow{d} \mathcal{N}(0,1).$$

This completes the proof for one-dimensional ϕ. The result for a vector of functions ϕ follows by applying the Cramér-Wold device.

□

E.2 Estimation for conditional variances

In this part, we detail the estimation of the conditional variances $\mathrm{Var}_P(\phi_P(X))$ and $\mathrm{Var}_P(\phi(X,Y,T) - \mathbb{E}_P[\phi(X,Y,T) \,|\, X])$ in the construction of our distribution shift measures, which is omitted from Appendix B.3.

Recall that P is the underlying distribution of the "source" site, and Q is that of the "target" site. We write the influence function in the most general from $\phi(X,Y,T)$, though in the datasets it is a function of

only (Y, T). We will use cross-fitting (Chernozhukov et al., 2018) with machine learning models for fitting the conditional mean functions.

The variance estimation only needs data from the source site, $\mathcal{D}_1 \overset{\text{i.i.d.}}{\sim} P$. We randomly split $\mathcal{D}_1$ into two folds $\mathcal{D}_1^{(1)} \cup \mathcal{D}_1^{(2)}$. For $k = 1, 2$, we use $\mathcal{D}_1^{(k)}$ to fit the conditional mean function $\widehat{\phi}^{(k)}(\cdot)$ for $\phi_P(\cdot) := \mathbb{E}_P[\phi(X, Y, T) \,|\, X = \cdot]$. Then, writing $\widehat{\varphi}(X_i) = \widehat{\varphi}^{(k)}(X_i)$ for $i \in \mathcal{D}_1 \backslash \mathcal{D}_1^{(k)}$, and $\widehat{\varphi}(X_j) = \widehat{\varphi}^{(k)}(X_j)$ for $j \in \mathcal{D}_2 \backslash \mathcal{D}_2^{(k)}$, we let

$$\widehat{s}^2_{Y|X} = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \big(\phi(X_i, Y_i, T_i) - \widehat{\varphi}(X_i)\big)^2,$$
$$\widehat{s}^2_{X} = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \widehat{\varphi}(X_i) \cdot \big(2\phi(X_i, Y_i, T_i) - \widehat{\varphi}(X_i)\big) - \left(\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \phi(X_i, Y_i, T_i)\right)^2.$$

We define $\sigma(x) = \mathrm{Var}_P(\phi(X, Y, T) \,|\, X = x)$. Our estimators converge in $n^{-1/2}$ rate under standard slow convergence rates of nuisance components.

Theorem E.1. *Suppose* $\|\widehat{\varphi}^{(k)} - \varphi\|_{L_2(P)} = o_P(n^{-1/4})$, *and* $\|\sigma \cdot (\widehat{\varphi}^{(k)} - \varphi)\|_{L_2(P)} = o_P(1)$ *for* $k = 1, 2$. *Then,*

$$\begin{pmatrix} \widehat{s}_{Y|X} \\ \widehat{s}_X \end{pmatrix} = \begin{pmatrix} s_{Y|X} \\ s_X \end{pmatrix} + \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \psi_1(X_i, Y_i, T_i) + o_P(1/\sqrt{n_1})$$

for some fixed function ψ *with mean zero. As a result, each element is consistent and asymptotically* $\sqrt{n}$*-normal, and the asymptotic variances can all be consistently estimated.*

Proof of Theorem E.1. For simplicity, we write $D_i = (X_i, Y_i, T_i)$. By definition,

$$\widehat{s}^2_{Y|X} = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \big\{\phi(D_i) - \widehat{\varphi}(X_i)\big\}^2$$
$$= \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \big\{\phi(D_i) - \varphi(X_i))\big\}^2 + \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} (\widehat{\varphi}(X_i) - \varphi(X_i))^2 - \frac{2}{n_1} \sum_{i \in \mathcal{D}_1} (\widehat{\varphi}(X_i) - \varphi(X_i)) \cdot (\phi(D_i) - \varphi(X_i)).$$

Since $\|\widehat{\varphi}^{(k)} - \varphi\|_{L_2(P)} = o_P(n^{-1/4})$, and by Markov's inequality, we know that for any fixed $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n_1/2} \sum_{i \notin \mathcal{D}_1^{(k)}} (\widehat{\varphi}(X_i) - \varphi(X_i))^2\right| > \epsilon \,\middle|\, \mathcal{D}_1^{(k)}\right]$$
$$\leq \frac{2}{\epsilon^2} \mathbb{E}\Big[(\widehat{\varphi}(X_i) - \varphi(X_i))^2 \,\Big|\, \mathcal{D}_1^{(k)}\Big] = 2\|\widehat{\varphi}^{(k)} - \varphi\|^2_{L_2(P)}/\epsilon^2 = o_P(n^{-1/2}).$$

This implies

$$\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} (\widehat{\varphi}(X_i) - \varphi(X_i))^2 = o_P(1/\sqrt{n}).$$

Also, conditional on $\mathcal{D}_1^{(k)}$, note that $(\widehat{\varphi}(X_i) - \varphi(X_i)) \cdot (\phi(D_i) - \varphi(X_i))$ is i.i.d. with mean zero for all $i \notin \mathcal{D}_1^{(k)}$.

Thus, by Markov's inequality, we have

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{1}{n_1/2}\sum_{i\notin\mathcal{D}_1^{(k)}}(\widehat{\varphi}(X_i)-\varphi(X_i))\cdot(\phi(D_i)-\varphi(X_i))\right|>\epsilon\,\middle|\,\mathcal{D}_1^{(k)}\right]\\
&\le\frac{1}{\epsilon^2}\mathbb{E}\left[\left(\frac{1}{n_1/2}\sum_{i\notin\mathcal{D}_1^{(k)}}(\widehat{\varphi}(X_i)-\varphi(X_i))\cdot(\phi(D_i)-\varphi(X_i))\right)^2\,\middle|\,\mathcal{D}_1^{(k)}\right]\\
&=\frac{4}{\epsilon^2 n_1^2}\sum_{i\notin\mathcal{D}_i^{(k)}}\mathbb{E}\Big[(\widehat{\varphi}(X_i)-\varphi(X_i))\cdot(\phi(D_i)-\varphi(X_i))^2\,\Big|\,\mathcal{D}_1^{(k)}\Big]\\
&=\frac{4}{\epsilon^2 n_1}\|\sigma\cdot(\widehat{\varphi}^{(k)}-\varphi)\|_{L_2(P)}^2.
\end{aligned}$$

Given that $\|\sigma\cdot(\widehat{\varphi}^{(k)}-\varphi)\|_{L_2(P)}=o_P(1)$, we know

$$\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}(\widehat{\varphi}(X_i)-\varphi(X_i))\cdot(\phi(D_i)-\varphi(X_i))=o_P(1/\sqrt{n}).$$

This means

$$\widehat{s}_{Y|X}^2-s_{Y|X}^2=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\big\{(\phi(D_i)-\varphi(X_i))^2-s_{Y|X}^2\big\}+o_P(1/\sqrt{n}).$$

Similarly, by definition, and due to the fact that $\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\phi(D_i)-\mathbb{E}_P[\phi]=O_P(1/\sqrt{n})$,

$$\begin{aligned}
\widehat{s}_X^2&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\widehat{\varphi}(X_i)(2\phi(D_i)-\widehat{\varphi}(X_i))-\Big(\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\phi(D_i)\Big)^2\\
&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\varphi(X_i)(2\phi(D_i)-\varphi(X_i))+\frac{2}{n_1}\sum_{i\in\mathcal{D}_1}(\widehat{\varphi}(X_i)-\varphi(X_i))(\phi(D_i)-\varphi(X_i))\\
&\quad-\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}(\widehat{\varphi}(X_i)-\varphi(X_i))^2-\Big(\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\phi(D_i)\Big)^2\\
&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\varphi(X_i)(2\phi(D_i)-\varphi(X_i))-\Big(\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\phi(D_i)\Big)^2+o_P(1/\sqrt{n})\\
&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\varphi(X_i)(2\phi(D_i)-\varphi(X_i))-(\mathbb{E}_P[\phi])^2-2\mathbb{E}_P[\phi]\Big(\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\phi(D_i)-\mathbb{E}_P[\phi]\Big)+o_P(1/\sqrt{n})\\
&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\big\{\varphi(X_i)(2\phi(D_i)-\varphi(X_i))-2\mathbb{E}_P[\phi]\phi(D_i)+(\mathbb{E}_P[\phi])^2\big\}+o_P(1/\sqrt{n}),
\end{aligned}$$

which further implies

$$\widehat{s}_X^2-s_X^2=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\big\{\varphi(X_i)(2\phi(D_i)-\varphi(X_i))-\mathbb{E}_P[\phi^2]-2\mathbb{E}_P[\phi](\phi(D_i)-\mathbb{E}_P[\phi])\big\}+o_P(1/\sqrt{n}).$$

By Delta method, the above two results imply

$$\begin{aligned}
\widehat{s}_{Y|X}-s_{Y|X}&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\frac{1}{2s_{Y|X}}\big\{(\phi(D_i)-\varphi(X_i))^2-s_{Y|X}^2\big\}+o_P(1/\sqrt{n}),\\
\widehat{s}_X-s_X&=\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}\frac{1}{2s_X}\big\{\varphi(X_i)(2\phi(D_i)-\varphi(X_i))-\mathbb{E}_P[\phi^2]-2\mathbb{E}_P[\phi](\phi(D_i)-\mathbb{E}_P[\phi])\big\}+o_P(1/\sqrt{n}).
\end{aligned}$$

These give us the desired asymptotic expansion of the resulting estimators, with

$$\psi(X_i, Y_i, T_i) = \begin{pmatrix} \frac{1}{2s_{Y|X}}\left\{(\phi(D_i) - \varphi(X_i))^2 - s^2_{Y|X}\right\} \\ \frac{1}{2s_X}\left\{\varphi(X_i)(2\phi(D_i) - \varphi(X_i)) - \mathbb{E}_P[\phi^2] - 2\mathbb{E}_P[\phi](\phi(D_i) - \mathbb{E}_P[\phi])\right\} \end{pmatrix}.$$

We thus conclude the proof of Theorem E.1. □