This appendix present more details regarding the technical results in the main paper, the data and the experimental processing pipeline. We also present a detailed version of the theorems on consistency results in the main paper. The proofs for those theorems (Thm. 2 and Thm. 3 in the main paper) can be found in the reference (1). We also discuss some technical aspects describing how to conduct hypothesis tests in the small sample size case, how to find a minimal d-separating set to control biases efficiently, and how to solve the optimization problem. We give the full proofs of Thm. 1 and Thm. 4 in the main paper here in the appendix. Finally we show the performance of the hypothesis tests and estimation with some simulations.

### S.1. Details regarding datasets from ADNI and W-ADRC

The ADNI dataset is obtained from the Alzheimer’s Disease Neuroimaging Initiative database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD).

The W-ADRC dataset is available to the research community. All requests are handled centrally by the W-ADRC. A researcher can access the data at [http://www.adrc.wisc.edu/apply-resources](http://www.adrc.wisc.edu/apply-resources). This includes the user agreement to ensure that investigators are in compliance with their local institutional IRB policy.

In our datasets W-ADRC and ADNI, there are 7 proteins available which we used in our analyses, including \( A\beta_{1-38}, A\beta_{1-40}, A\beta_{1-42}, t-\text{tau}, p-\text{tau}_{181} \), neurofilament light (NFL) and neurogranin (NG) (2). We describe the assays used to get the measurements of those proteins. The reader can find more details regarding the assays in the literature (3). In W-ADRC, \( A\beta_{1-38}, A\beta_{1-40}, A\beta_{1-42} \) were quantified by electrochemiluminescence (ECL) using an \( A\beta \) triplex assay/Meso Scale Discovery (MSD). For the ADNI data acquired at University of Pennsylvania (UPenn), the procedure used was 2D-UPLC-tandem mass spectrometry method. In W-ADRC, \( A\beta_{1-42}, t-\text{tau}, p-\text{tau}_{181} \) were quantified with INNOTEST, whereas in the ADNI data, UPenn uses AlzBio3 xMAP to measure \( A\beta_{1-42}, t-\text{tau}, p-\text{tau}_{181} \). Both W-ADRC and ADNI data use a sandwich ELISA method to measure NFL and electrochemiluminescence technology/MSD to measure neurogranin.

To obtain the ADNI datasets, we downloaded “UPENN CSF Biomarker Master” file and select each person’s baseline data at batch UPENNBIOMK to get \( A\beta_{1-42}, t-\text{tau}, p-\text{tau}_{181} \). We downloaded the “UPENN 2D-UPLC tandem mass spectrometry measurement” file to get MSD \( A\beta_{1-38}, A\beta_{1-40}, A\beta_{1-42} \) values. We downloaded the “Blennow Lab CSF Ng” and “Blennow Lab CSF NFL” files to get neurogranin and NFL. We downloaded the “Key ADNI tables merged into one table” file to get the hippocampus volume, diagnosis status, age and sex. Then, we merge all tables by patient ID as the primary key and we only include samples with all measurements available, which gives us 311 samples. Similarly, we get a dataset from W-ADRC only including samples with all measurements available, which gives us 154 samples. Since sample sizes of ADNI and W-ADRC datasets are very different for participants with age larger than 85 and smaller than 55, we decided to only consider subjects with ages between 55 and 85, which leads to 284 samples for ADNI and 125 samples for W-ADRC.

### S.2. Details of theorems from Section "Tests for Correcting Distributional Shift"

In the main paper, we avoided listing all technical assumptions and details for brevity, such as constant terms for consistency results in Theorem 2 and 3. We provide a detailed version of this information here in this appendix. Following notations used in the main paper, we denote samples from the two datasets separately as \( X_S \) and \( X_T \) with sample sizes \( n_S \) and \( n_T \) from population distributions \( P_S \) and \( P_T \). We call transformation classes \( h^S(\cdot) \) for \( X_S \) and \( g^T(\cdot) \) for \( X_T \), depending on parameters \( \lambda \) or \( \theta \). We define the search region for \( \lambda \) as \( \Omega_\lambda \) which belongs to \( \mathbb{R}^p \); similarly, the search region for \( \theta \) is defined as \( \Omega_\theta \in \mathbb{R}^{p_\theta} \). In our results, we assume that \( \Omega_\lambda, \Omega_\theta \) live in Euclidean space. The results, however, may be generalized to other spaces with similar techniques.

#### Assumptions S.1

We require three assumptions for the consistency results (Thm. 2 and Thm. 3 in the main paper).

1. The search regions \( \Omega_\lambda, \Omega_\theta \) of \( \lambda, \theta \) are bounded.

Hao Henry Zhou, Vikas Singh, Sterling C. Johnson, Grace Wahba, and the Alzheimer’s Disease Neuroimaging Initiative
The kernel function \( K \) of RKHS used for MMD is non-negative, characteristic and bounded by a constant \( k \).

(3a) \( \|K(h^x(x), \cdot) - K(h^y(x), \cdot)\|_H \leq L_k \|\lambda_1 - \lambda_2\|_2^6 \) with constants \( L_k, r_h \), for any \( x \) in the support of \( P_S \) and \( \forall \lambda_1, \lambda_2 \in \Omega_\lambda \).

(3b) \( \|K(g^x(x), \cdot) - K(g^y(x), \cdot)\|_H \leq L_g \|\theta_1 - \theta_2\|_s^s \) with constants \( L_g, r_g \), for any \( x \) in the support of \( P_T \) and \( \forall \theta_1, \theta_2 \in \Omega_g \).

The Assumption S.1 (2) is used for most consistency analyses based on MMD (4). The Assumption S.1 (3a) and (3b) are satisfied for a big class of transformations with differentiable radial basis kernel. We have following sufficient conditions to satisfy Assumptions S.1 (3a) and (3b).

Lemma S.2. If the following two conditions are satisfied, then Assumptions S.1 (3a) and (3b) hold.

1. The kernel function \( K(\|x - y\|_2) \) is a radial basis kernel where \( \partial K(\cdot) \) is bounded in a neighborhood of 0.
2. The transformation \( h^\lambda(x) \) is Holder-continuous as a function of \( \lambda \) with ratio \( r_h \) for any \( x \) in the support of \( P_S \). Similarly, the transformation \( g^\theta(x) \) is Holder-continuous as a function of \( \theta \) with ratio \( r_g \) for any \( x \) in the support of \( P_T \).

Based on Assumptions S.1, we are guaranteed to satisfy the consistency results (Thm. 2 and Thm. 3) in the main paper. The proof of Thm. S.3 is presented in the reference (1).

Theorem S.3. [Thm. 3 in the main paper] Given Assumptions S.1, when \( H_0 \) is true, with probability at least \( 1 - \alpha \),

\[
\overline{\mathbb{MMD}}(h^\lambda(X_S), g^\theta(X_T)) \leq \sqrt{\frac{2k(n_S + n_T)\log \alpha^{-1}}{n_S n_T}} + 2\sqrt{\frac{k}{n_S} + 2\sqrt{\frac{k}{n_T}}}
\]

When \( H_A \) is true, with probability at least \( 1 - \alpha \),

\[
|\overline{\mathbb{MMD}}(h^\lambda(X_S), g^\theta(X_T)) - C_*| \leq \sqrt{\frac{k}{n_S} (4 + \sqrt{C^h.\alpha + \frac{d_k}{2r_h} \log n_S})} + \sqrt{\frac{k}{n_T} (4 + \sqrt{C^g.\alpha + \frac{d_g}{2r_g} \log n_T})}
\]

where \( C_* = \min_{\lambda, \theta} \overline{\mathbb{MMD}}(h^\lambda(P_S), g^\theta(P_T)) \) is a positive constant when \( H_A \) holds (4). Here, \( C^h.\alpha = \log(2|\Omega_\lambda|) + \log \alpha^{-1} + \frac{d_k}{2r_h} \log \frac{L_k}{\sqrt{r_h}} \) and \( C^g.\alpha = \log(2|\Omega_\theta|) + \log \alpha^{-1} + \frac{d_g}{2r_g} \log \frac{L_g}{\sqrt{r_g}} \).

The proof of Thm. S.3 is presented in the reference (1).

S.3. Hypothesis Test in the Small to Moderate Sample Size Case

The threshold given in Thm. S.3 can be used for performing hypothesis tests when the sample size is large enough. But when the sample sizes are small to moderate, a data-driven method may perform better. The same observation is discussed in related papers (4) for the MMD-based two sample test.

For the data-driven method, we require one transformation class to always be the identity. In other words, we consider \( g^\theta(X_T) \) to always be \( X_T \) itself and we transform \( X_S \) using \( h^\lambda(X_S) \) to match \( X_T \). Before presenting the method, let us recall our definition of the null and alternative hypotheses.

\[ H_0 : \text{There exists a } \lambda \text{ such that } h^\lambda(X_S) \text{ matches } X_T \]
\[ H_A : \text{There is no } \lambda \text{ which matches } h^\lambda(X_S) \text{ and } X_T \]

To test these hypotheses, we construct a bootstrap-type procedure here to calculate p-values:

Algorithm S.1. Bootstrap-type hypothesis testing procedure

1. Define \( B \) to be the total number of bootstraps
2. For \( b \) in 1 : \( B \) do
3. Randomly generate \( n_S \) and \( n_T \) samples: \( \tilde{X}_S^b, \tilde{X}_T^b \) from empirical distribution based on \( X_T \) with replacement.
4. Calculate the empirical distance between two distribution, \( \text{mmd} \), by \( \text{mmd} = \overline{\mathbb{MMD}}(\tilde{X}_S^b, \tilde{X}_T^b) \)
5. Compute p-value given the test statistics \( \overline{\mathbb{MMD}}(h^\lambda(X_S), X_T) \) based on the empirical distributions of (mmd\( ^1 \), ..., mmd\( ^B \)).

Remark. We can speed up the loop in Alg. S.1 by computing distances between \( n_S + n_T \) samples and permuting them for each iteration.

We now show that our procedure is a valid hypothesis test given the following results:

Lemma S.4. Given \( n_S \) samples \( \tilde{X}_S \) from \( P_T \), we have that (from Thm. 7 in (4)) with probability at least \( 1 - \alpha \),

\[
\overline{\mathbb{MMD}}(\tilde{X}_S, X_T) \leq \sqrt{\frac{2k(n_S + n_T)\log \alpha^{-1}}{n_S n_T}} + 2\sqrt{\frac{k}{n_S} + 2\sqrt{\frac{k}{n_T}}} \tag{S.1}
\]

Meanwhile, if \( H_0 \) is true: there exists a \( \lambda_0 \) such that \( h^{\lambda_0}(X_S) \) come from the same distribution as the underlying distribution \( P_T \) for \( X_T \). Therefore, \( \overline{\mathbb{MMD}}(h^{\lambda_0}(X_S), X_T) \) is identically distributed as \( \overline{\mathbb{MMD}}(\tilde{X}_S, X_T) \),

\[
\overline{\mathbb{MMD}}(h^{\lambda_0}(X_S), X_T) \sim \overline{\mathbb{MMD}}(\tilde{X}_S, X_T) \tag{S.2}
\]
We provide additional details regarding finding a minimal d-separating set when the graphical causal model is not too complicated. Before stating the algorithm, we introduce some notations first. For any node set $A$, we define $\mathcal{A}_n(A)$ to be the ancestral set containing set $A$, that is, $\mathcal{A}_n(A) = A \cup \{\text{all ancestors of } u\}$. We call the directed subgraph composed only of nodes from $\mathcal{A}_n(A)$ as $\mathcal{DAG}_{\mathcal{A}_n(A)}$. A so-called moral graph is formed from a directed acyclic graph by adding edges between all pairs of nodes that have a common child, and then making all edges in the graph undirected. We denote the moral graph of $\mathcal{DAG}_{\mathcal{A}_n(A)}$ as $m\mathcal{DAG}_{\mathcal{A}_n(A)}$.

In order to find a minimal d-separating set $Z$ for the measurements of interest $X$ and bias nodes $E_P$ and $E_B$, as mentioned in the main paper, we only need to consider this question on the undirected graph $(m\mathcal{DAG}_{\mathcal{A}_n(X\cup E_P\cup E_B)})$ instead of the full directed acyclic graph (as shown in (5)). We give an example for finding the minimal d-separating set of the graph in the main paper, which is shown in Fig. S1. Instead of using Fig. S1, the minimal d-separating set can be found from Fig. S1(b), which represents $m\mathcal{DAG}_{\mathcal{A}_n(X\cup E_P\cup E_B)}$.

The authors (5) show that searching a d-separating set for $X$ and the bias nodes on the original DAG is equivalent to finding a node set that can block any path between $X$ and bias nodes in the moral graph of the DAG. Therefore, we can run a Breadth First Search (BFS) algorithm on the undirected graph $m\mathcal{DAG}_{\mathcal{A}_n(X\cup E_P\cup E_B)}$ to check whether $Z$ is the minimal d-separating set and adjust it if it is not. For our example in Fig. S1, by using BFS we can see that $Z = \{D, \text{age}\}$ will block any path from $X$ to the bias nodes $E_P$ and $E_B$ in Fig. S1(b).

### S.5. Optimization

We give a special case of the optimization problem in the main paper in [1] when we set $\mathcal{K}$ to be a Gaussian kernel, $g^\theta(\cdot)$ as the identity and $h^W(x) = WT_r(x)$ as a linear function for parameters $W$. Here, $T_r(x)$ is a known transformation on $x$, for example, $T_r(x) = x^2$ relates to the second order polynomial transformation. Recall that the general form of the optimization is

$$
(\hat{\lambda}, \hat{\theta}) = \arg \min_{\lambda \in \Omega_\lambda, \theta \in \Omega_\theta} \left\{ \frac{1}{n_S} \sum_{i=1}^{n_S} \mathcal{K}(h(x^{(i)}_S), \cdot) - \frac{1}{n_T} \sum_{j=1}^{n_T} \mathcal{K}(g(x^{(j)}_T, \theta), \cdot) \right\}^2
$$

For our example case, the simplified objective can be written as

$$
\hat{W} = \arg \min_{W \in \Omega_W} F(W)
$$

$$
= \arg \min_{W \in \Omega_W} \left\{ \frac{1}{n_S} \sum_{i=1}^{n_S} \mathcal{K}(WT_r(x^{(i)}_S), \cdot) - \frac{1}{n_T} \sum_{j=1}^{n_T} \mathcal{K}(x^{(j)}_T, \cdot) \right\}^2
$$

$$
= \arg \min_{W \in \Omega_W} \left\{ \frac{1}{n_S} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \mathcal{K}(WT_r(x^{(i)}_S), WT_r(x^{(j)}_S)) + \frac{1}{n_T} \sum_{j=1}^{n_T} \sum_{i=1}^{n_S} \mathcal{K}(x^{(j)}_T, x^{(i)}_T) - \frac{2}{n_ST_n} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \mathcal{K}(WT_r(x^{(i)}_S), x^{(j)}_T) \right\}
$$

$$
= \arg \min_{W \in \Omega_W} \left\{ \frac{1}{n_S} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \exp(\|WT_r(x^{(i)}_S) - WT_r(x^{(j)}_S)\|^2_2) - \frac{2}{n_ST_n} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \exp(\|WT_r(x^{(i)}_S) - x^{(j)}_T\|^2_2) \right\}
$$
This is a continuous optimization and its gradient with respect to $F$ is:

$$\nabla F(W) = \frac{2}{n_S} \sum_{i=1}^{n_S} \sum_{j=1}^{n_S} \exp(||W(T_r(x'_i) - T_r(x'_j))||^2) tr(W^TW_r(T_r(x'_i) - T_r(x'_j))(T_r(x'_i) - T_r(x'_j))^T) - \frac{4}{n_SN_T} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \exp(||WT_r(x'_i) - x'_j||^2) (T_r(x'_i)^T W_r(T_r(x'_i) - x'_j))$$  \[S.5\]

Since the objective is continuous, the optimization can be performed by gradient descent or stochastic gradient descent. A useful prior based on some domain knowledge can also help the optimization scheme since the objective is non-convex.

We now provide a result to setup a confidence region to choose a good initial point for the optimization.

**Lemma S.5** (Lemma 4.4 in (1) supplement). Under $H_0$, the identity $g^\theta(\cdot)$ with $h^\theta(x) = WT_r(x)$, we have

$$\Omega_W := \left\{ \frac{1}{\min(n_S, n_T)} \sum_{i=1}^{\min(n_S, n_T)} ||x'_i - WT_r(x'_i)||_2^2 \leq 3 \text{tr}(VAR(P_T)) + \epsilon \right\},$$

where $\forall \theta \in \mathbb{R}$ is the variance and $\text{tr}(\cdot)$ is the trace. For any $\epsilon, \alpha > 0$ and a sufficiently large sample size, a neighborhood of the true $W_0$ is contained in $\Omega_W$ with probability at least $1 - \alpha$.

**S.6. Proof for Identifiability Conditions (Thm. 1 in the main paper)**

In this subsection, we provide a proof for the identifiability conditions.

**Theorem S.6.** [Thm. 1 in the main paper] The distribution shift correction is identifiable if there exists a known set of variables $X$ such that the following three conditions are all concurrently satisfied:

1) $X$ d-separates $Z$ and $E_B$ (sample selection bias) and also d-separates $X$ and $E_P$ (population characteristic difference);

2) The conditional probability $P(X|Z)$, after appropriate transformations on $X$, is the same across multiple participating sites ($S$ and $T$);

3) The distribution of $Z$ has a non-trivial overlap across multiple sites ($S$ and $T$), which means that there exists an interval $[a, b]$ such that $P(a \leq Z \leq b) \geq 0.5$ for all sites.

**Proof.** We denote measurements of interest as $X_S$ in dataset 1 and $X_T$ in dataset 2. We assume $E_B = 1$, $E_P = 1$ for dataset 1 and $E_B = 2$, $E_P = 2$ for dataset 2 to represent the biases between the two datasets. Without loss of generality, we assume that a transformation $h^\lambda(\cdot)$ can resolve the distributional shift, that is, marginal distributions of $h^\lambda(X_S)$ and $X_T$ are the same. However, when biases exist, we have,

$$P(h^\lambda(X_S)|E_B = 1, E_P = 1) \neq P(X_T|E_B = 2, E_P = 2).$$ \[S.6\]

Therefore, we may not be able to find the correct transformation by matching the distributions (1). Then, the correction problem becomes non-identifiable if we do not have any additional information. The situation changes when we have a set of variables $X$ which satisfy the three identification conditions in Thm. S.6 (Thm. 1 in the main body). The following explanation describes why.

$$P(h^\lambda(X_S)|E_B = 1, E_P = 1) = E_{Z|E_B=1,E_P=1}[P(h^\lambda(X_S)|Z, E_B = 1, E_P = 1)]$$ \[S.7\]

$$= E_{Z|E_B=1,E_P=1}[P(h^\lambda(X_S)|Z)]$$ \[S.8\]

The second equation holds because of condition 1) in Thm. S.6. Similarly, we have,

$$P(X_T|E_B = 2, E_P = 2) = E_{Z|E_B=2,E_P=2}[P(X_T|Z)]$$ \[S.9\]

Then, since condition 2) holds, we have $P(h^\lambda(X_S)|Z) = P(X_T|Z)$, that is, they are identical functions of $Z$. Therefore, the only difference between the two datasets are $P(Z|E_B = 1, E_P = 1)$ and $P(Z|E_B = 2, E_P = 2)$. We should keep in mind that this difference is similar to $[S.6]$. However, there is no longer an unknown transformation $h^\lambda(\cdot)$ in the relation. To address this issue, we can conduct a subsampling procedure at random $Z$ to approximately align $P(Z|E_B = 1, E_P = 1)$ with $P(Z|E_B = 2, E_P = 2)$. The condition 3) in Thm. S.6 shows that this is possible. After the subsampling procedure, we will approximately have,

$$P(h^\lambda(X_S)|E_B = 1, E_P = 1, SSP) = P(X_T|E_B = 2, E_P = 2, SSP).$$ \[S.10\]

Therefore, the correction problem becomes identifiable since we can now learn $h^\lambda$ by matching the distributions.
S.7. Proof of extended IJ estimators

We provide a proof of our extension of the infinitesimal Jackknife estimator (6) to multiple groups. The proofs proceed in two steps.

**Proof. Step 1.**
Before describing the proof, we define some notations. We have two datasets $X_S$, $X_T$ and $d$ groups for each dataset. Similar to the discussion in the main paper, we define our samples from one dataset as

$$X_S = (x_S^{(1,1)}, ..., x_S^{(1,n_S)}, ..., x_S^{(d,1)}, ..., x_S^{(d,n_S)})$$

and samples from the other dataset as

$$X_T = (x_T^{(1,1)}, ..., x_T^{(1,n_T)}, ..., x_T^{(d,1)}, ..., x_T^{(d,n_T)}).$$

Based on $X = (X_S, X_T)$, we can generate bootstrap samples for every group and we call all such generated samples as $V = (V_S, V_T)$. In this proof (similar to (6)), we assume that the bootstrap sample sizes are the same as the original samples, that is,

$$V_S = (v_S^{(1,1)}, ..., v_S^{(1,n_S)}, ..., v_S^{(d,1)}, ..., v_S^{(d,n_S)})$$

and

$$V_T = (v_T^{(1,1)}, ..., v_T^{(1,n_T)}, ..., v_T^{(d,1)}, ..., v_T^{(d,n_T)}).$$

Based on $V = (V_S, V_T)$, we define the count variables $N(V) = (N(V_S), N(V_T))$ and probability variables $P(V) = (P(V_S), P(V_T))$, which are,

$$N(V_{u})^{(i,j)} = \# \{ k = 1, ..., n_u | V_{u}^{(i,k)} = X_{u}^{(i,j)} \}, \text{ for any } u \in \{S, T\}, i \in \{1, ..., d\}, j \in \{1, ..., n_u\}.$$

$$P(V_{u})^{(i,j)} = \frac{1}{n_u} N(V_{u})^{(i,j)}, \text{ for any } u \in \{S, T\}, i \in \{1, ..., d\}, j \in \{1, ..., n_u\}.$$

In other words, $N(V_{u})^{(i,j)}$ records how many times $X_{u}^{(i,j)}$ appears in the bootstrap samples for each iteration. The probability variable $P(V_{u})^{(i,j)}$ is the normalized $N(V_{u})^{(i,j)}$ such that $\sum_{j=1}^{n_u} P(V_{u})^{(i,j)} = 1$. The bootstrap process is repeated $B$ times. Later, when we have superscript $b$ for $V$, $N(V)$ and $P(V)$, it implies that those variables are related to the $b^{th}$ iteration of the bootstrap process. Now, we define another term to be the baseline for the probability variables, which is,

$$P(X) = (P(X_S), P(X_T)), \text{ where } P(X_{u})^{(i,j)} = \frac{1}{n_u}, \text{ for any } u \in \{S, T\}, i \in \{1, ..., d\}, j \in \{1, ..., n_u\}.$$

We can see that $P(X)$ represents the uniform distribution for each group of $X$, which only depends on the original samples $X$. On the other hand, $P(V^b)$ depends on $X$ and the bootstrap samples $V^b$ for the $b^{th}$ iteration.

The probability variables $P(V^b)$ and the baseline probability variables $P(X)$ are connected by multinomial distributions as

$$n_u \times P(V^b) \sim Multi_{n_u}(n_u, P(X)^{(i,j)}), \text{ for any } u \in \{S, T\}, i \in \{1, ..., d\}, b \in \{1, ..., B\}, \tag{S.11}$$

where $P(V^b)^{(i,j)}$ is related to the $i^{th}$ group from dataset $V^b_u$ and $P(X)^{(i,j)}$ is related to the $i^{th}$ group from dataset $X_u$. In $Multi_{n_u}(n_u, P(X)^{(i,j)})$, $n_u$ is the total number of variables in the $i^{th}$ group whereas $P(X)^{(i,j)}$ is the uniform distribution for this group. This means that $n_u \times P(V^b)^{(i,j)}$ can be viewed as $n_u$ samples generated from the multinomial distribution on $n_u$ discrete values with equal probability to be drawn.

Let us consider an estimation function $f$. For the $b^{th}$ iteration in the bootstrap process, we obtain an estimator $\hat{f}^b$ from samples $V^b$. We assume that, given $X$, $\hat{f}$ only depends on $P(V^b)$, which means that we can represent $\hat{f}^b$ by $f(P(V^b))$. Therefore, we can approximate $\hat{f}^b$ via the tangent hyperplane that goes through $f(P(X))$ at $P(X)$, which gives us that

$$f_{TAN}(P(V^b)) = f(P(X)) + \langle P(V^b) - P(X), \mathcal{U} \rangle \tag{S.12}$$

where $\langle P(V^b) - P(X), \mathcal{U} \rangle = \sum_{u=S,T} \sum_{i=1}^{d} \langle P(V^b_{u})^{(i)} - P(X_{u})^{(i)}, \mathcal{U}^{(i)}_{u} \rangle$, and $\mathcal{U}^{(i)}_{u} = (\mathcal{U}^{(i,1)}_{u}, ..., \mathcal{U}^{(i,n_u)}_{u})$,

$$\mathcal{U}^{(i,j)}_{u} = \lim_{\epsilon \to 0} \frac{f(P(X_{u})^{(i)} + \epsilon e_{j} - P(X_{u})^{(i)})) - f(P(X_{u})^{(i)}))}{\epsilon}. \tag{S.13}$$

where $e_{j}$ is a vector with all zeros except one on the $j^{th}$ position.

Because our bootstrap samples are drawn independently from every group, we have that

$$E[f_{TAN}(P(V^b)) - f(P(X))]^2 = \sum_{u=S,T} \sum_{i=1}^{d} E[(P(V^b_{u})^{(i)} - P(X_{u})^{(i)}, \mathcal{U}^{(i)}_{u})]^2. \tag{S.14}$$
Further, using the multinomial distribution relation between $\mathcal{P}(V^b)$ and $\mathcal{P}(X)$ in [S.11], we get
\[
E[(P(V^b)^{(i)}) - P(X)^{(i)}, U^b_u)]^2 = (U^b_u)^t E[(P(V^b)^{(i)}) - P(X)^{(i)}, (P(V^b)^{(i)} - P(X)^{(i)})^t] U^b_u
\]
\[
= \frac{1}{n_u^2} \sum_{i,j=1}^{n_u^i} (\mathcal{U}_u^{(i,j)})^t (I - \frac{1}{n_u^2}) \mathcal{U}_u^{(i,j)} = \sum_{i,j=1}^{n_u^i} \frac{1}{n_u^2} (\mathcal{U}_u^{(i,j)})^2 - \frac{1}{n_u^2} (\sum_{j=1}^{n_u^i} \mathcal{U}_u^{(i,j)})^2,
\]
where $(\cdot)^t$ represents the transpose operation on a vector or matrix, and $I$ is the all one vector. Further, since $\sum_{j=1}^{n_u^i} (X_j - \mathcal{P}(X)^{(i)}) = 0$, we have $\sum_{j=1}^{n_u^i} \mathcal{U}_u^{(i,j)} = 0$ from its definition in [S.13]. Therefore, we get an approximation for the mean squared error (MSE) of $\hat{b}$ for estimation $f(\mathcal{P}(X))$, that is
\[
E[f_{\text{TAN}}(P(V^b)) - f(\mathcal{P}(X))]^2 = \sum_{u=S,T} \sum_{i=1}^{d} \frac{1}{n_u^2} (\sum_{j=1}^{n_u^i} \mathcal{U}_u^{(i,j)})^2.
\]

Further, if $E[f(P(V^b))] = f(\mathcal{P}(X))$, then this MSE is the bootstrap estimation for the variance of $f(\mathcal{P}(X))$ (6).

**Step 2.**
Now, we consider the estimation regarding the variance of $\hat{\lambda} := \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^b$. In this proof, we assume that $B$ is large enough that it exactly covers all finite possibilities for the bootstrap samples. Therefore, we can consider $\hat{\lambda}$ as the value of a function $f(\cdot)$ at $\mathcal{P}(X)$. In other words, we consider $f(P(X))$ in Step 1 to be $\hat{\lambda}$. Let us imagine that we generate bootstrap samples to estimate the variance of $\hat{\lambda}$ as in Step 1, which is a second layer of bootstrap since $\hat{\lambda}$ is already based on bootstrap samples. Again, we can use the tangent hyperplane scheme in Step 1 to approximate the bootstrap estimation of the variance. For this estimator, the relation $E[f(P(V^b))] = f(P(X)) = \hat{\lambda}$ holds because we assume that $B$ covers all the possibilities. Therefore, the only remaining issue is to calculate $\mathcal{U}_u^{(i,j)}$. We define a probability measure on $X$ which is
\[
\mathcal{P}(V_u)^{(i,j)} = \{P(X)^{(i)} + \epsilon_j - P(X)^{(i)}\} \quad \text{for the group } X_u^{(i)},
\]
and $\mathcal{P}(X_v)^{(k)}$ for other groups $X_v^{(k)}$, for any $v \in \{S,T\}, k \in \{1,2,...,d\}$.

Because $B$ bootstrap samples covers all possibilities, we have that
\[
f(\mathcal{P}(V_u)^{(i,j)}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^b w^b(\mathcal{P}(V_u)^{(i,j)})
\]
where $w^b(\mathcal{P}(V_u)^{(i,j)}) = \frac{\mathbb{P}(V^b|^\mathcal{P}(V^b))}{\mathbb{P}(V^b|^\mathcal{P}(X))}$. The probability $\mathbb{P}(V^b|\mu)$ is defined to be the probability of $V^b$ under the probability measure $\mu$. We can further simplify $w^b(\mathcal{P}(V_u)^{(i,j)})$ by
\[
w^b(\mathcal{P}(V_u)^{(i,j)}) = \frac{\mathbb{P}(V^b|^\mathcal{P}(V_u)^{(i,j)})}{\mathbb{P}(V^b|^\mathcal{P}(X))}
\]
\[
= (1 + (n_u^i - 1)\epsilon) \prod_{l=1, l \neq j}^{n_u^i} (1 - \epsilon)^n_u^l \mathbb{P}(V^b)^{(i,l)}
\]
\[
= (1 + (n_u^i - 1)\epsilon)^{n_u^i} \mathbb{P}(V^b)^{(i,j)} (1 - \epsilon)^n_u^i - n_u^j \mathbb{P}(V^b)^{(i,j)}
\]
Let $\epsilon \to 0$, we have that
\[
[w^b(\mathcal{P}(V_u)^{(i,j)}) - 1]/\epsilon = [(1 + (n_u^i - 1)\epsilon)^{n_u^i} \mathbb{P}(V^b)^{(i,j)} (1 - \epsilon)^n_u^i - n_u^j \mathbb{P}(V^b)^{(i,j)} - 1]/\epsilon
\]
\[
\to (1 + n_u^i \epsilon)^{n_u^i} \mathbb{P}(V^b)^{(i,j)}(1 - \epsilon)^n_u^i - n_u^j \mathbb{P}(V^b)^{(i,j)} - 1)/\epsilon
\]
\[
\to n_u^i \mathbb{P}(V^b)^{(i,j)} - 1)
\]
Therefore, using [S.13], we know that
\[
\frac{1}{n_u^2} \mathcal{U}_u^{(i,j)} = \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^b (n_u^i \mathbb{P}(V_u)^{(i,j)} - 1)
\]
By plugging in that form in equation [S.16], we get the approximation for the bootstrap estimation for the variance of $\hat{\lambda}$, which is
\[
\text{VAR}(\hat{\lambda}) = \sum_{u=S,T} \sum_{i=1}^{d} \sum_{j=1}^{n_u^i} (\text{COV}(\hat{\lambda}, \mathcal{N}(V_u)^{(i,j)}))^2,
\]
where $\text{COV}(\hat{\lambda}, \mathcal{N}(V_u)^{(i,j)}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^b ((V_u^b)^{(i,j)} - 1)$.
For finite $B$, $\frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^b((V_u^b)^{(i,j)} - 1)$ is an approximation of $\text{COV}(\hat{\lambda}, N(V_u)^{(i,j)})$. For the subsampling case, the expectation of $(V_u^b)^{(i,j)}$ is no longer 1 but $\frac{a_i}{n_c}$. As a result, we adjust the covariance term to $\text{COV}(\hat{\lambda}, N(V_u)^{(i,j)}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^b((V_u^b)^{(i,j)} - \frac{a_i}{n_c})$ as in (7). This leads to the Thm. 4 in the main paper.

\section{S.8. Simulations}

Our first set of experiments were designed to check the efficacy of the hypothesis test whereas the second experiment evaluated the estimation consistency.

\textbf{Simulation for the Hypothesis Test.} We generated samples from the standard normal distribution $N(0,1)$ to synthesize the first dataset $X_S$ and use the normal distribution $N(10,2)$, with mean 10 and standard deviation 2, to synthesize another dataset $X_T$. Notice that, under the correct transformation class $h^{(a,b)}(X_S) = aX_S + b$, we can correct the distribution shift and we should accept $H_0$. We consider two types of variations that can potentially affect the correction and check whether the hypothesis test indeed rejects $H_0$ at a high power.

1) In Fig. S2(a), we always choose the transformation class $h^{(a,b)}(X_S) = aX_S + b$ and one dataset $X_T$ comprised of samples from $N(10,2)$. But we vary the generating distribution for the other dataset choosing between $N(0,1)$, $\text{Laplace}(0,1)$, and $\text{Exponential}(1)$. We also vary the sample sizes. The hypothesis test with significance level 0.05 is performed on 100 repetitions and the acceptance rate curves are plotted. From Fig. S2(a) we see that our hypothesis test does accept $H_0$ at a high rate when it is true (red curve) and tends to reject $H_0$ with an increase in power as the sample size increases whenever the generating distributions are not Normal (blue and black curves).

2) In Fig. S2(b), we always choose one dataset $X_T$ composed of samples from $N(10,2)$ and the generating distribution for the other dataset $X_S$ is set to be $N(0,1)$. Then, we vary the transformation class $h^{(a,b,c)}(X_S)$ choosing between $aX_S + bX_S + c$ and $a\log(|x|) + b$. Here, note that $aX_S^2 + bX_S + c$ includes the true transformation whereas $a\log(|x|) + b$ corresponds to an incorrect transformation class. Again, we vary the sample sizes and repeat the hypothesis test 100 times and plot the acceptance curves in Fig. S2(b). Similar to the first setting, we observe that our hypothesis test accepts $H_0$ at a high rate when it is true (red curve) and tends to reject $H_0$ with high power as the sample size increases in the setting where the transformation class is wrong (blue curve).

\textbf{Simulation for Estimation Consistency.} In this simulation, we assume that the distributional shift is the only variation across the datasets, $X_S$ and $X_T$. We check the estimation consistency of the transformation $h^W(\cdot)$ which minimizes the MMD distance between $h^W(X_S)$ and $X_T$.

In Fig. S2(c), we perform the experiments for two models. For model 1, the generating distributions for two datasets are,

$X_S \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$

$X_T^{\text{raw}} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$

$X_T = \begin{pmatrix} 1 & 2 & 10 \\ 2 & 1 & -20 \end{pmatrix} \begin{pmatrix} X_T^{\text{raw}} \\ 1 \end{pmatrix}$

In other words, we generate samples for $X_S$ and $X_T^{\text{raw}}$, and we transform $X_T^{\text{raw}}$ to get $X_T$. The transformation class we consider is

$X_T = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} X_S \\ 1 \end{pmatrix}$
We define the quadratic mean of the estimation error for "Model 1, first row" in Fig. S2(c) as

$$\sqrt{\frac{(a_{11} - 1)^2 + (a_{12} - 2)^2 + (a_{13} - 10)^2}{3}}$$

Similar, the quadratic mean of the estimation error is defined for the second row \((a_{21}, a_{22}, a_{23})\). The plot shows us that the estimation error is small and decreases as the sample size increases. We also check this behavior under different types of generating distributions, including

\[
X_T^{raw} \sim \left( \frac{N(0, 1)}{\chi^2_1} \right)
\]

\[
X_S \sim \left( \frac{N(0, 1)}{\chi^2_1} \right)
\]

\[
X_T = \begin{pmatrix} 1 & 2 & 10 \\ 2 & 1 & -20 \end{pmatrix} \begin{pmatrix} X_T^{raw} \\ 1 \end{pmatrix}
\]

We call this model 2 and observe a similar trend as model 1 as shown in Fig. S2(c).

S.9. ADNI acknowledgments

Apart of the data used in our experiments come from the ADNI study. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lundbeck; Merck & Co., Inc.; Mesobore Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

1. Zhou HH, et al. (2016) Hypothesis testing in unsupervised domain adaptation with applications in Alzheimer’s disease in Advances in Neural Information Processing Systems 29.
2. Shaw LM, et al. (2009) Cerebrospinal fluid biomarker signature in Alzheimer’s disease neuroimaging initiative subjects. Annals of neurology 65(4).
3. Mattsson N, et al. (2011) The Alzheimer’s Association external quality control program for cerebrospinal fluid biomarkers. Alzheimer’s & Dementia 7(4).
4. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Simola A (2012) A kernel two-sample test. Journal of Machine Learning Research 13(Mar).
5. Tian J, Paz A, Pearl J (1998) Finding minimal d-separators, Technical report.
6. Efron B (2014) Estimation and accuracy after model selection. Journal of the American Statistical Association 109(507).
7. Wager S, Athey S (2017) Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association (just-accepted).