A recursive clustering scheme for identifying transportable subgroups between multiple RCT populations

by

Marin Jaić

To obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on July 9th, 2025 at 09:00 AM.

Student number:	6075185
Thesis advisor:	Dr. ir. Jesse Krijthe
Daily supervisors:	Rickard Karlsson
	Dr. ir. Jim Smit
External committee member:	Dr. ir. Sicco Verwer
Project duration:	November, 2024 - July, 2025

An electronic version of this thesis is available at `http://repository.tudelft.nl/`

TU Delft

A recursive clustering scheme for identifying transportable subgroups between multiple RCT populations

Marin Jaić

July 2025

Abstract

Combining data from *Randomized Controlled Trials* (RCTs) is a widely used method to estimate causal treatment effects. In order to combine data, the property of transportability, under which different covariate vectors exhibit similar treatment benefit, must hold between the RCTs. However, differences in study design, execution, and the underlying effect modifier distributions can violate transportability which could in turn lead to estimating incorrect causal treatment effect estimates. This thesis addresses the challenge of validating transportability between multiple RCTs and identifying subsets of RCTs between which transportability holds. Our contributions include studying a linear regression-based framework for testing transportability between multiple RCTs and a clustering-based approach for identifying transportable RCT subgroups. Through simulations and analysis of real-world RCTs concerning corticosteroid treatment for Community-acquired pneumonia (CAP), we evaluate the power, robustness, and limitations of our proposed framework.

1 Introduction

Causal inference focuses on estimating how changes in a single variable, the treatment, affect another variable, the outcome. A key task in this field is estimating the *Conditional Average Treatment Effect* (CATE), a function which predicts the expected treatment benefit given a set of covariates.

The CATE is instrumental for administering individualized treatment. Hence, it is of practical use in many spheres, from marketing to medicine [1–3]. In medicine, specifically, strides have been made to personalize treatment based on the patient's features [4]. For instance, Smit et al. [5] have recently studied the benefit of corticosteroid treatment for *Community-acquired pneumonia* (CAP).

Randomized Controlled Trials (RCTs) are considered the gold standard for assessing treatment effects and developing treatment guidelines [6–9]. Treatment assignments in RCTs are random, which allows for unbiased estimation of causal effects from the observed data. RCTs are typically powered for estimating main effects, but are often lacking in power for inferring conditional causal effects, such as the CATE [10].

To address this limitation, researchers often combine data from multiple sources [3, 11, 12]. For instance, in their analysis, Smit et al. [5] pooled different RCT data. Such meta-analyses can grant increased statistical power for estimating the CATE by increasing the total sample count. However, combining RCT data ought to be done carefully.

When combining data from multiple RCTs, estimating conditional treatment effects requires observing all treatment-modulating covariates, called effect modifiers. In practice, unobserved effect modifiers could have differing distributions across RCTs. Pooling data in such circumstances can result in biased estimates. Additional challenges arise from study design variations like treatment protocols or control conditions [13, 14].

When the underlying distributions of effect modifiers and the RCT procedures are compatible, data from multiple trials can be freely pooled. We then say that the RCTs populations are *transportable*. Thus, to correctly estimate causal effects, validating transportability between populations is an important step before combining data.

There have been several proposed methods for validating transportability. Recently Hussain [15] introduced a pairwise test that can be used to validate transportability between an observational study and an RCT, which is easily adaptable to testing two RCTs. For validating transportability

between multiple RCTs, there are options such as Racine's significance of categorical variables test [16] and Luedtke's omnibus test [17].

However, even if transportability does not hold between an entire group of tested RCTs, it still might hold between certain subsets. None of the listed methods provide the means for identifying potentially transportable subgroups of RCTs. To confirm if a subset of RCTs is transportable, we would have to resort to additional testing which would increase the chance of observing a false result.

To address this limitation, we propose a simple parametric model which allows for transportability validation, and enables identification of transportable subgroups of RCTs via clustering.

The questions we then pose are; (i) How effectively does our testing framework detect transportability violations? (ii) Can clustering reliably recover transportable RCT subsets when full-group transportability is violated?

The contributions of this work are twofold. (i) A simple parametric framework for modeling response surfaces and validating transportability between multiple RCTs. Using simulations, we evaluate the power of this approach and showcase its shortcomings. (ii) Using the proposed model we develop a clustering scheme for identifying transportable subgroups of RCTs when transportability is found not to hold within the whole group.

Additionally, we employ our method to analyze real world data, using a subset of the CAP RCTs previously analyzed by Smit et al. [5].

In the following sections, we first introduce the relevant terminology and definitions in Section 2. Section 3 describes all the relevant methodology used in our analysis. The conducted experiments are described in Section 4. Notable findings, limitations and possible future work are discussed in Section 5. Finally the conclusions are outlined in Section 6.

2 Background

Given m RCT datasets. Let $X \in \mathcal{X}$ denote a vector of covariates, i.e. features of a subject, with n dimensions. When administered a treatment $A \in \{0, 1\}$ the subject exhibits an outcome $Y \in \mathcal{Y}$. Depending on the treatment, we observe one of the two possible outcomes Y_0 or Y_1. These are referred to as *counterfactual potential outcomes*. Counterfactual potential outcomes can be either continuous or binary in nature. Additionally, assume that observing a higher value corresponds to a more positive outcome. Finally, let S in the range of $\{1, ..., m\}$ indicate which of the RCTs a data point belongs. We refer to the group of m RCTs as a *family*. Data points comprise a covariate vector X, administered treatment A, observed outcome Y and S denoting which RCT the data point is a member of.

To be able to estimate causal relations the following assumptions must hold for all of the RCT datasets:

1. *Ignorability*: $Y_a \perp\!\!\!\perp A|X, S, \forall a \in \{0, 1\}$
2. *Consistency*: $A = a \Rightarrow Y_a = Y, \forall a \in \{0, 1\}$
3. *Positivity*: $0 < P(A = 1|X = x, S = s) < 1, \forall x \in \mathcal{X}, \forall s \in \{1, ..., m\}$

Ignorability ensures that for a fixed X the counterfactual outcomes are independent of the assigned treatment A. This property follows from the randomized treatment assignment. The consistency assumption is enforced by strictly defining what is entailed by receiving treatment A. It ensures that the observed counterfactual matches the potential counterfactual outcome for the received treatment. Positivity ensures that every subject has a non-zero chance of receiving treatment.

The listed assumptions together lead to the property of *identifiability*; allowing for causal effects to be identified from observed data.

If identifiability holds, the CATE can be estimated. It is defined as $\tau(X) = E[Y_1 - Y_0|X]$. CATE is a function which predicts the expected benefit of administering treatment conditioned on X.

Let us now present three possible transportability conditions, increasing in strength:

1. *Mean exchangeability of contrast*: $E[Y_1 - Y_0|X, S] = E[Y_1 - Y_0|X]$
2. *Mean exchangeability of counterfactual outcome*: $E[Y_a|X, S] = E[Y_a|X]$
3. *Conditional exchangeability of RCTs*: $Y_a \perp\!\!\!\perp S|X$

Mean exchangeability of contrast signifies that two populations adhere to the same CATE function. It is implied by mean exchangeability of counterfactual outcome which states that the expected counterfactual outcome for a fixed X is the same across all of the sources. Finally, conditional exchangeability of RCTs is the strongest transportability condition, implying both mean exchangeability of counterfactual outcome and exchangeability of contrast.

For combining RCTs based on their perceived benefit when administering the treatment, mean exchangeability of contrast is a sufficient condition.

We would thus, like to validate it directly, since it could hold even when the stronger conditions (mean exchangeability of counterfactual outcome and conditional exchangeability of RCTs) do not. However validating mean exchangeability of contrast directly is not trivial as we generally do not have access to the ground truth values for the CATE. Because of that limitation, our scheme is able to validate mean exchangeability of contrast only indirectly, by validating mean exchangeability of counterfactual outcome.

As our scheme can only validate mean exchangeability of counterfactual outcome, from now on when referring to two RCTs as transportable we imply that mean exchangeability of counterfactual outcome holds between them.

3 Method

3.1 Model

The question of whether transportability holds between multiple sources can be rephrased in a different manner. Does conditioning the model on the trial membership variable S yield statistically significant predictive value? Formally put this equates to testing $E[Y_a|X, S] = E[Y_a|X]$.

To assess the influence of trial membership on outcome prediction, we estimate two models: one which conditions on membership and one which does not. Both models are specified as generalized linear models (GLMs) for estimating the outcome. Assume that the covariate vector X includes an intercept term.

1. **Stratified Model**

$$\eta = \sum_{i=1}^{m} \mathbb{I}\{S = i\} \left(\beta_i^\top X + A \cdot \gamma_i^\top X\right) \tag{1}$$

2. **Unstratified Model**

$$\eta = \beta^\top X + A \cdot \gamma^\top X \tag{2}$$

In a GLM the relationship between the outcome Y and the transformed outcome η is established with the *link function* g, such that $E[Y|X] = \mu = g^{-1}(\eta)$.

For continuous outcomes, we assume the outcome model follows a Gaussian distribution, $Y \sim \mathcal{N}(\mu, \sigma^2)$, with an identity link function $f(\mu) = \mu$. The corresponding response surface is then given by $\mathbb{E}[Y \mid X] = \eta$. This corresponds to a linear regression model estimated via ordinary least squares (OLS).

For binary outcomes, we assume $Y \sim \mathcal{B}(\mu)$. The link function in this case is the logit function, $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$. The corresponding response surface is given by $\mathbb{E}[Y \mid X] = \frac{1}{1+e^{-\eta}}$, resulting in a logistic regression model. Model parameters are estimated by maximizing the Bernoulli log-likelihood.

Regardless of outcome distribution, we refer to the first model as the *stratified* model and to the second as the *unstratified* model. The stratified model incorporates information about trial membership variable S when predicting the outcome. The unstratified model is trained on all data by ignoring the categorical variable S. Furthermore, we will refer to the β coefficients as the *main effect coefficients*, and the γ coefficients as the *treatment-covariate interaction coefficients*.

Note that a GLM assumes a linear relationship between the predictors X and the transformed outcome η. If said assumption does not hold, the ability of this framework to correctly model the outcome, and by extension, validate transportability is diminished.

3.2 Transportability validation methods

3.2.1 Chow's test

In the simplest case of having a family of two RCTs, one should turn to a pairwise test. One such test, particularly well-suited for linear regression models, is Chow's Test [18].

Chow's Test assesses whether two independent linear regression models (e.g., trained on two RCT datasets) share the same true regression coefficients. It is commonly used to detect structural breaks i.e., changes in model parameters across subgroups. For our purposes, this is equivalent to validating mean exchangeability of counterfactual outcome.

Suppose we have two datasets of sizes n_1 and n_2. Each dataset is used to fit a linear regression model equivalent to an unstratified model from Equation 2. Presume that X contains an intercept term:

- Dataset 1: $Y_1 = \beta_1^\top X_1 + A_1 \cdot \gamma_1^\top X_1$
- Dataset 2: $Y_2 = \beta_2^\top X_2 + A_2 \cdot \gamma_2^\top X_2$

We then test whether there are any significant differences between the parameters of two models. Under identifiability, this equates to testing:

$$H_0 : E[Y_a|X, S] = E[Y_a|X]$$
$$H_1 : E[Y_a|X, S] \neq E[Y_a|X]$$

Rejecting H_0 suggests that pooling the two RCTs may not be valid due to structural differences in

the outcome prediction.

Finally, we need to calculate the test statistic. To do so, let us define a pooled dataset combining both RCTs, fitted with a single model, also equivalent to an unstratified model:

$$Y = \beta^\top X + A \cdot \gamma^\top X, \quad n = n_1 + n_2$$

Let RSS_p, RSS_1 and RSS_2 denote the Residual Sum of Squares of the pooled model, the model fit on the first RCT and the model fit on the second RCT, respectively. Let k denote the number of estimated parameters. The Chow test statistic is:

$$F = \frac{\left(\mathrm{RSS}_p - (\mathrm{RSS}_1 + \mathrm{RSS}_2)\right)/k}{(\mathrm{RSS}_1 + \mathrm{RSS}_2)/(n_1 + n_2 - 2k)}$$

Under H_0, the test statistics has an F-distribution with k and $n_1 + n_2 - 2k$ degrees of freedom.

The main drawback of Chow's test is that it can only test two datasets at a time, meaning that it requires multiple tests to check a family of more than two datasets. This is best avoided, as repeated pairwise testing increases the probability of observing a false test result. To correct for having to test multiple hypotheses, the Bonferroni correction can be used. This, however, comes at an expense of power.

Note that Chow's test does not leverage the stratified model, defined in Section 3.1. It is used as a reference point when evaluating the performance of ANOVA in the upcoming experiments.

3.2.2 Likelihood-ratio test

Unlike Chow's test, the Likelihood-ratio test (LRT) [19] can validate transportability across multiple RCT datasets by comparing a stratified model against an unstratified model. It assumes that one model is a special case of another, i.e. the reduced model is nested within the full model.

We employ this test when using a maximum likelihood model, such as logistic regression, where the outcome Y is binary. In this setting, the stratified model serves as the full model, and the unstratified model as the reduced model.

The LRT evaluates whether including group-specific effects improves the fit of the model. This again corresponds to testing mean exchangeability of outcome. For our purposes the hypotheses are defined as:

$$H_0 : E[Y_a|X, S] = E[Y_a|X]$$
$$H_1 : E[Y_a|X, S] \neq E[Y_a|X]$$

Let ℓ_0 and ℓ_1 denote the log-likelihoods of the reduced and full model, respectively. Furthermore, let k_0 and k_1 denote the number of parameters in the reduced model and the full model, respectively. The LRT statistic is defined as:

$$\Lambda = -2(\ell_0 - \ell_1)$$

Under the null hypothesis H_0, the test statistic Λ follows a χ^2 distribution with $k_1 - k_0$ degrees of freedom.

3.2.3 ANOVA

In case the outcome Y is continuous, the LRT can be simplified to an ANOVA F-test. The setup is the same as in the LRT. Being a special case of LRT, ANOVA also validates mean exchangeability of counterfactual outcome.

For our purposes the null and alternative hypotheses are formulated as follows:

$$H_0 : E[Y_a|X, S] = E[Y_a|X]$$
$$H_1 : E[Y_a|X, S] \neq E[Y_a|X]$$

To compute the test statistic let RSS_0 and RSS_1 denote the Residual Sum of Squares for the restricted and full model, respectively. The total number of observations is denoted as n. Finally, let k_0 and k_1 denote the number of parameters in the reduced model and the full model, respectively. The ANOVA F-statistic is calculated as:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(k_1 - k_0)}{\mathrm{RSS}_1/(n - k_1)}$$

Under the null hypothesis H_0, the test statistic follows an F-distribution with $(k_1 - k_0)$ and $(n - k_1)$ degrees of freedom.

Note, for $m = 2$, ANOVA is equivalent to Chow's test.

3.3 Transportable subgroup detection

3.3.1 Hierarchical clustering

Unlike Chow's test, ANOVA and LRT can validate transportability for a group of RCTs by conducting a single test. If the test fails however, these methods lend no information as to which RCT or RCTs caused us to discard the null hypothesis, nor if there are transportable RCT subgroups.

In order to detect the transportable RCT subgroups, additional testing would be necessary. For instance, we could check every pair of datasets in a subgroup. Repeated pairwise testing, especially when the number of datasets is bigger, significantly increases the risk of observing a false positive. A Bonferroni correction should be employed to address this. However, this comes at the cost of

lowered power. It would thus be ideal to minimize or completely eliminate any additional testing in the process of finding transportable subgroups.

We propose clustering the regression coefficients obtained from the stratified model described in Equation 1, in which a separate set of parameters is estimated for each RCT. Each set of coefficients corresponds to a point in the clustering space.

Since the outcome is modeled using both main effect β and treatment-covariate interaction coefficients γ, clustering based on mean exchangeability of counterfactual outcomes requires inclusion of both coefficient sets. In contrast, the CATE depends solely on the treatment-covariate interaction coefficients γ, so clustering RCTs based on mean exchangeability of contrasts requires only these coefficients.

There are three particularly challenging obstacles in our clustering problem, in the context of the CAP RCTs dataset application:

1. We have relatively few data points (6 real-world RCTs)
2. The number of clusters (in the case of our real-world data) is unknown
3. Clusters may contain only a single point

Determining the number of clusters is a common challenge in unsupervised learning. It can be approached using iterative heuristics like the elbow method [20], which searches for an optimal number based on a chosen metric, typically the *Within-Cluster Sum of Squares* (WCSS: $\sum_{j=1}^{k}\sum_{\mathbf{x}\in C_j}(\mathbf{x}-\mu_j)^2$, where $\mu_j = \frac{1}{|C_j|}\sum_{\mathbf{x}\in C_j}\mathbf{x}$), or by using clustering algorithms that infer the number of clusters automatically.

However, most metrics used to evaluate a clustering, WCSS included, measure the distance between the elements of a cluster. This makes singleton clusters a problematic edge case, as their within-cluster distances equate to 0. They can artificially inflate the apparent quality of a clustering. Similarly, automatic clustering algorithms are generally not fit for dealing with singleton clusters as they can mark outlier points as noise, or penalize creation of smaller clusters.

To address the listed problems, we employ a hierarchical bottom-up approach [21]. With no introduced adjustments, its main downsides are its time and space complexity, $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$ respectively. Due to the data scarce nature of our setting, these complexities are a non-issue.

The process is a greedy bottom-up grouping method. At the beginning each data-point is placed in its own singleton cluster. A pairwise distance matrix is computed using a chosen distance metric $d(a, b)$ which quantifies the similarity between individual data points. As our data points are continuous and have the same importance, we opt for Euclidean distance. At each iteration, the algorithm selects the two clusters R and S that are closest to each other according to a specified linkage criterion $l(R, S)$. We opt for the complete-clustering linkage criterion $l(R, S) = max_{r\in R, s\in S} d(r, s)$ as it creates compact and distinct clusters. These two clusters are then merged into a single new cluster. After the merge, the distance matrix is updated to reflect the distances between the new cluster and all remaining clusters. This process is repeated iteratively: at each step, the algorithm identifies and merges the most similar pair of clusters and updates the distances accordingly. The procedure continues until all data points are contained in a single cluster, producing a hierarchical tree or dendrogram.

The end result is a hierarchy of possible clusterings, which can be cut at any level to obtain a clustering with a desired number of clusters or to match a specific cut-off distance. For our purposes, we use a cut-off distance t, which returns the clustering such that the distance (according to the linkage criterion) of no two merged clusters is greater than t.

The cut-off distance is chosen by arranging the distances of all of the merged clusters in a list $[d_0, ..., d_n]$ and then computing a difference list $[d_1 - d_0, ..., d_n - d_{n-1}]$. We set $t = \frac{d_i + d_{i+1}}{2}$ for i such that $i = argmax_j(d_{j+1} - d_j)$. The intuition for this approach is that we cut at the point when the most dissimilar clusters were joined.

3.3.2 Recursive clustering

Due to the greedy nature of hierarchical clustering, once joined, there is no way to break up a cluster. This means there is no way to recover from mistakenly grouping points which ought to be apart. To mitigate this potential issue, we devise a recursive clustering scheme which combines hierarchical clustering and transportability validation.

Given a set of RCT datasets, the algorithm first checks whether the set contains only a single dataset. If so, it returns a singleton cluster corresponding to that dataset. Otherwise, it proceeds by fitting a stratified and an unstratified model on the data as described in Section 3.1. Next, the algorithm conducts a transportability test, as described in Section 3.2.2, to validate whether the trials in the current set are transportable. If the test concludes that the trials are transportable, the algorithm returns the entire set as a single

cluster. If the set is not transportable and contains exactly two datasets, it returns two singleton clusters—one for each dataset. Otherwise, the algorithm performs hierarchical clustering on the treatment-interaction coefficients of the stratified model. It then recursively applies the same clustering procedure to each identified subgroup of trials. This process continues until all resulting subsets are either deemed transportable or reduced to singleton clusters. The algorithm finally returns the collection of identified clusters as the output.

This approach repeatedly validates transportability and hierarchically clusters the subgroups of RCTs among which transportability was found not to hold. This allows the method to recover from falsely grouping RCTs which are not transportable, as the algorithm can break up the clusters at further depths.

It is important to note however, if two RCTs which are in the same true group are mistakenly clustered apart at a lower depth, they can not be joined and correctly grouped at a greater depth. Put simply, this algorithm can recover from impure clusters, but it can not recover from incomplete clusters.

4 Experiments

This work explores the effectiveness of using the proposed models to validate transportability using LRT (simplified to ANOVA, for continuous outcomes), and whether RCTs can correctly be grouped into transportable subgroups using only stratified model coefficients, minimizing the need for additional tests. We present four experiments to address these questions.

First, in Experiment 4.2, we analyze how the size of the family of RCTs and the individual RCT dataset sizes affect our proposed transportability validation method and compare it to a pairwise test. Second, we quantify the impact of a misspecified model on our transportability validation tests in Experiment 4.3.

To analyze the performance of recursive clustering we also conduct two experiments. In Experiment 4.4 we take a look at how much do the RCT dataset generation schemes need to differ to correctly cluster different transportable subgroups. In Experiment 4.5, we explore how the grouping structures of transportable RCTs affects our ability to correctly identify them.

All hypotheses are tested at a significance level of 0.05.

4.1 Experimental setup

4.1.1 Clustering metrics

As we synthesized the data for testing of our methods, we had access to the ground truth when clustering. This made metrics such as pairwise precision, pairwise recall and the Rand index specifically practical, as they are easily interpretable and can be calculated only with knowledge of the ground truth.

To be able to define pairwise precision and pairwise recall, we must first define the pairwise confusion matrix.

Given the ground truth clustering G and the estimated clustering S, each pair of data points in S is evaluated, and counted towards one of the following groups based on the ground truth clustering G:

1. True positive (TP):
   A TP pair is a transportable pair of RCTs correctly grouped together

2. False positive (FP):
   An FP pair is a non-transportable pair of RCTs falsely grouped together

3. True negative (TN):
   A TN pair is a non-transportable pair of RCTs correctly grouped apart

4. False negative (FN):
   An FN pair is a transportable pair of RCTs falsely grouped apart

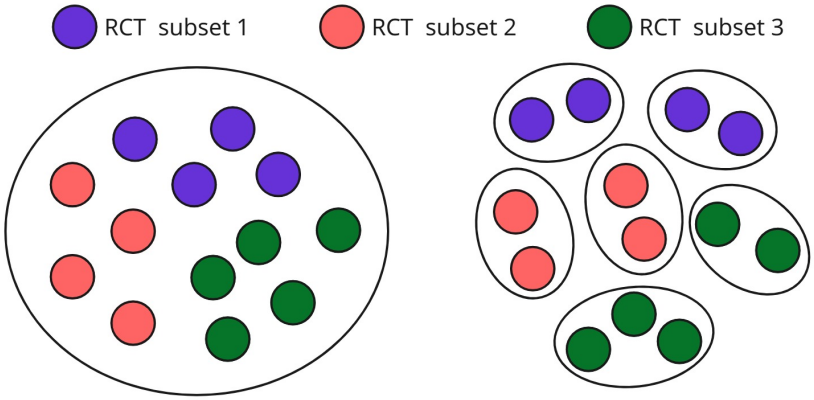
We can now define pairwise precision p and and pairwise recall r as:

$$p = \begin{cases} \frac{TP}{TP+FP}, & TP+FP>0 \\ 0, & TP+FP=0 \end{cases} \tag{3}$$

$$r = \begin{cases} \frac{TP}{TP+FN}, & TP+FN>0 \\ 0, & TP+FN=0 \end{cases} \tag{4}$$

Both pairwise precision and pairwise recall take values in the domain of $\mathbb{R} \in [0,1]$, where 1 is the perfect score. Intuitively, precision measures the purity of the clusters, whereas recall measures the completeness of the clusters. These concepts are visualized in Figure 1.

As only pairs of points are evaluated in the generation of the confusion matrix, singleton clusters can not contribute to the TP count. This becomes an issue in our setting, where singleton clusters may appear in the data. To combat this, we introduce the adjusted confusion matrix calculation which includes one simple change: if a point is in a singleton cluster in the ground truth and it is grouped in a singleton cluster it contributes to the TP count.



(a) Naive maximized recall: r = 1, p << 1 (b) Naive maximized precision: r <<1, p = 1

Figure 1: Example (a) is an extreme example of having a recall of 1 at the expense of precision. Recall is maximized if all members of a true group are placed in the same final cluster, it is not lowered by having multiple such groups in a single cluster. Recall is thus a measure of completeness. Precision is maximized by clustering pairs from the same true group in the same cluster, where there are no other members of a different true group. Precision is not lowered if members of a true group are distributed between different clusters, so long as there are no members of a different group in said clusters. This can be seen in example (b) which has precision of 1 at the expense of recall. Precision measures the purity of the clusters.

Finally, we use the Rand index. Let G be the ground truth clustering and S the predicted clustering. After generating the confusion matrix. The Rand index is then defined as:

$$R = \frac{TP + TN}{TP + FP + FN + TN}$$

The rand index measures the ratio of agreements the between the predicted clustering S and the ground truth G. It takes values on the domain of $[0, 1]$, where 1 is perfect clustering.

4.1.2 Data generation

We first evaluate our approach on synthesized data. Each vector in an RCT dataset is formatted as $[S, X_1, ...X_n, A, Y]$. The data is generated using the following model.

$$Y_s(X, A) = B_s(X) + \tau_s(X) \cdot A + \epsilon$$

We will refer to function B_s as the baseline function. Its effect is always observed regardless of the treatment assignment. τ_s represnts the CATE function of its respective RCT s. The noise factor ϵ is a mean-zero normally distributed variable, $\epsilon \sim \mathcal{N}(0, 1)$. The outcome is synthesized as a continuous variable $Y \in \mathbb{R}$.

For our purposes the baseline function is always modeled as linear:

$$B_s(X) = \beta_s^\top X$$

In Experiments 4.2, 4.4 and 4.5 the CATE function is also linear:

$$\tau_s(X) = \gamma_s^\top X \tag{5}$$

In Experiment 4.3 we use two different CATE functions when synthesizing the data:

$$\tau_1(X) = (1 - p) \sum_i X_i + p \, \frac{25}{3} \cos(3 \sum_i X_i) \tag{6}$$

$$\tau_2(X) = (1 - p) \sum_i X_i + p \, 5e^{-\frac{(\sum_i X_i)^2}{0.98}} \tag{7}$$

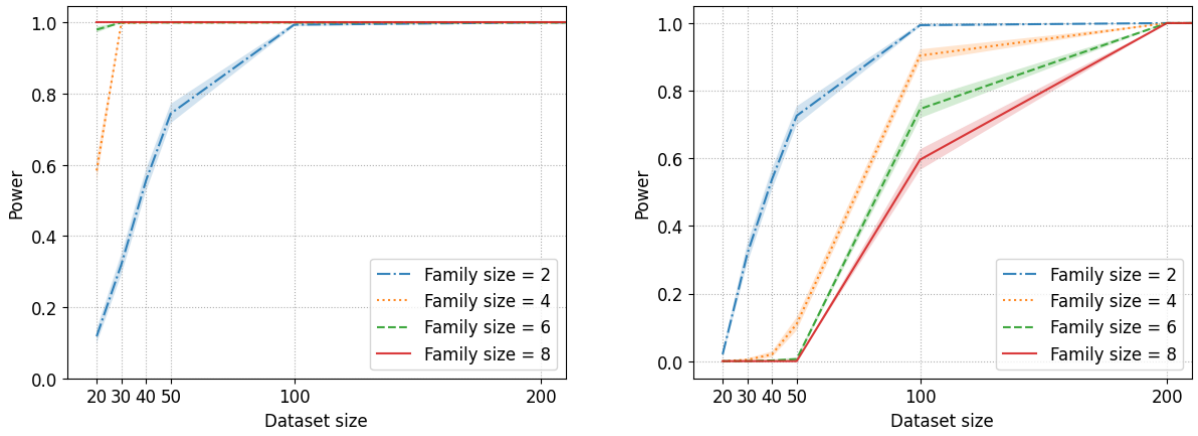
We refer to parameter p as the degree of misspecification. When p is close to 0, the CATE appears to be linear. As p reaches 1, the CATE becomes entirely non-linear.

Unless stated otherwise, the covariate vector X will contain 8 dimensions independently drawn from a mean zero multivariate normal distribution, $X \sim \mathcal{N}(0, I)$.

4.2 Dataset and family size analysis

To determine how the size of the individual datasets and the dataset family we conduct the following experiment affect our proposed transportability testing method we conduct the following experiment. We systematically vary both family and dataset size. For each combination of sizes we track either type I error or power, based on the scenario, averaged over 1000 runs. Since the data generation scheme is continuous, we use ANOVA to validate transportability. We compare it to Chow's test by testing all in the group.

We examine two scenarios:



(a) ANOVA power

(b) Chow's test family-wise power

Figure 2: Figures 2a and 2b depict how the power, averaged over 1000 runs, behaves as the sizes of individual datasets increases for ANOVA and Chow's test family-wise testing respectively.

1. **Transportable scenario**: All datasets are generated from the same distribution, making them mutually transportable. In this scenario, we track the type I error rate. For pairwise testing, all pairs within the family are tested, and a single false positive renders the entire run a false positive. The goal of this scenario is to show that the tests are well calibrated.

2. **Non-transportable scenario**: No two datasets in the family are generated from the same distribution. Here, we measure power, i.e., the probability of correctly rejecting the false null hypothesis. For pairwise testing, a run is considered a false negative if any pairwise test fails to reject the null hypothesis.

To account for the accumulation of type I errors, we use the Bonferroni correction for setting the significance value of Chow's test. We hypothesise that ANOVA will perform better than Chow's test on larger families, but the inverse to be true for smaller families.

The results can be seen in Figure 2. Both ANOVA and Chow's test maintain a type I error rate close to the nominal level of 0.05 as can be seen in Figure 7. ANOVA exhibits good power for all but the smallest dataset and family sizes. Chow's test on the other hand, exhibits a family-wise power of almost 0 for any family size above 2 and dataset size beneath 100. The power quickly rises to nearly 1 as the RCT dataset sizes rise to 200.

4.3 Misspecified model

Our approach assumes a linear relation between the covariates and the transformed outcome. However, this assumption need not be correct and can result in estimating a misspecified outcome model. This can hinder our ability to correctly test transportability and further group the trials.

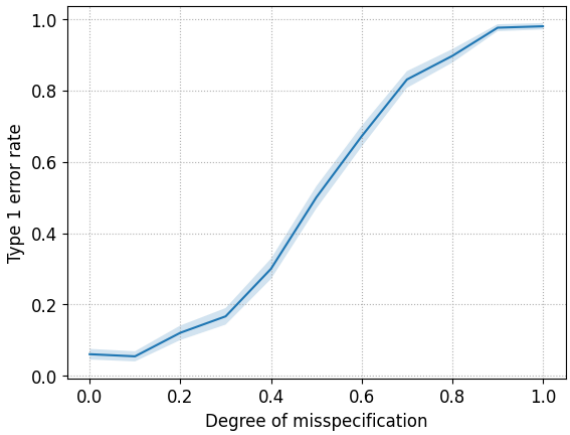
The goal of this experiment is to quantify how much the degree of misspecification of our model affects the reliability of our transportability validation approach.

Consider the following setting with a family of 6 RCTs.

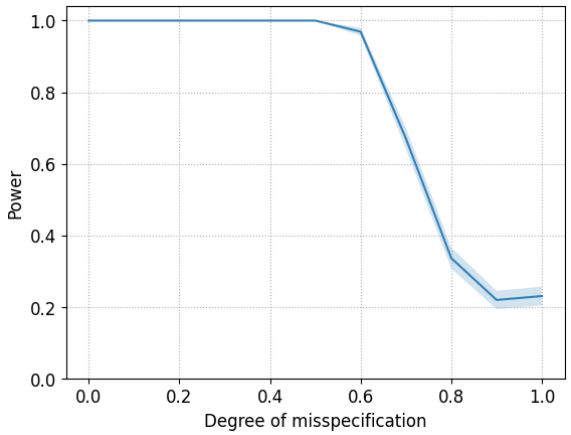
- **Scenario A**: All RCTs are transportable, with the CATE defined in Equation 7. The RCTs are divided into two equally sized groups, where covariates are drawn from multivariate normal distributions with means -0.5 and 0.5, respectively. The degree of misspecification p is gradually increased from 0 to 1. As the degree of misspecification grows, the CATE becomes increasingly non-linear. We track the type I error rate over 1000 runs. We hypothesize that type I error rate will increase with the degree of misspecification, as the linear models will try to approximate different regions of a non-linear function. This should result in diverging estimates of the same true function as more non-linearity is introduced. This concept is visualized in the appendix (Figure 8).

- **Scenario B**: The family is divided into two non-transportable groups of size 3, with CATE functions defined by Equations 6 and 7, respectively. As in the first scenario, we incrementally introduce model misspecification. Covariates in both groups are drawn from a multivariate mean-zero normal distribution. We evaluate the power of ANOVA over 1000 runs, which we hypothesize to decrease substantially as misspecification increases.

Additionally, to confirm that the test is well calibrated in Scenario B, retaining a nominal type



(a) Scenario A: type I error rate



(b) Scenario B: power

Figure 3: Figure 3a depicts the average type I error when validating transportability on a family of 6 RCTs which have the same CATE, but are split in 2 groups of equal sizes which have different covariate distributions. Figure 3b depicts the power when testing whether transportability holds on a family of 6 datasets which are split into 2 equally sizes transportable groups. The covariates are sampled from the same distribution. In both experiments, the CATE is gradually made less linear as the degree of misspecification is increased.

I error of 0.05, we track the type I error when testing a group of 6 RCTs. The RCTs follow same non-linear CATE and sample their covariates from the exact same mean zero normal distribution.

The results are depicted in Figure 3. In Figure 3a we observe that having a misspecified model with different covariate distributions between the RCT populations will increase the type I error rate of ANOVA. The type I error rate reaches almost 1 as the degree of misspecification grows. Furthermore, ANOVA's power starts decreasing significantlly after the degree of misspecification reaches 0.6 as can be seen in Figure 3b. The power drops to a value of about 0.25 when the CATE is completely non-linear. In this case the test retains a nominal type I error rate of 0.05 no matter the degree of specification as visualized in Figure 7.

4.4 Dataset difference analysis

To evaluate the sensitivity of the recursive clustering procedure in detecting non-transportable datasets, we assess how distinct the data-generating mechanisms must be for the method to reliably differentiate between them. While such differences are difficult to quantify in real-world settings, synthetic data allow precise control over and observation of differences in data generation across datasets.

Suppose we have a family of 6 synthetically generated RCTs, which form 3 transportable groups. Suppose $s \in \{1,2\}$ is group 1, $s \in \{3,4\}$ group 2, and $s \in \{5,6\}$ group 3. When generating the data, we increasingly vary the linear CATE function, generated as per Equation 5. The

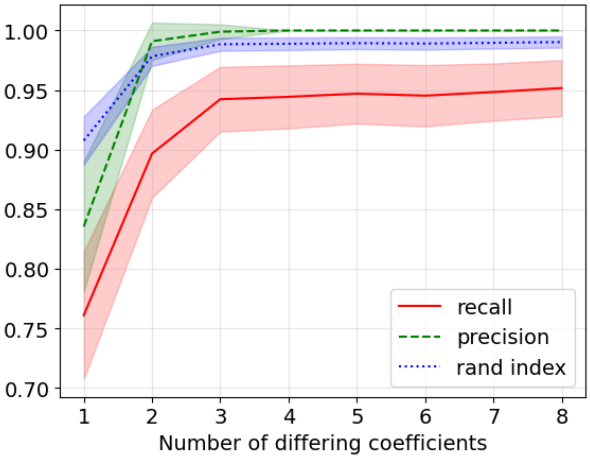


Figure 4: The figure depicts the average recall, precision and rand index with their respective 95% confidence intervals averaged over 1000 runs. The x-axis tracks the number of different parameters in the treatment effect parameter vector which is described in Section 4.1.2.

starting configuration of γ_s for all of the groups is a vector of ones, $\gamma_s \in \{1\}^8$. Iteratively, we change the parameter vectors, one element at a time at the same index, so that in the final iteration $\gamma_{s\leq 2} \in \{1\}^8$, $\gamma_{3\leq s\leq 4} \in \{0\}^8$, $\gamma_{5\leq s\leq 6} \in \{-1\}^8$.

At each iteration, we track how the recall, precision and Rand index of behave as the outcome generation schemes become increasingly different. We hypothesise that the method will perform better as the datasets start to differ more in the later iterations. The clustering is performed on the covariate-treatment interaction coefficients.

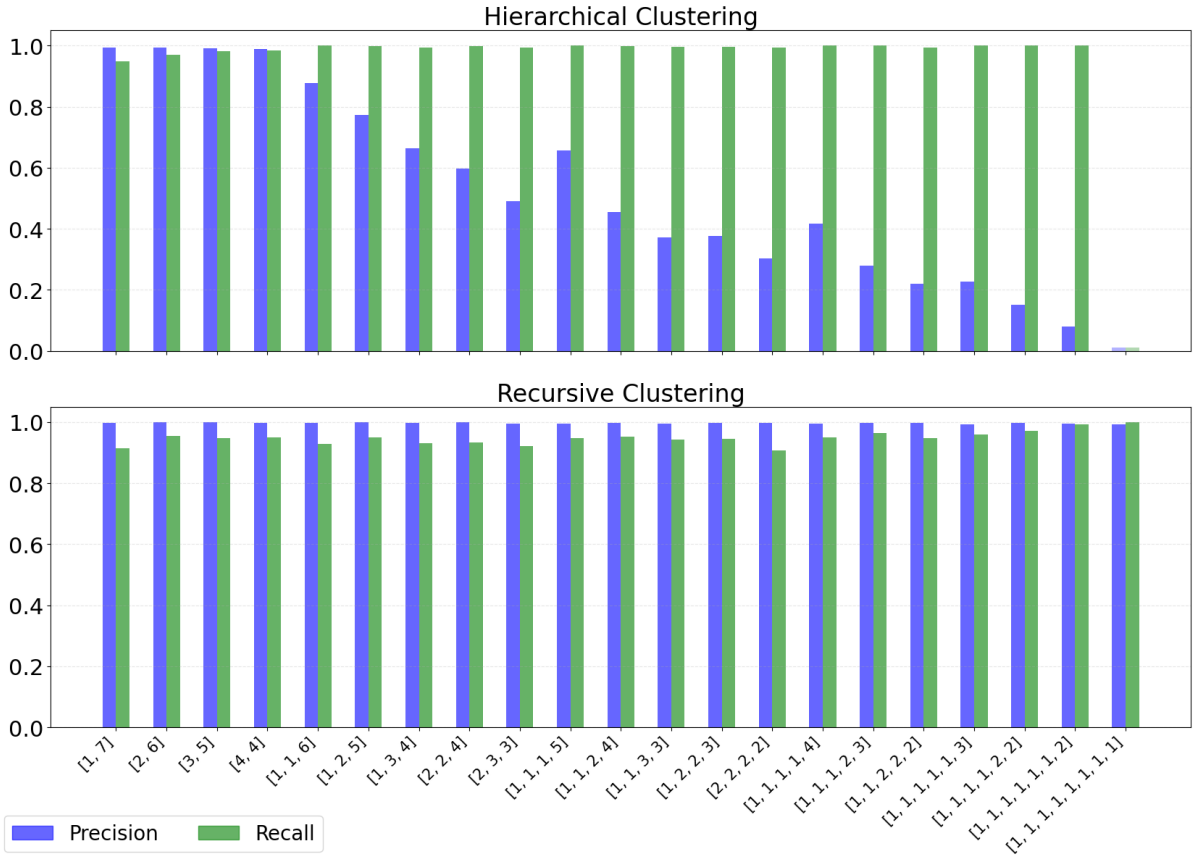


Figure 5: Mean recall and precision for hierarchical clustering and recursive clustering. For clarification, {1, 1, 2} signifies the existence of 3 transportable RCT subgroups, two singleton RCTs and one subgroup with two RCTs

The results can be seen in Figure 4. Precision and Rand index converge to a value of 1 and 0.98 respectively with very narrow confidence intervals as is seen in Figure 4, after just 3 differing coefficients. Recall reaches a stable value of 0.95 after 5 differing coefficients, but has wider confidence intervals.

4.5 Grouping structures

Aside from the differences in the data-generating processes, both the number of true clusters and the underlying grouping structure of transportable RCTs may influence the algorithm's ability to detect violations of transportability and correctly cluster datasets.

By grouping structure, we refer to the possible ways to group a set of m RCTs into transportable subsets. For instance, a family of size 7 partitioned into 3 clusters can yield the following configurations: {1,1,5}, {1,2,4}, {1,3,3}, and {2,2,3}.

To evaluate the impact of grouping structure on the performance of the recursive clustering algorithm, we conduct the following experiment. For a family of 8 RCT datasets, we incrementally vary the number of transportable groups from 2 to 8. For each number of clusters, we enumerate all possible grouping structures and generate synthetic data accordingly. Recursive clustering is done on the treatment-covariate interaction coefficients, repeated 1000 times for each configuration. For comparison, we repeat the same experiment using the hierarchical clustering method from Section 3.3.1.

Results, visualized in Figure 5, show that the recursive clustering overall performs significantly better than just hierarchical bottom-up clustering, averaging a precision of almost 1 and a recall greater than 0.92 over all grouping structures. Hierarchical clustering performs badly precision-wise as the number of true groups grows, but exhibits almost perfect recall in all but a single case. However, when the true number of transportable RCT groups is 2, hierarchical clustering slightly outperforms recursive clustering.

4.6 Real world data

This experiment demonstrates the usage of our method on real data and observing which transportablity trends can be detected on the real world CAP RCTs. The test includes training a

Table 1: Transportability test results across preprocessing strategies. RCTs are labeled {1, 2, 3, 4, 5} with starting sizes: 1 (304), 2 (785), 3 (401), 4 (213), 5 (794). Transportability is tested using ANOVA, the RCTs are grouped using recursive clustering. In the clustering column, {1, 2} would denote RCTs 1 and 2 being placed in the same cluster.

Method	**p-value**	**RCTs Dropped**	**Total Size**	**Remaining Features**	**Clustering**
Method 1	0.08	4, 5	908	20	{1, 2, 3}
Method 2	$8.2 \cdot 10^{-10}$	\	1649	16	{1}, {2, 3}, {4}, {5}
Method 3	$6.75 \cdot 10^{-7}$	\	2497	16	{1}, {2}, {3}, {4}, {5}
Method 4	$6.33 \cdot 10^{-7}$	\	2497	20	{1}, {2}, {3}, {4}, {5}

Method 1: Row Dropping **Method 2:** Column Pruning and Row Dropping **Method 3:** Column Pruning and Within-RCT Imputation **Method 4:** Cross-RCT and Within-RCT Imputation

logistic regression model as described in Section 3.1. Transportability is tested using LRT and transportable subgroups are identified using recursive clustering, if transportability is found not to hold.

We have access to 6 real-world RCTs, containing a total of 2617 samples. Every row contains 20 covariates, some of which can be missing. From the very beginning, one RCT is excluded from the analysis due to the stratified model being unable to converge if it was included during optimization. This occurred because the model could perfectly predict the outcomes in said RCT. Since the likelihood function grows as the coefficients grow, preventing convergence to a finite solution by pushing coefficients towards $\pm\infty$ during optimization.

To handle missing values, we employ four approaches: (i) Row dropping: Remove all rows with any missing values. This is the simplest approach to dealing with missing values. It can, however, cause severe data loss, which could lead to biased estimates. (ii) Column pruning row dropping: Drop columns entirely missing in any RCT, then drop remaining rows with missing values. This approach mitigates the severe data loss of row dropping, however, it can cause us to ignore important predictors due to them not being present in a single RCT. (iii) Column pruning and within-RCT imputation: Drop columns which are always missing in any RCT. Impute remaining missing values using a random forest trained on the same RCT. This method drops columns, but preserves the total number of samples in the data. However, it could induce bias by projecting the distribution of complete rows when filling in the missing values. (iv) Cross-RCT and within-RCT imputation: If a column is always missing in an RCT, impute it using a random forest trained on other RCTs where it's observed; otherwise, impute within-RCT. This method preserves both the total number of predictors and the total number of samples. However, similarly to method (iii) it could induce bias by projecting the distributions of the fully observed data.

The results are shown in Table 1. Due to missing values, the row dropping approach removes two more sources from the analysis as their every row had at least one missing value. The remaining 3 RCTs are found to be transportable with a p-value of 0.08. When using column pruning and row dropping the family is found to be not transportable with a p-value of $8.2 \cdot 10^{-10}$. The RCTs are clustered into 3 singleton clusters and one cluster of size 2, indicating that 2 RCTs form a transportable subgroup. The remaining data processing methods both find transportability not to hold and place all of the RCTs in singleton clusters. Column pruning and within-RCT imputation rejects the null hypothesis with a p-value of $6.75 \cdot 10^{-10}$, whereas cross-RCT and within-RCT imputation does so with a p-value of $6.33 \cdot 10^{-7}$. The column pruning methods remove four columns (creatine, cancer indicator, liver disease indicator and renal disease indicator).

5 Discussion

Our findings suggest that ANOVA generally outperforms Chow's test for transportability validation. However, we did not consider every possible way that a family of RCTs could be partitioned into transportable subgroups (when evaluating the power with a family of 4 RCTs we only look at the case where none are transportable, we could also have 2 groups of 2 transportable RCTs). In cases where most RCTs are transportable and only a few are outliers (e.g., one out of seven), a pairwise testing scheme may be more appropriate.

We also find that rejection of the null hypothesis can stem from model misspecification in

conjunction with a covariate shift between the RCTs, rather than true differences in their CATEs. When covariate distributions differ, the stratified linear model fits hyperplanes around the sampling region of each RCT. If covariates are sampled from sufficiently distant regions, the approximated hyperplanes can differ, potentially leading to false rejections. This issue is illustrated in the appendix (Figure 8). Moreover, when non-transportable RCTs follow different CATE functions but are locally approximated by similar linear hyperplanes, our method may fail to reject transportability. These limitation could be mitigated by transforming covariates before fitting the model when the relationship between the covariates and the outcome is suspected to be non-linear. Variable selection methods like forward selection could be used to find meaningful transformations. If data is not transformed, a non-parametric test such as Racine's test [16] ought to be utilized to validate transportability.

Hierarchical clustering performs well only when there are two transportable groups. In all other cases it exhibits lowered precision while retaining almost perfect recall. This indicates that the method rarely splits RCTs which are in the same true group, but often groups RCTs which are in different true groups. Thus, we conclude that hierarchical clustering generally produces only two clusters, which may not align with the true underlying structure.

Recursive clustering consistently demonstrates high recall and precision and substantially outperforms hierarchical clustering. However, the results show that precision is consistently higher than recall. Interestingly, recursive clustering is outperformed by hierarchical clustering only when there are exactly two transportable RCT subgroups. Both findings can be attributed to the additional transportability validation step that recursive clustering performs upon forming a cluster. Even if a formed cluster contains only RCTs from the same transportable group, a false positive will cause the method to unnecessarily split the cluster. This preserves the precision but reduces recall.

The previously mentioned finding highlights an important limitation of recursive clustering. Although it reduces the amount of additional tests required for identifying transportable subgroups, it does not eliminate them, making it susceptible to error accumulation. Ideally, one would want to cluster all transportable RCTs correctly in a single pass without employing the recursion. This could be achieved by improving the robustness of hierarchical clustering, specifically by designing a better strategy for choosing the cluster cut-off distance, allowing it to discover more than two true subgroups. Alternative clustering strategies could also be explored. Furthermore, incorporating the uncertainty of main effect and treatment-covariate interaction coefficients into the clustering process could also be explored in future work.

In the real-world experiment, dropping rows with missing values did not lead to rejecting transportability. However, this method retained only about one-third of the original data. Furthermore, all other data processing methods reject transportability, having p-values which are several magnitudes of order more extreme. We thus conclude that, in this case, dropping rows with missing values removes a lot of samples which would invalidate transportability if present. Methods (iii) and (iv) differ only in how they handle missing columns, but produce very similar p-values, leading us to hypothesize that the dropped columns do not significantly contribute towards rejecting transportability.

Observing the appendix of Smit et al. [5], we note that certain distributions of covariates visually differ between the RCTs (e.g., diabetes and renal disease). In conjunction with a misspecified model, the differences between the distributions of covariates could explain the strong rejection of our method. To diagnose the covariate shift between the RCTs explicitly, a test such as Maximum Mean Discrepancy (MMD) [22] ought to be utilized.

Our method is currently limited to simple parametric models. While they do have the upside of being highly interpretable, their ability to capture more complex, non-linear relationships is limited. Future work could explore extensions using more flexible models such as neural networks, or causal forests. In such cases, we hypothesize that clustering could be done on model representations, such as a vector of CATE estimates.

Another limitation of our approach is that we can validate mean exchangeability of contrast only ndirectly, by testing mean exchangeability of counterfactual outcomes which implies it. Although it may seem intuitive to construct models that estimate the CATE directly, this invalidates likelihood or RSS based statistical tests, since the model is no longer fitted to the observed outcome Y. As a result, likelihood ratio and ANOVA tests become ill-defined. To directly test mean exchangeability of contrast, a test such as the method proposed by Hussain et al. [15] could be utilized. However, as Hussain's method is a pairwise test, this would require repeated testing.

Finally, our experiments relied on synthetic data generation, including linear CATEs for Experiments 4.2, 4.4, 4.5 and non-linear functions for Experiment 4.3. While these choices enabled con-

trolled evaluation of model misspecification, they may not fully capture the complexity of real-world treatment effect heterogeneity. Similarly, our covariate distributions were limited to multivariate Gaussians without structured missingness patterns. Future work could explore more diverse data generation schemes or sparsity-inducing distributions, bridging the gap between simulations and reality.

6 Conclusion

This thesis addresses the challenge of validating transportability across multiple RCT datasets. Furthermore, it explores an approach for identifying transportable subgroups when transportability does not hold across an entire family of tested RCTs. To that end, we proposed a linear modeling framework for testing transportability and introduced a recursive clustering algorithm that effectively identifies transportable clusters within non-transportable families of RCTs.

Through extensive simulations, we demonstrated the advantages and weaknesses of our transportabiltiy validation framework, and the recursive clustering algorithm. In particular, we showed that even for smaller families our transportability testing approach which utilizes ANOVA outperforms pairwise testing power-wise while maintaining a nominal type I error rate. Furthermore, we found that our transportability validation scheme is sensitive to model misspecification, especially so when there are differences in covariate distributions between the RCTs. The effect of model misspecification could be alleviated by domain expertise or variable selection techniques in order to recognize when applying transformations to the data is required. In case transportability is found not to hold between a family of RCT datasets, we show that our recursive clustering can reliably detect underlying transportable structures, especially when differences between the CATE functions are sufficiently pronounced. Finally we employ our method on a real world family of RCTs and draw the conclusion that transportablity does not hold across the entire family. However, the results are highly dependent on the data processing technique. Furthermore, the rejections could also be caused by a misspecification of our model.

In sum, this thesis contributes a simple hyperparameter-free framework for validating transportability and detecting transportable subgroups between RCT datasets, with promising results. With further refinement, the proposed approach could be used to enhance the reliability of meta-analyses and advance personalized healthcare.

References

[1] H. R. Varian. "Causal inference in economics and marketing". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7310–7315. DOI: 10.1073/pnas.1510479113. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1510479113. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1510479113.

[2] L. Norman. "Rethinking causal explanation in interpretive international studies". In: *European Journal of International Relations* 27.3 (2021), pp. 936–959.

[3] I. J. Dahabreh, R. Hayward, and D. M. Kent. "Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence". In: *International journal of epidemiology* 45.6 (2016), pp. 2184–2193.

[4] L. H. Goetz and N. J. Schork. "Personalized medicine: motivation, challenges, and progress". In: *Fertility and sterility* 109.6 (2018), pp. 952–963.

[5] J. M. Smit et al. "Predicting benefit from adjuvant therapy with corticosteroids in community-acquired pneumonia: a data-driven analysis of randomised trials". In: *The Lancet Respiratory Medicine* (2025).

[6] I. Chalmers, D. G. Altman, et al. *Systematic reviews*. BMJ Publishing London, 1995.

[7] H. C. Sox. "Defining comparative effectiveness research: the importance of getting it right". In: *Medical care* 48.6 (2010), S7–S8.

[8] B. Djulbegovic and G. H. Guyatt. "Progress in evidence-based medicine: a quarter century on". In: *The lancet* 390.10092 (2017), pp. 415–423.

[9] D. L. Sackett et al. *Evidence based medicine: what it is and what it isn't*. 1996.

[10] P. L. Bedard et al. "Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings". In: *Journal of clinical oncology* 25.23 (2007), pp. 3482–3487.

[11] K. E. Rudolph and M. J. Laan. "Robust estimation of encouragement design intervention effects transported across sites". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.5 (2017), pp. 1509–1525.

[12] D. Westreich et al. "Transportability of trial results using inverse odds of sampling weights". In: *American journal of epidemiology* 186.8 (2017), pp. 1010–1014.

[13] U. Klinge et al. “Bias-variation dilemma challenges clinical trials: inherent limitations of randomized controlled trials and meta-analyses comparing hernia therapies”. In: *International Journal of Clinical Medicine* 5.13 (2014), pp. 778–789.

[14] V. Smail-Faugeron et al. “Meta-analyses frequently include old trials that are associated with a larger intervention effect: a meta-epidemiological study”. In: *Journal of Clinical Epidemiology* 145 (2022), pp. 144–153.

[15] Z. Hussain et al. “Falsification of internal and external validity in observational studies via conditional moment restrictions”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 5869–5898.

[16] J. S. Racine, J. Hart, and Q. Li. “Testing the significance of categorical predictor variables in nonparametric regression models”. In: *Econometric Reviews* 25.4 (2006), pp. 523–544.

[17] A. Luedtke, M. Carone, and M. J. van der Laan. “An omnibus non-parametric test of equality in distribution for unknown functions”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81.1 (2019), pp. 75–99.

[18] G. C. Chow. “Tests of equality between sets of coefficients in two linear regressions”. In: *Econometrica: Journal of the Econometric Society* (1960), pp. 591–605.

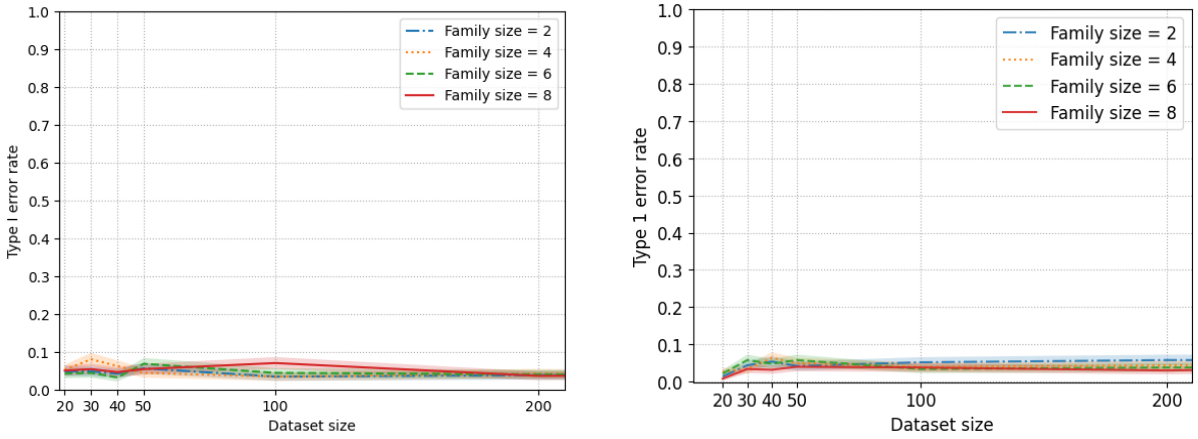
[19] J. Neyman and E. S. Pearson. “IX. On the problem of the most efficient tests of statistical hypotheses”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[20] R. L. Thorndike. “Who belongs in the family?” In: *Psychometrika* 18.4 (1953), pp. 267–276.

[21] J. H. Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[22] A. Gretton et al. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

7 Appendix A: Supplementary figures



(a) ANOVA type 1 error rate

(b) Chow's test family-wise type I error rate

Figure 6: Average type I error rates for ANOVA and Chow's test

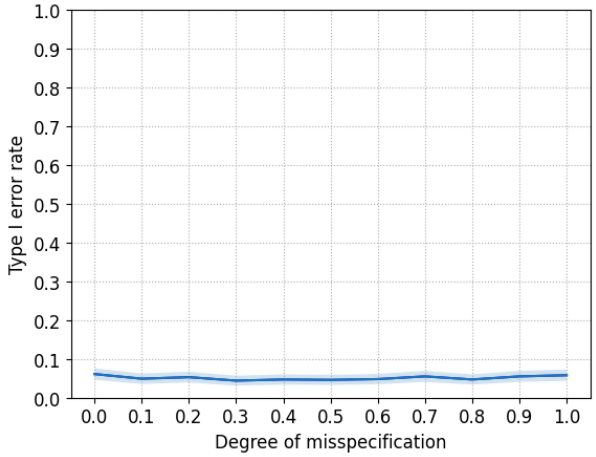


Figure 7: Average type I error rate when testing a family of 6 RCT datasets with the same covariate distribution as the degree of misspecification is increased.

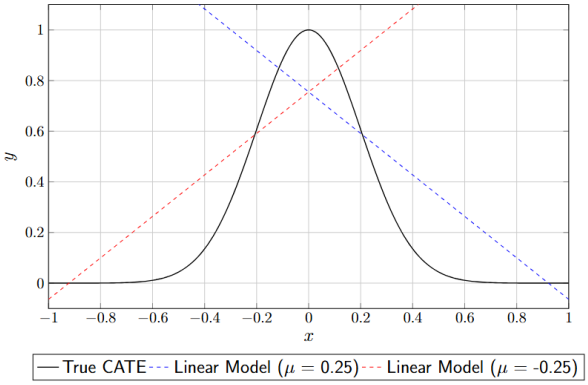


Figure 8: The figure demonstrates how a covariate shift can detrimentally impact the performance of a misspecified model model on a dummy example. The red and blue lines are the fit on two different populations which follows the same CATE and have a size one covariate vector X. However, the red population's covariate distribution is $X_r \sim \mathcal{N}(-0.25, 1)$, whereas blue's covariate distribution is $X_b \sim \mathcal{N}(0.25, 1)$. The resulting models then estimate vastly different slopes, even though their true CATE is the same.

8 Appendix B: Missing Data Handling Strategies

This appendix provides a detailed description of the four preprocessing methods used to handle missing values across multiple RCT datasets.

1. **Row dropping**: Removes all rows that contain any missing values.
2. **Column Pruning and Row Dropping**: Columns that are completely missing in at least one RCT are dropped from the dataset. Subsequently, any remaining rows with missing values are also removed.
3. **Column Pruning with Within-RCT Imputation**: Columns that are entirely missing within any RCT are dropped. The remaining missing values are imputed using a supervised machine learning model (random forest regressor or classifier), trained separately within each RCT using only the observations which are not missing the value being imputed. When training the random forest regressor or classifier, any missing values (other than the value being predicted) are filled with the mean value of the column for that source.
4. **Cross-RCT and Within-RCT Imputation**: For columns that are entirely missing within a given RCT, imputation is performed using a model trained on other RCTs where the column is observed. All other missing values in the training set are filled with the mean value of the value in that sources. For all other missing values, within-RCT imputation is applied as in method (iii).