My journey to a robust method for causal inference in the new-user cohort design with relatively small sample:

Preliminary results of the experiments with Representation Learning-Propensity score model

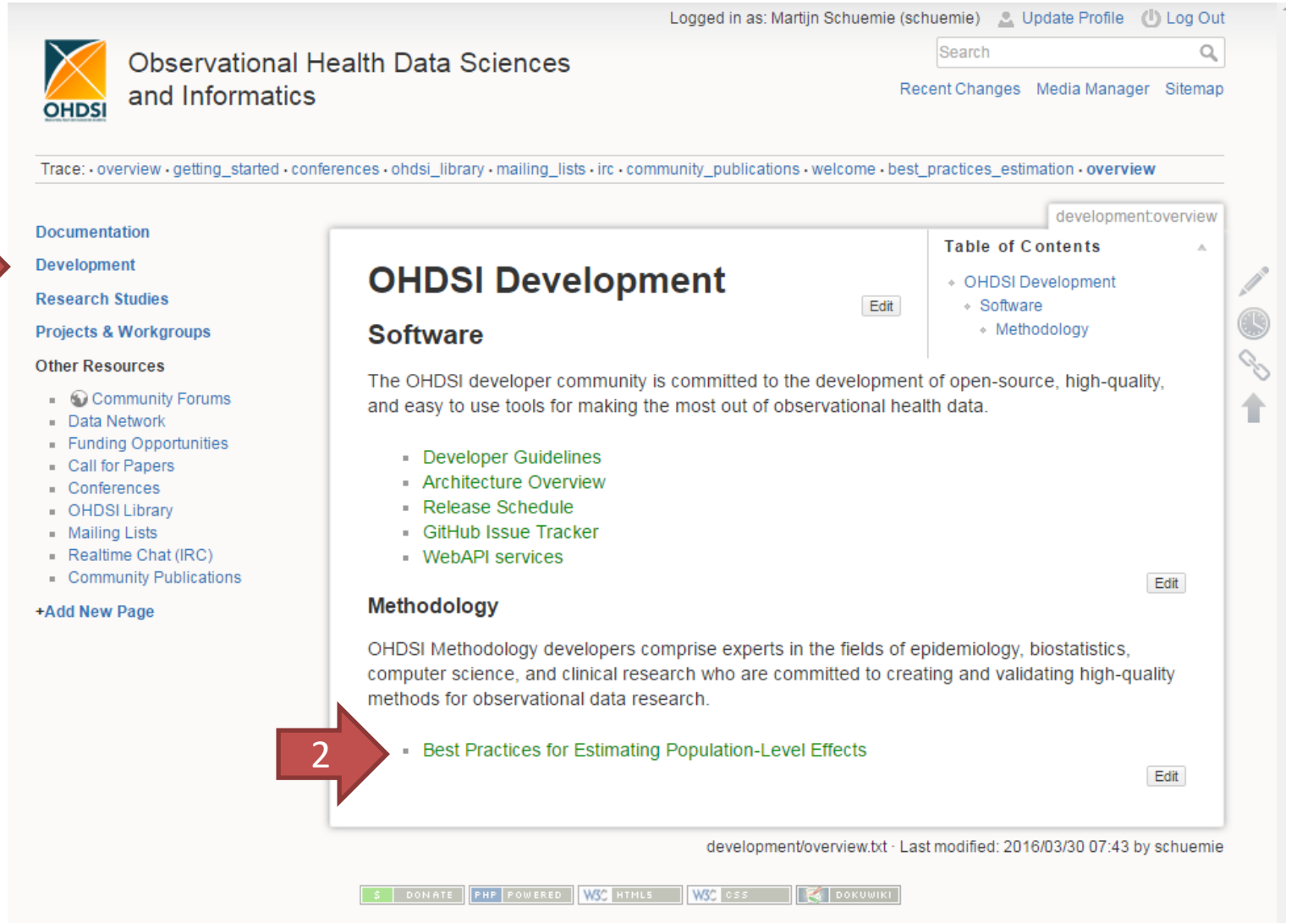
Caution: the results presented here are preliminary and might not be replicable

Seng Chan You

Causal inference

- For the observational data, the core question is how to get the counterfactual outcome. This is challenging for two reasons
  - 1. We only observe the factual outcome and never the counterfactual outcomes
  - 2. Treatments are typically not assigned at random in observational data

OHDSI best practices



https://www.ohdsi.org/web/wiki/doku.php?id=development:best_practices_estimation

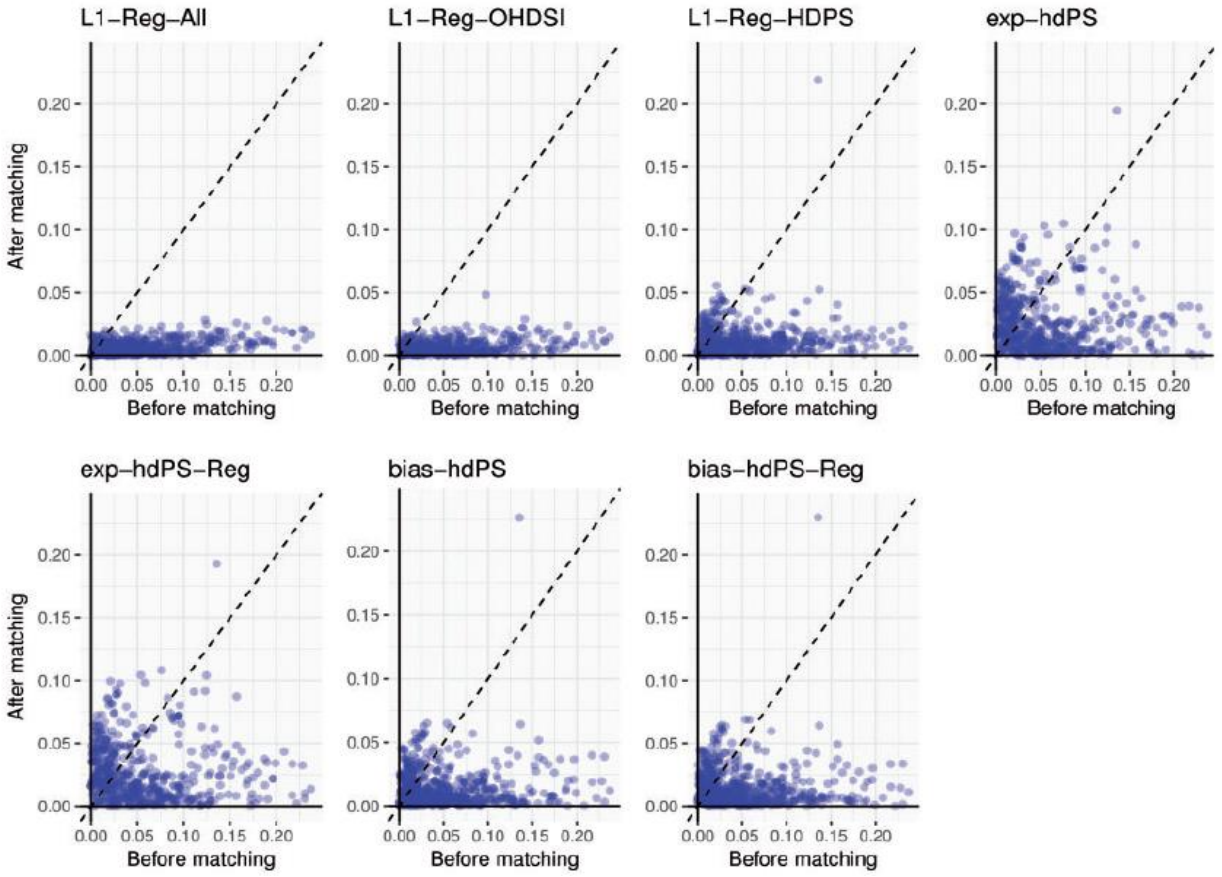
OHDSI best practice for new-user cohort design

- Use **propensity scores** (PS)
- Build PS model using **regularized regression** and a **large set of candidate covariates** (as implemented in the CohortMethod package)
- Use either **variable-ratio matching** or **stratification** on the PS
- **Compute covariate balance** after matching for all covariates, and **terminate study if a covariate has standardized difference > 0.1**

https://www.ohdsi.org/web/wiki/doku.php?id=development:best_practices_estimation

Why OHDSI recommends large-scale PS matching?

- Large-scale PS matching uses L_1 **statistical regularization** (LASSO) conducting a penalized likelihood regression with all covariates simultaneously
- Large-scale PS matching provides improved confounding control as compared with the high-dimensional PS for propensity score model selection



Tian and Schuemie et al., Int J Epidemiology, 2018

Re-consider Propensity score matching

- The propensity score **collapses the covariates of an observational study into a single measure** summarizing their joint association with treatment conditions.
- Like propensity scores, **prognostic scores** can reduce the dimension of the covariate, yet causal inferences conditional on them are as valid as are inferences conditional only on the unreduced covariate.
- Current OHDSI large-scale propensity score matching usually employees more than 5,000 covariates for each comparison.
  - **When the number of covariates is large relative to the number of observations, controlling for all observed covariates become infeasible** and selection based on substantive knowledge becomes impractical

Hanse et al., Biometrika, 2008

Leveraging prognostic score (disease risk score)

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; **21**(S2): 138–147
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.3231

ORIGINAL REPORT

Role of disease risk scores in comparative effectiveness research with emerging therapies

Robert J. Glynn*, Joshua J. Gagne and Sebastian Schneeweiss

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

ABSTRACT

Background Usefulness of propensity scores and regression models to balance potential confounders at treatment initiation may be limited for newly introduced therapies with evolving use patterns.

Objectives To consider settings in which the disease risk score has theoretical advantages as a balancing score in comparative effectiveness research because of stability of disease risk and the availability of ample historical data on outcomes in people treated before introduction of the new therapy.

Methods We review the indications for and balancing properties of disease risk scores in the setting of evolving therapies and discuss alternative approaches for estimation. We illustrate development of a disease risk score in the context of the introduction of atorvastatin and the use of high-dose statin therapy beginning in 1997, based on data from 5668 older survivors of myocardial infarction who filled a statin prescription within 30 days after discharge from 1995 until 2004. Theoretical considerations suggested development of a disease risk score among nonusers of atorvastatin and high-dose statins during the period 1995–1997.

Results Observed risk of events increased from 11% to 35% across quintiles of the disease risk score, which had a *C*-statistic of 0.71. The score allowed control of many potential confounders even during early follow-up with few study endpoints.

Conclusions Balancing on a disease risk score offers an attractive alternative to a propensity score in some settings such as newly marketed drugs and provides an important axis for evaluation of potential effect modification. Joint consideration of propensity and disease risk scores may be valuable.

Table 4. Crude and adjusted relative odds of recurrent myocardial infarction, stroke, or death within 1 year after initiation of statins among myocardial infarction survivors, 1997–2005; 5189 statin initiators, 1851 with atorvastatin, and 922 with high-dose statins

	Odds ratio	95%CI
Model: atorvastatin versus other		
Crude estimate	0.92	0.80–1.05
Adjusted for disease risk	0.93	0.81–1.07
Model: high dose versus other		
Crude estimate	0.93	0.78–1.11
Adjusted for disease risk	0.94	0.79–1.12

Leveraging prognostic score (disease risk score)

- While the DRS can be more stable over time, modeling the DRS in practice also presents unique challenges that are not shared by PS
  - Unlike PS, which models covariate associations with treatment, the DRS models covariate associations with the potential outcome under the control or comparator treatment
  - In practice, however, this potential outcome is not observed for all individuals in the study population, but only for those receiving the comparator treatment
  - Can covariates used for DRS be really independent from the treatment allocation? Eg, when compare the GI bleeding between Warfarin versus NOAC
    - NOAC is related with lower risk of future GI bleeding
    - INR testing or valvular heart disease might be associated with the prescription of warfarin. The large-scale DRS would learn these variables to predict GI bleeding event.

R Wyss and Til Stumer et al., Pharmacoepidemiol Drug Saf, 2015

Leveraging prognostic score (disease risk score)

KEY POINTS

- In theory, the degree of overlap in the distribution of disease risk across treatment groups will always be at least as large as the overlap in the propensity score (PS) across treatment groups.
- Controlling for a high-dimensional set of covariates can improve confounding control but increases separation between the PS distributions of the treatment groups while having less impact on the separation between the disease risk distributions of the treatment groups.
- Matching on the disease risk score (DRS) can allow researchers to evaluate the treatment effect within a larger proportion of treated individuals, compared with matching on the PS. However, accurately modeling the DRS can be challenging compared with the PS, even in settings involving newly introduced treatments.

Table 3. Empirical results comparing new users of dabigatran with new users of warfarin in preventing combined ischemic stroke and all-cause mortality in the Medicare population between 19 October 2010 and 31 December 2012

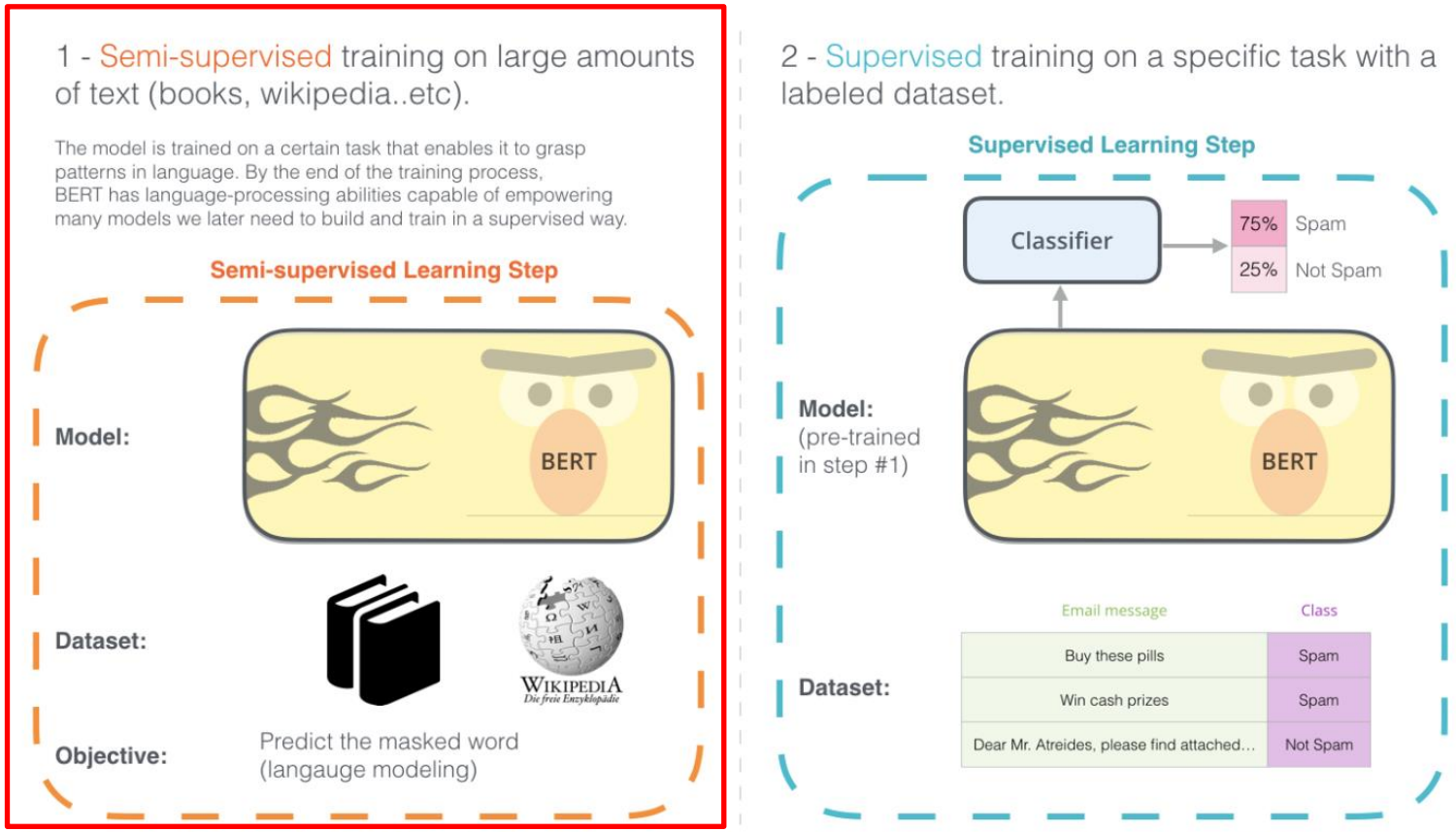
Sample size*	# Covs†	Method	Hazard ratio‡	Standard. error§	95%CI	% Matched	Model fit¶		
							c-Stat	*p*-Value	ASAMD‖
20% sample		Unadjusted	0.48	0.02	(0.46, 0.50)	—	—	—	0.14
	37	PS match	0.75	0.03	(0.70, 0.80)	99.9	0.68	0.16	<0.01
		DRS match	0.73	0.03	(0.69, 0.77)	100	0.73	<0.01	—
	237	PS match	0.88	0.04	(0.81, 0.95)	99.2	0.73	0.18	<0.01
		DRS match	0.87	0.04	(0.81, 0.94)	99.7	0.78	<0.01	—
1% sample		Unadjusted	0.47	0.07	(0.41, 0.54)	—	—	—	0.17
	37	PS match	0.75	0.14	(0.57, 0.99)	98.5	0.71	0.49	0.01
		DRS match	0.74	0.14	(0.57, 0.98)	99.1	0.73	0.18	
	237	PS match	0.89	0.19	(0.61, 1.29)	92.0	0.79	0.47	0.01
		DRS match	0.85	0.16	(0.62, 1.16)	98.5	0.78	<0.01	—

*20% ($N = 67\,667$) and 1% ($N = 3383$) samples of the Medicare data. The 20% sample consisted of 11 407 dabigatran new users. The 1% sample consisted of 576 dabigatran new users.
†Number of covariates in propensity score (PS) and disease risk score (DRS) model.
‡RELY trial relative risk for 150 mg dabigatran versus warfarin: 0.76 (0.60, 0.98) for ischemic stroke and 0.88 (0.77, 1.00) for death from any cause. In the current study, >90% of the outcomes were death from any cause.
§Bootstrapped standard errors. Hazard ratio estimates are the mean of the bootstrapped sampling distribution.
¶*c*-Statistic and *p*-value for each PS model and DRS model.
‖The average standardized absolute difference (ASAMD) of covariates across PS-matched treatment groups. Because the DRS does not balance covariates across treatment, the ASAMD was only calculated for PS models. The unadjusted ASAMD was calculated for all 237 covariates.

R Wyss and Til Stumer et al., Pharmacoepidemiol Drug Saf, 2015

Background of dimension reduction using data-driven representation learning

- 1. Fundamentally, large-scale propensity score model is a data-driven dimension reduction method agnostic about the exposure-outcome association and the effects of the covariates on this association
- 2. **BERT** model, leading recent advance in natural language process (NLP), leveraged **representation learning** with **large unlabeled data** → Then, fine tuning with labeled data for specific task of interest



The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. [Source for book icon].

Size of dimension for word piece embedding: 30,000

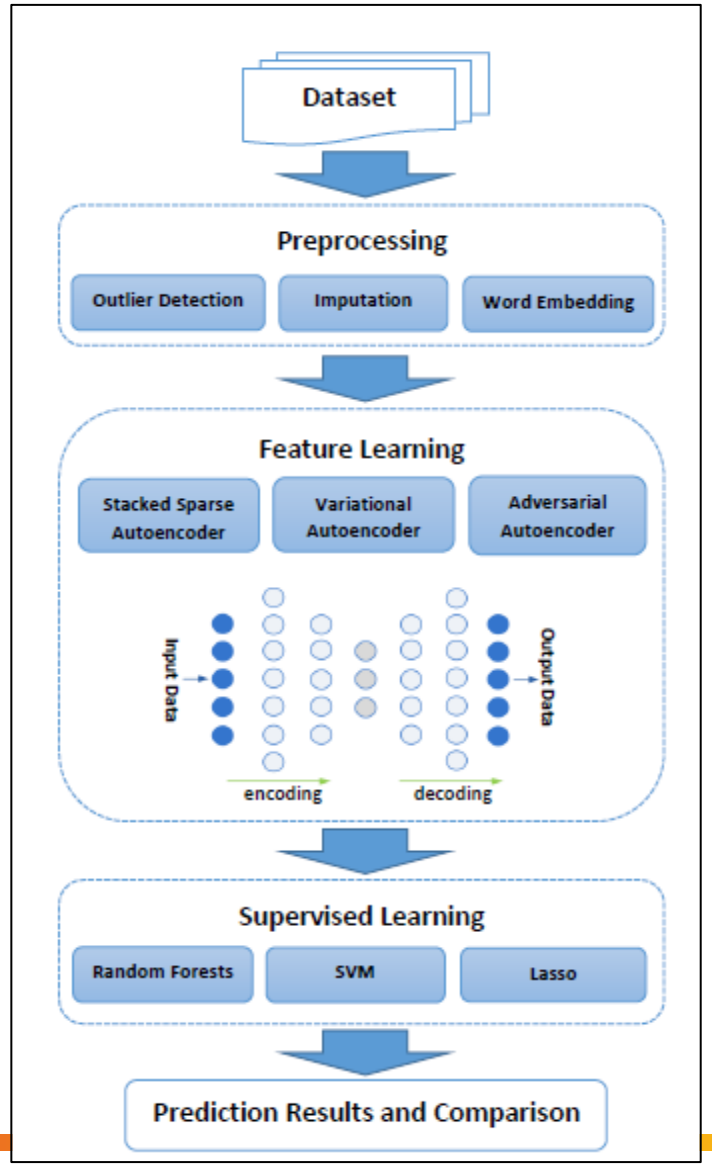
Using 3.3 billion word corpus

http://jalammar.github.io/illustrated-bert/

Previous attempts for Data-driven dimension reduction using autoencoders for EHRs

Representation Learning with Autoencoders for Electronic Health Records: A Comparative Study

Najibesadat Sadati[a] , Milad Zafar Nezhad[a], Ratna Babu Chinnam[a], Dongxiao Zhu[b,*]
Department of Industrial and Systems Engineering, Wayne State University[a]
Department of Computer Science, Wayne State University[b]
Corresponding author*, E-mail address: dzhu@wayne.edu



Dimension reduction and shrinkage methods for high dimensional **disease risk scores** in historical data

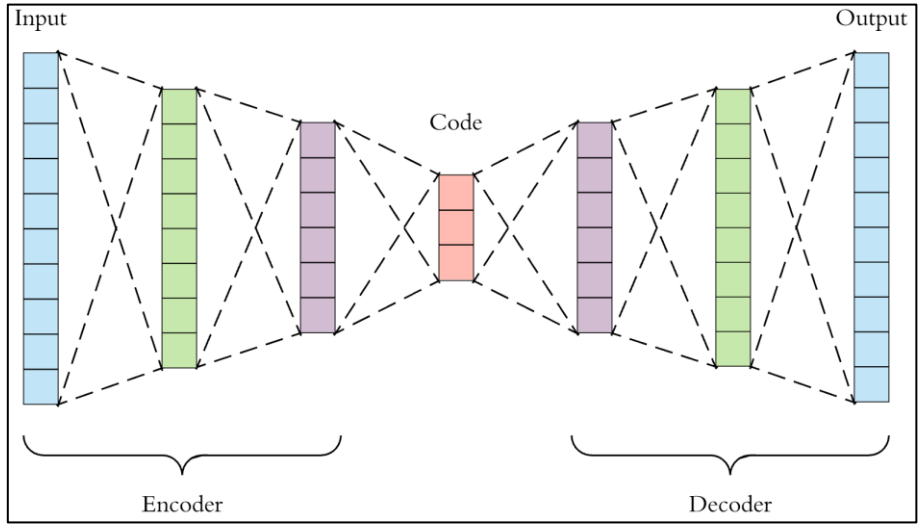
- In a high dimensional data setting, empirical selection of hundreds of potential confounders and modeling of DRS in the historical cohort can lead to over-fitting and reduced predictive performance in the study cohort
- Kumamaru and Schneeweiss et al. found that the use of **combination of dimension reduction** (PCA) and **shrinkage methods** (lasso or ridge regression) in high-dimensional DRS model had higher c-statistics and closer odd ratios to the benchmark estimates than an unreduced model in hd-DRSs from historical data in two empirical study examples (dabigatran vs warfarin; coxibs vs NSAIDs)

➔ How about combination of dimension reduction (deep learning autoencoder) based on historical data and shrinkage methods (lasso) for propensity score model for emerging therapy?

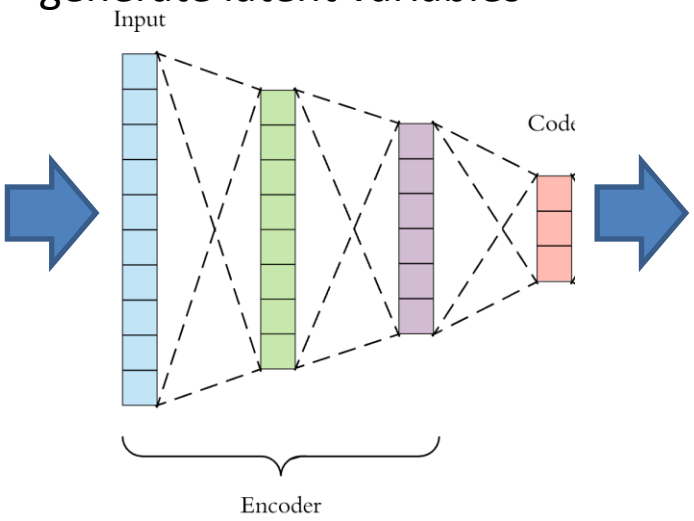
Kumamaru and Schneeweiss et al., Emerg Themes Epidemiol, 2016

My proposal: Representation Learning-Propensity score model (RLPS)

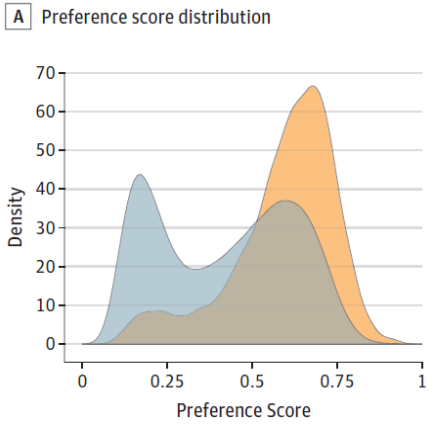
Step 1. Training autoencoder in large historical data

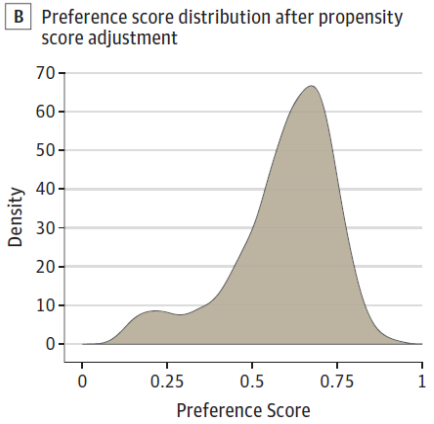


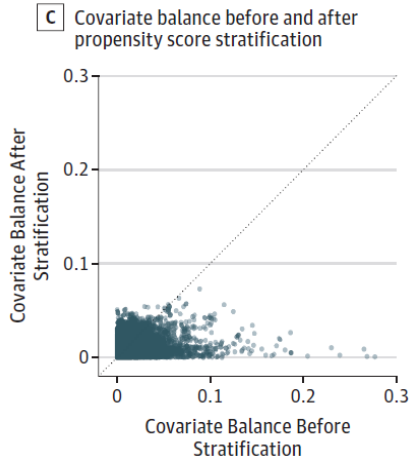
Step 2. Applying encoder against small study population to generate latent variables



Step 3. Building PS model with latent variables with shrinkage method and adjustment







Experiment 1) Coxib Vs NSAIDS

- According to the vignette of the CohortMethod
- I generated Coxib and NSAIDs cohorts from EHR database

3.2 Preparing the exposures and outcome(s)

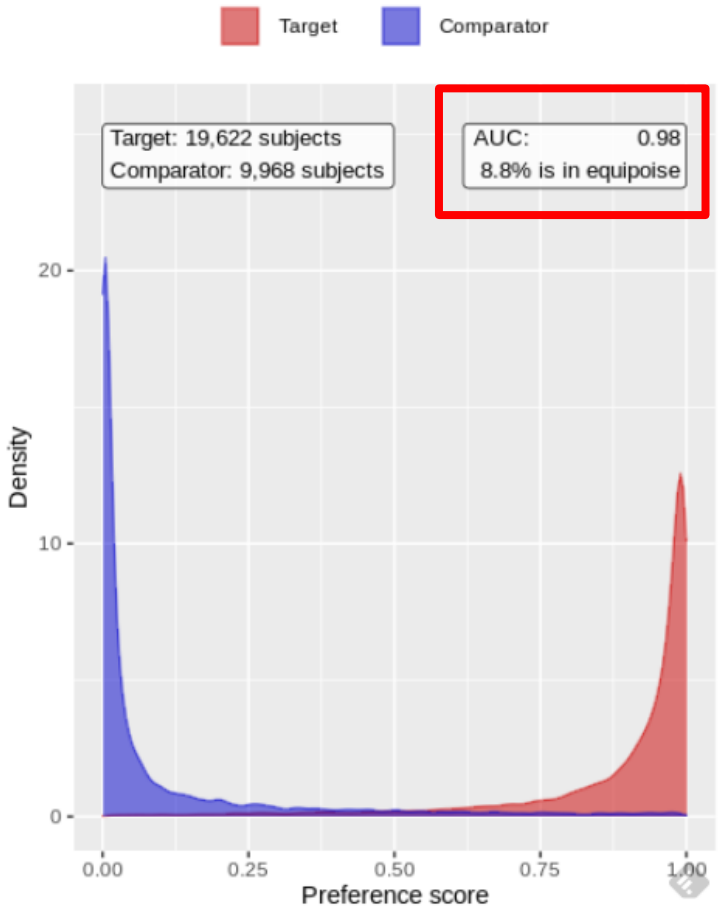
We need to define the exposures and outcomes for our study. One could use an external cohort definition tools, but in this example we do this by writing SQL statements against the OMOP CDM that populate a table of events in which we are interested. The resulting table should have the same structure as the `cohort` table in the CDM. This means it should have the fields `cohort_definition_id`, `cohort_start_date`, `cohort_end_date`,and `subject_id`.

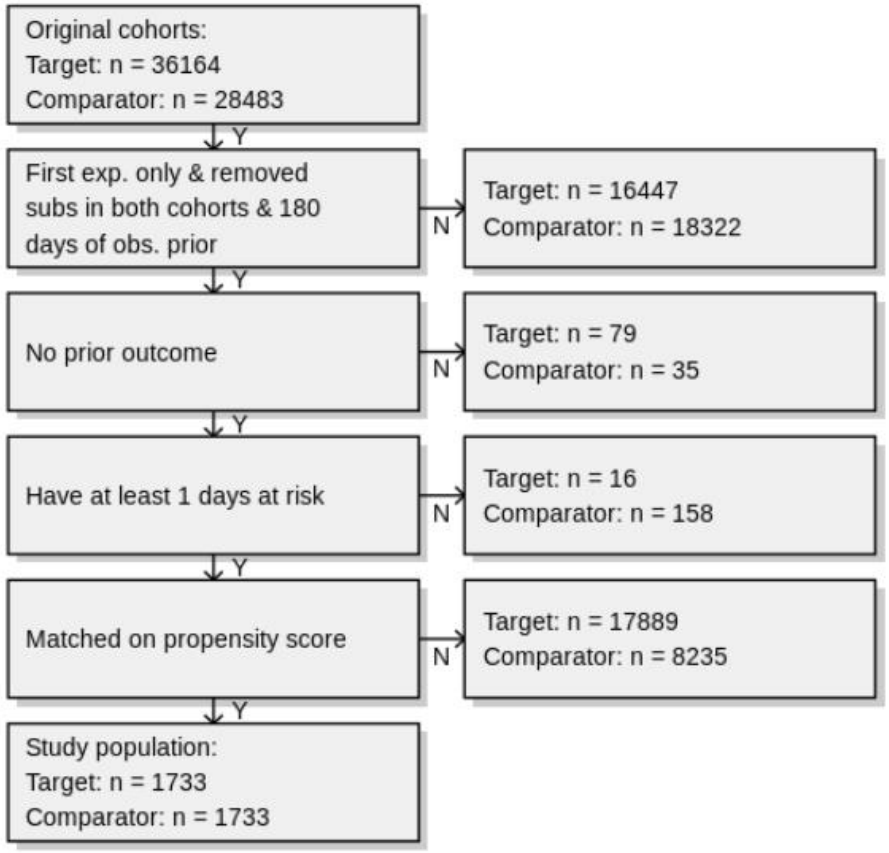
For our example study, we have created a file called *coxibVsNonselVsGiBleed.sql* with the following contents:

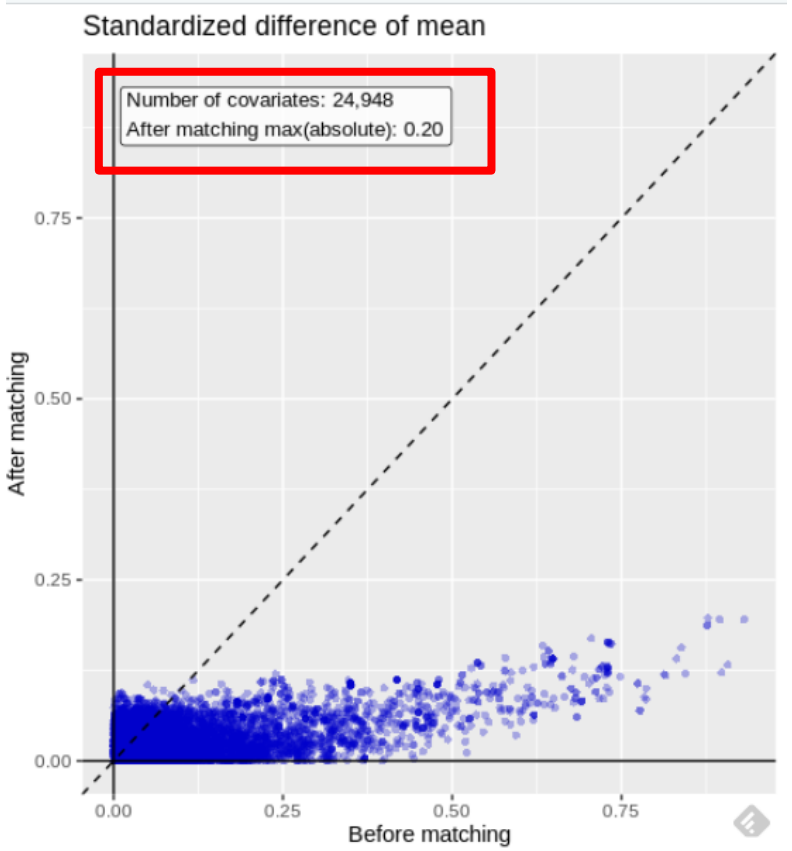
```
/************************************
File coxibVsNonselVsGiBleed.sql
************************************/

IF OBJECT_ID('@resultsDatabaseSchema.coxibVsNonselVsGiBleed', 'U') IS NOT NULL
  DROP TABLE @resultsDatabaseSchema.coxibVsNonselVsGiBleed;
```

Experiment 1) 1:1 Large-scale PS matching using Full population

- Number of study population = 29,590



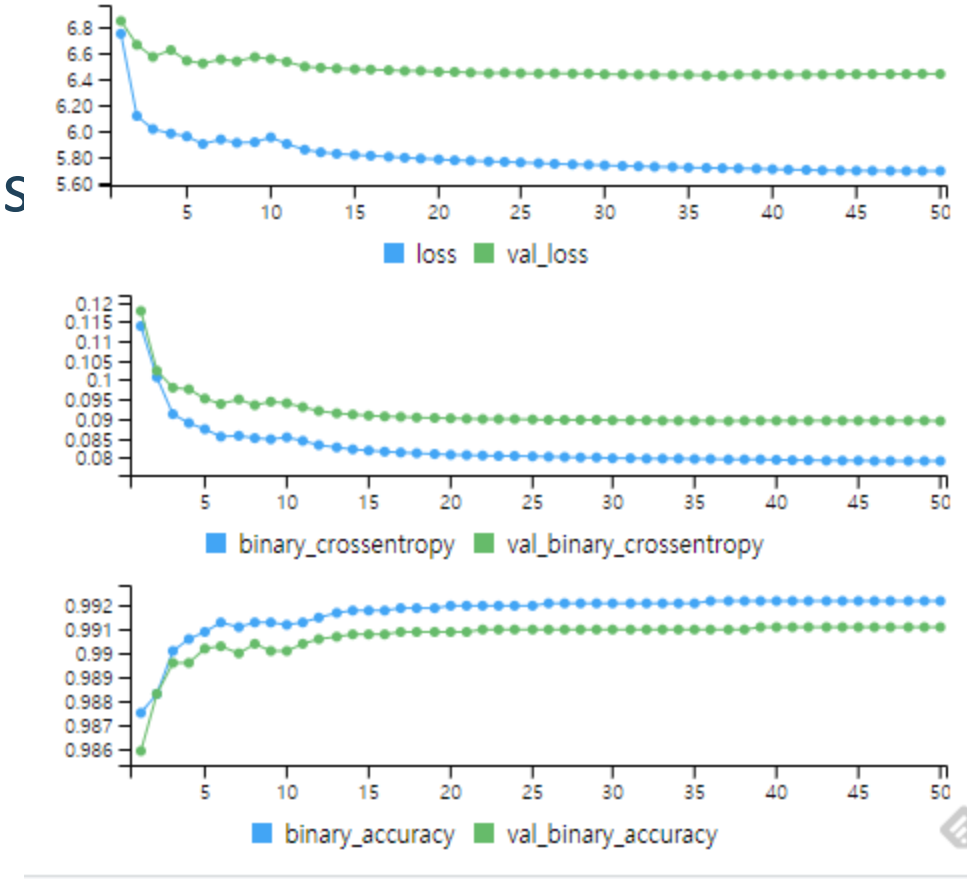




Study population were not in empirical equipoise. Even large-scale PS model with full study population cannot balance the baseline characteristics

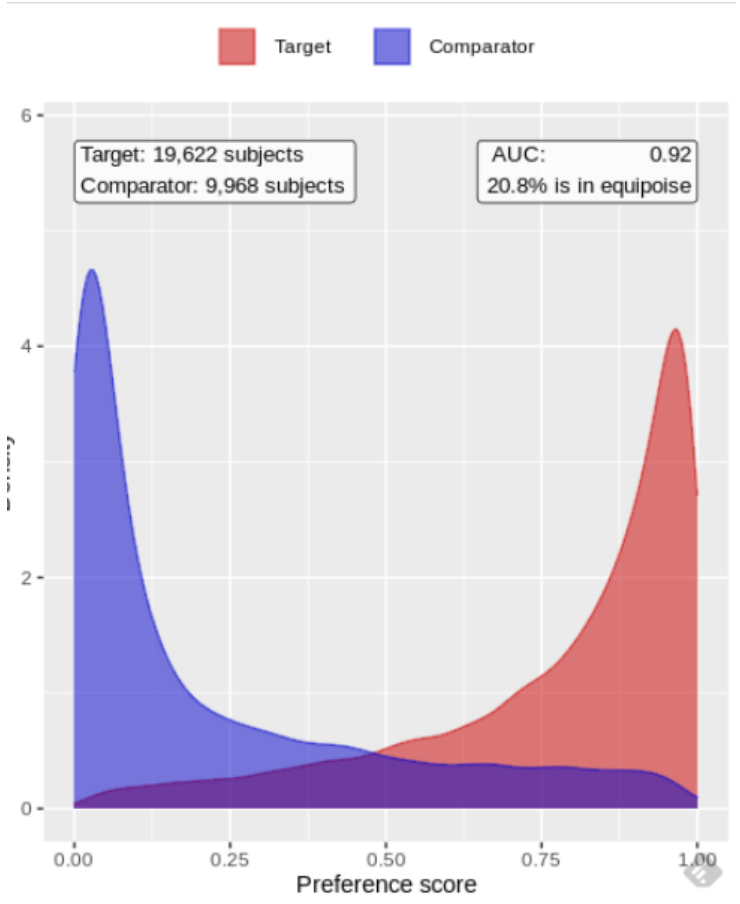
Dimension reduction of covariates using autoencoder

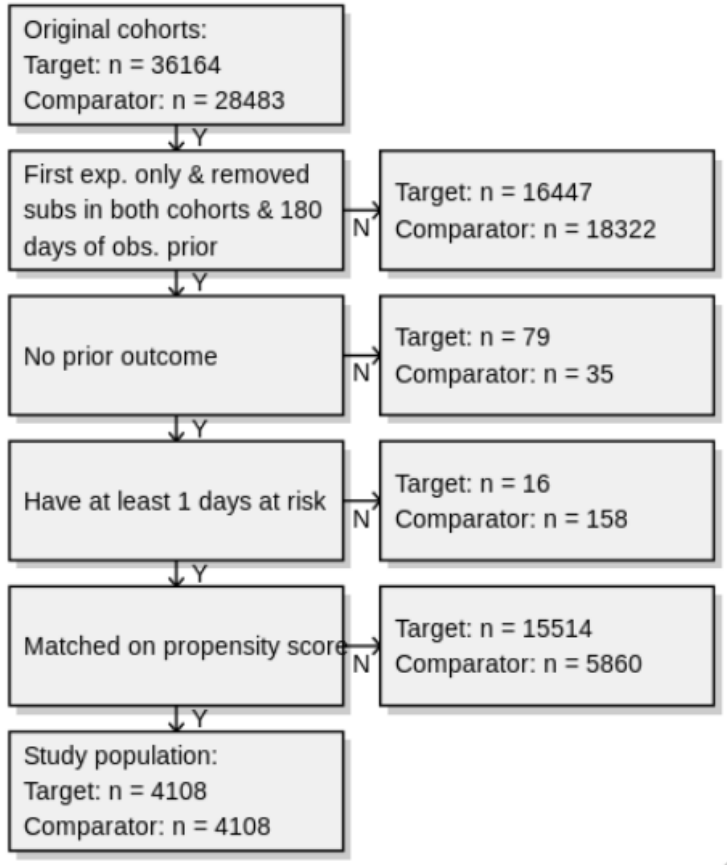
- Simple autoencoder with 1 layer
  - Using L1 regularization (to avoid over-fitting)
  - **Custom loss function** for weighted binary cross entropy
  - Reduce the dimension of covariates from 24,948 to **50**, by using 29,878 population data

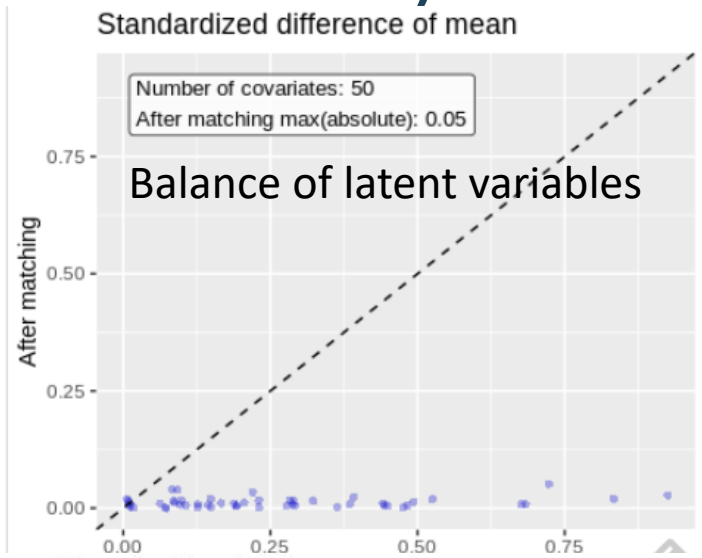


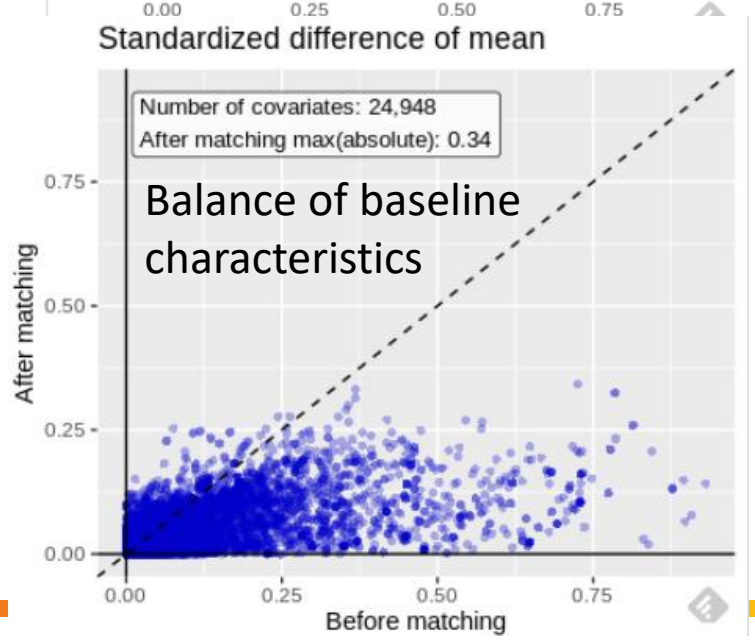
Experiment 1) Representation learning-PS matching using Full population (with autoencoder)

- Number of study population = 29,590





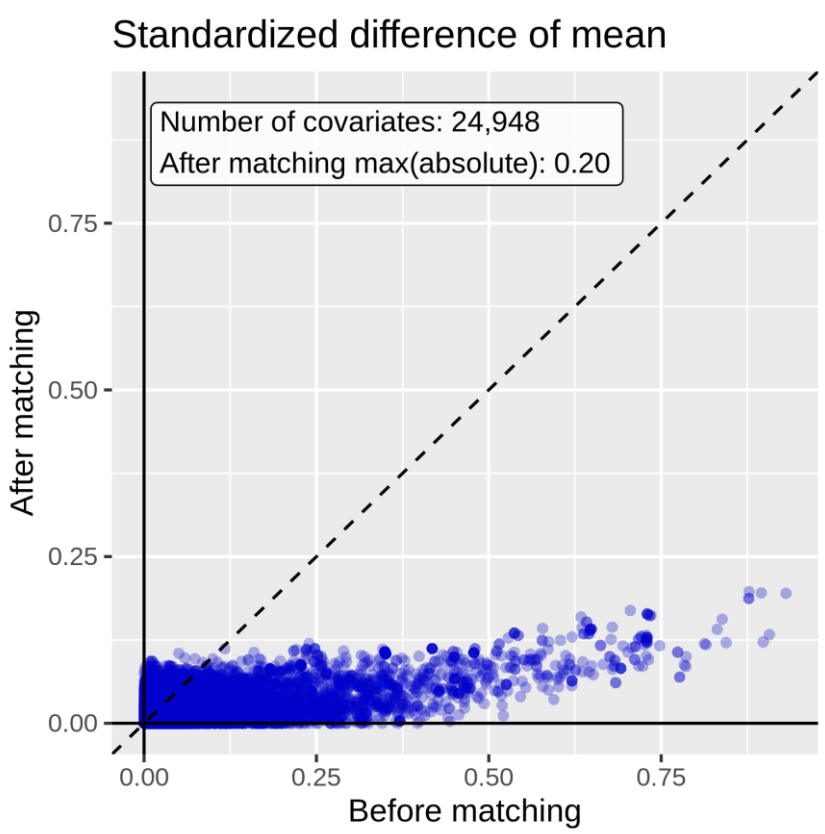




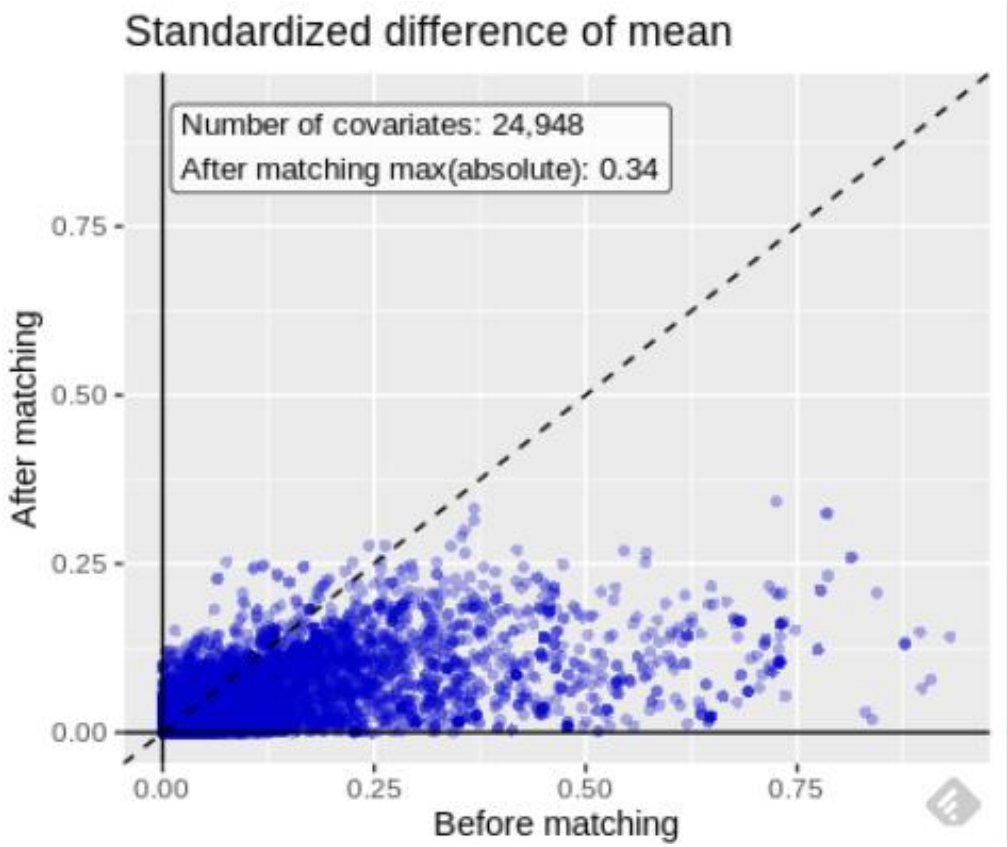
Experiment 1) Large-scale PS matching vs RLPS in full population

- Number of study population = 29,590

Balance in LSPS

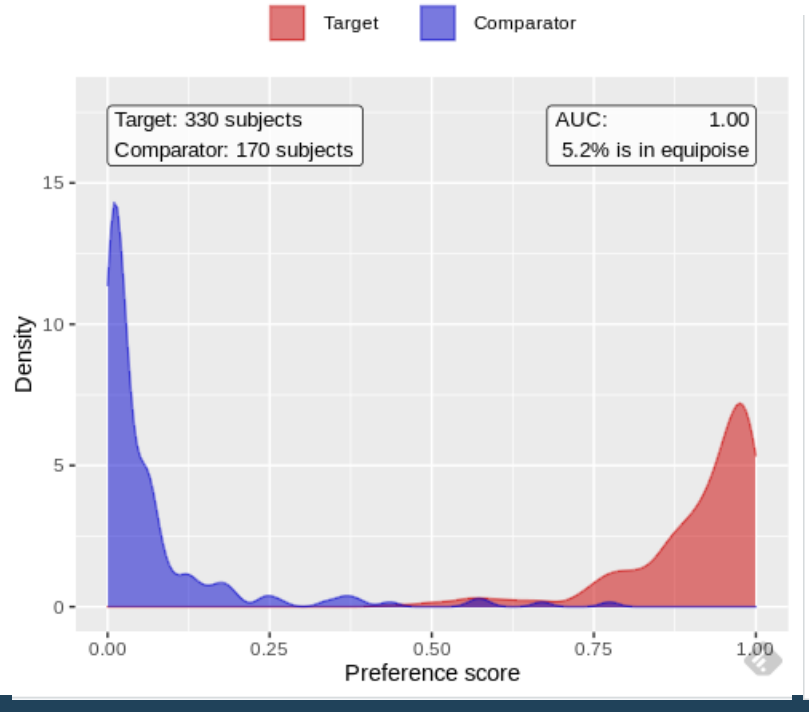


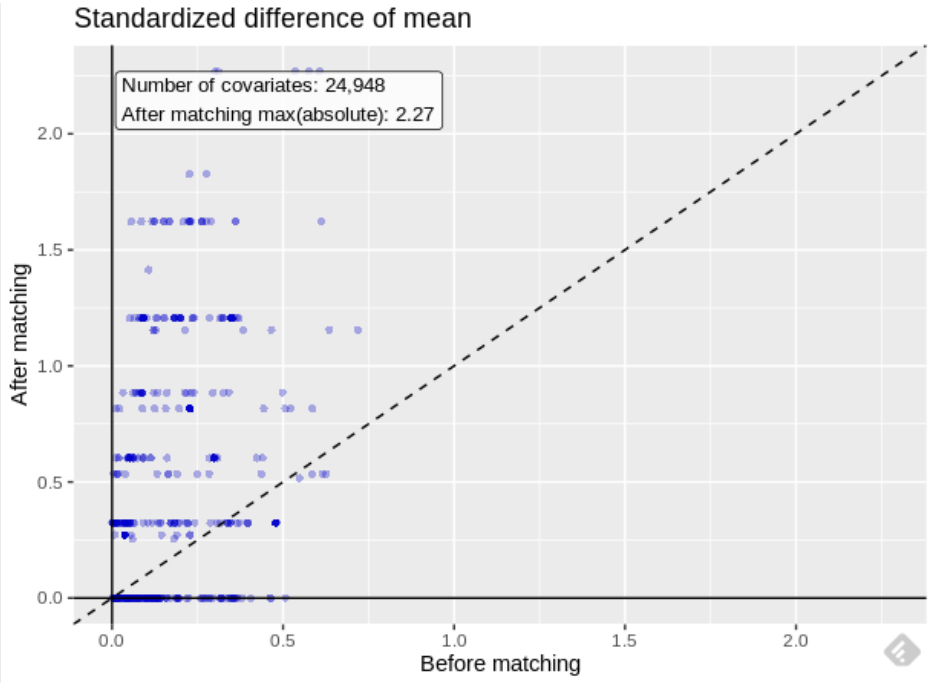
Balance in RLPS

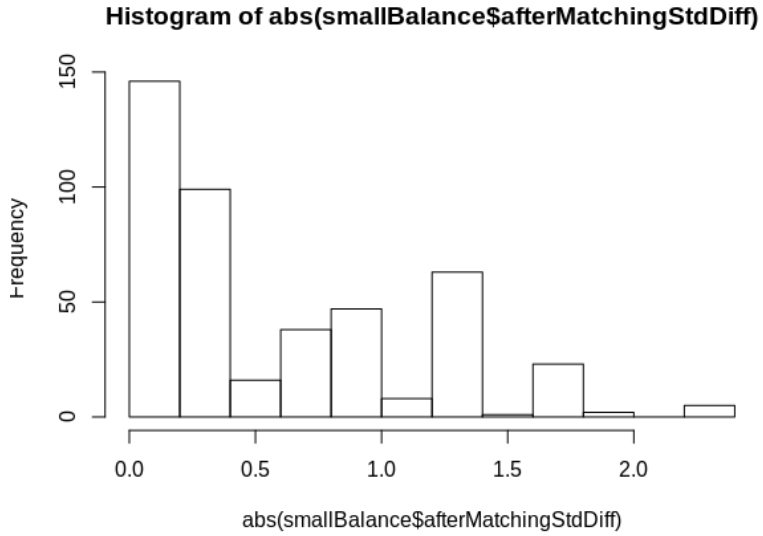


Experiment 1) Large-scale PS matching in small study population

- Number of study population = 500
- Trying large-scale PS matching
  - Number of covariates with Abs Std. Diff <= 0.1 : 146
  - Number of covariates with Abs Std. Diff > 0.1 : 302

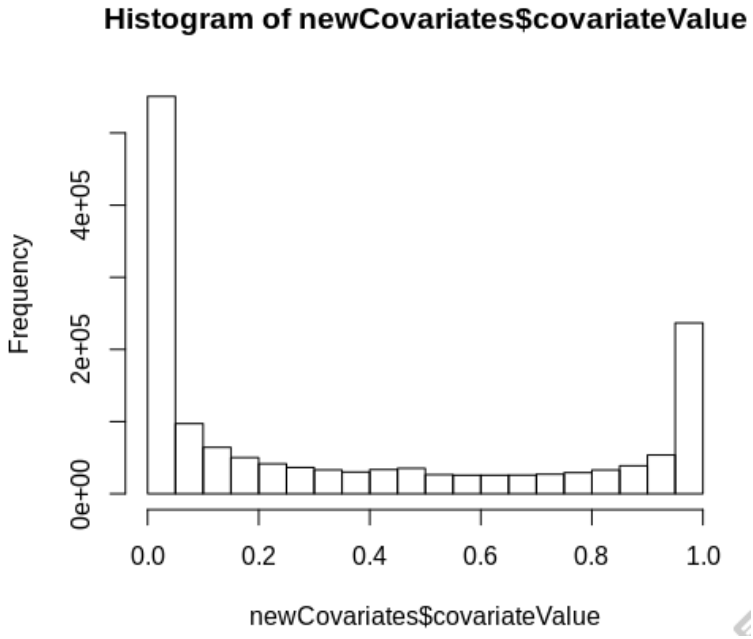


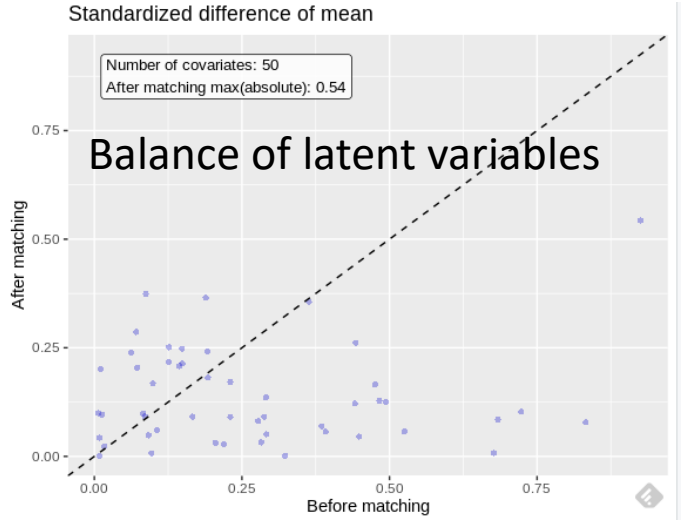


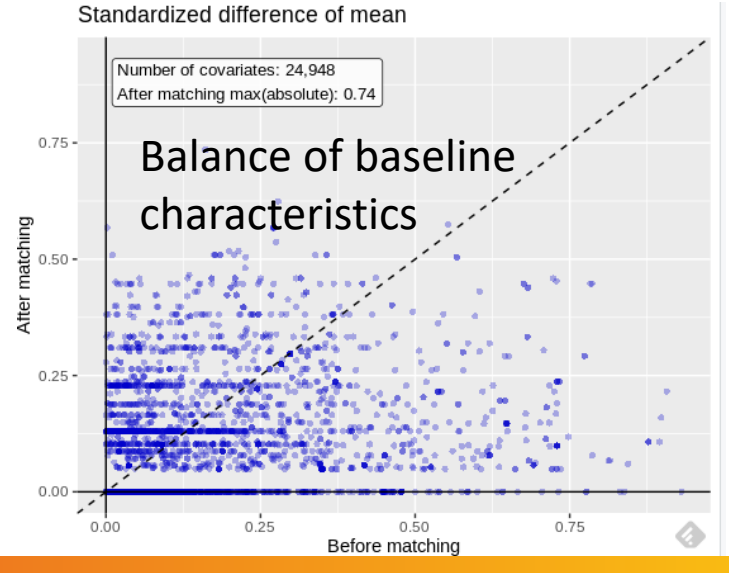


Experiment 1) Representation learning-PS matching in small study population (with autoencoder)

Distribution of latent variables



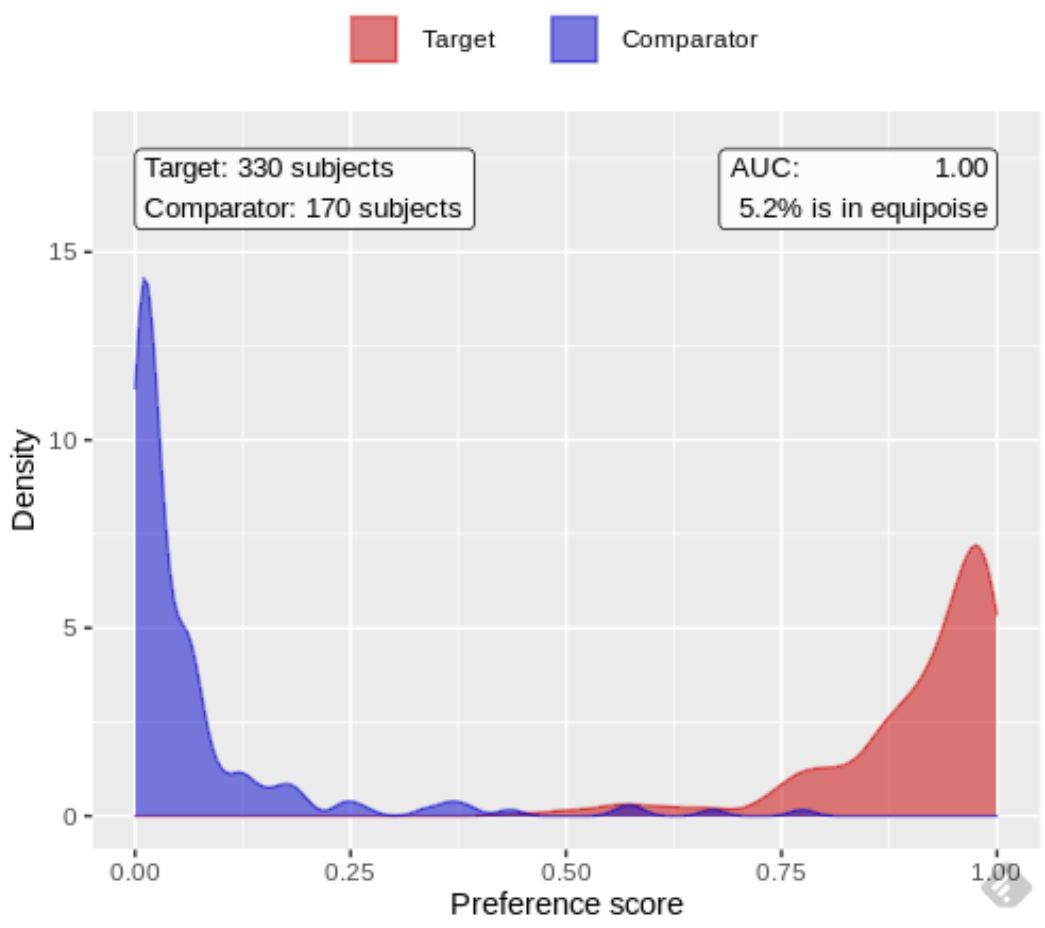




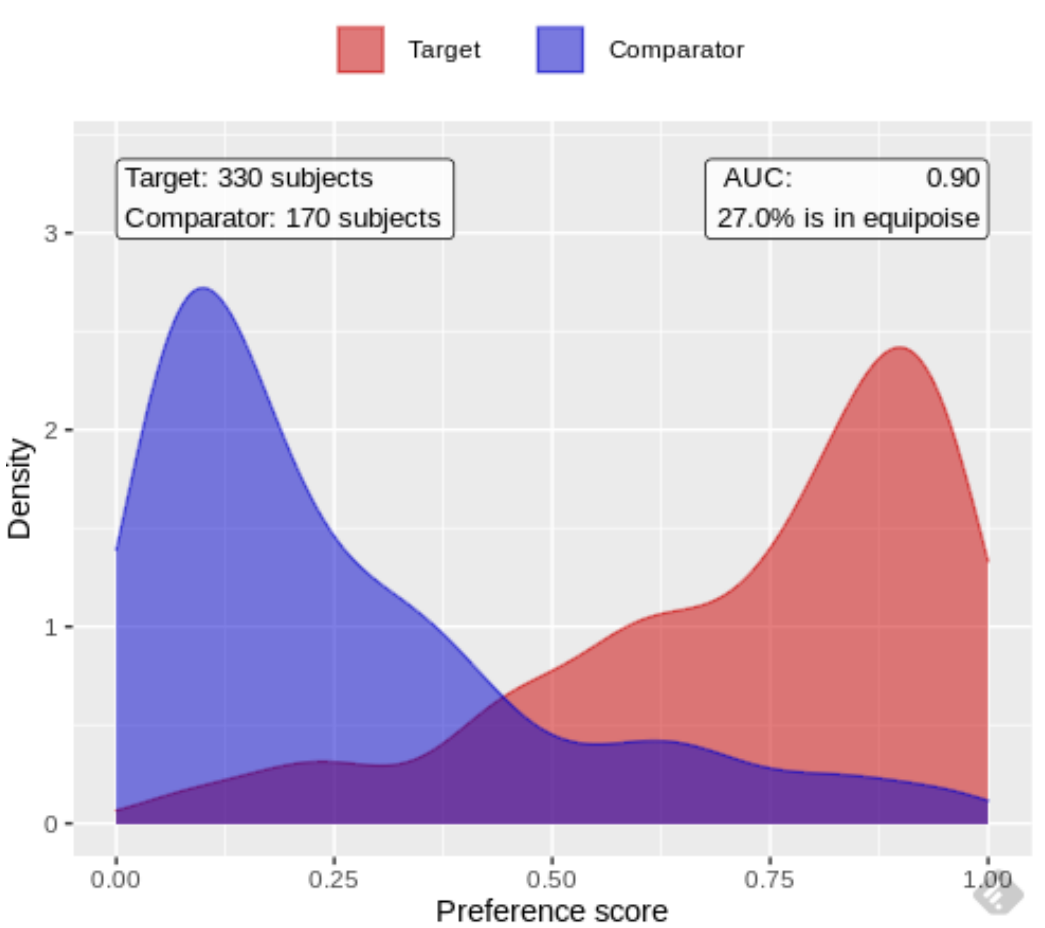
Experiment 1) Large-scale PS matching vs RLPS in small population (n=500)

- Preference score distribution

LSPS



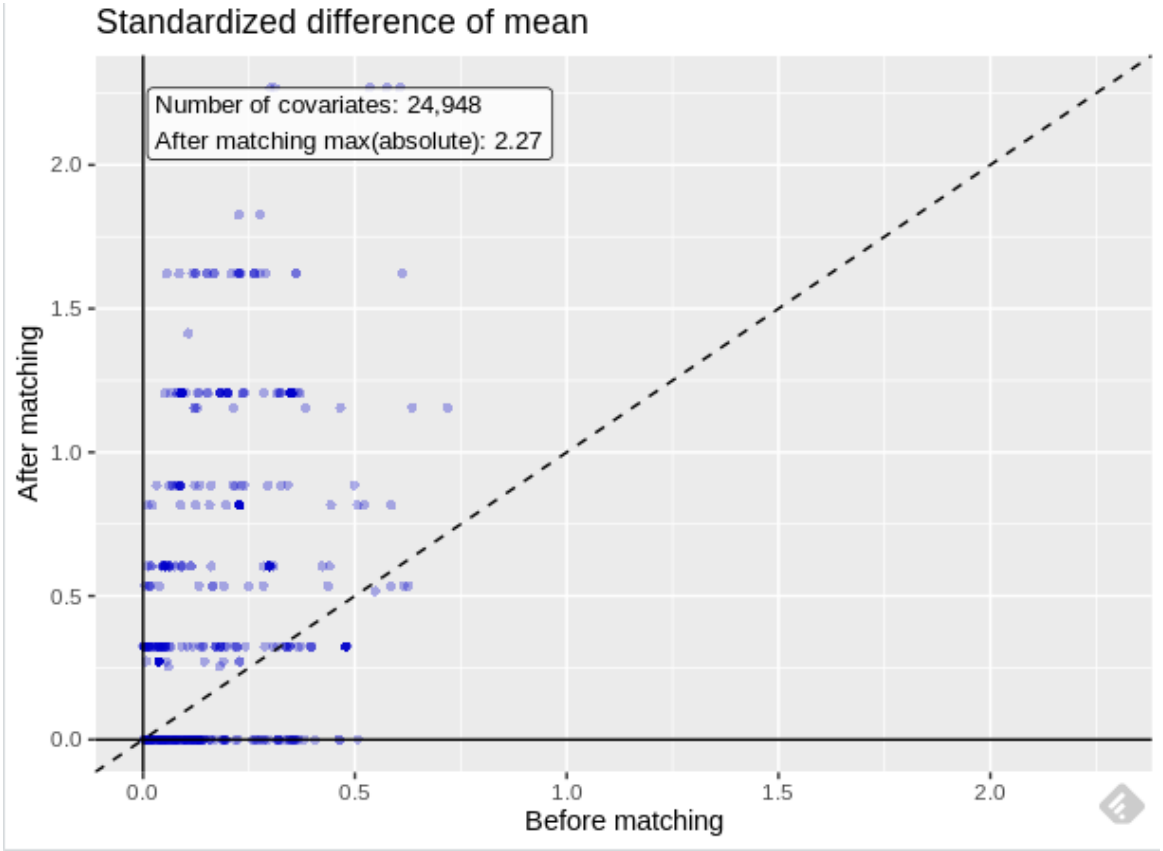
RLPS



Experiment 1) Large-scale PS matching vs RLPS in small population (n=500)

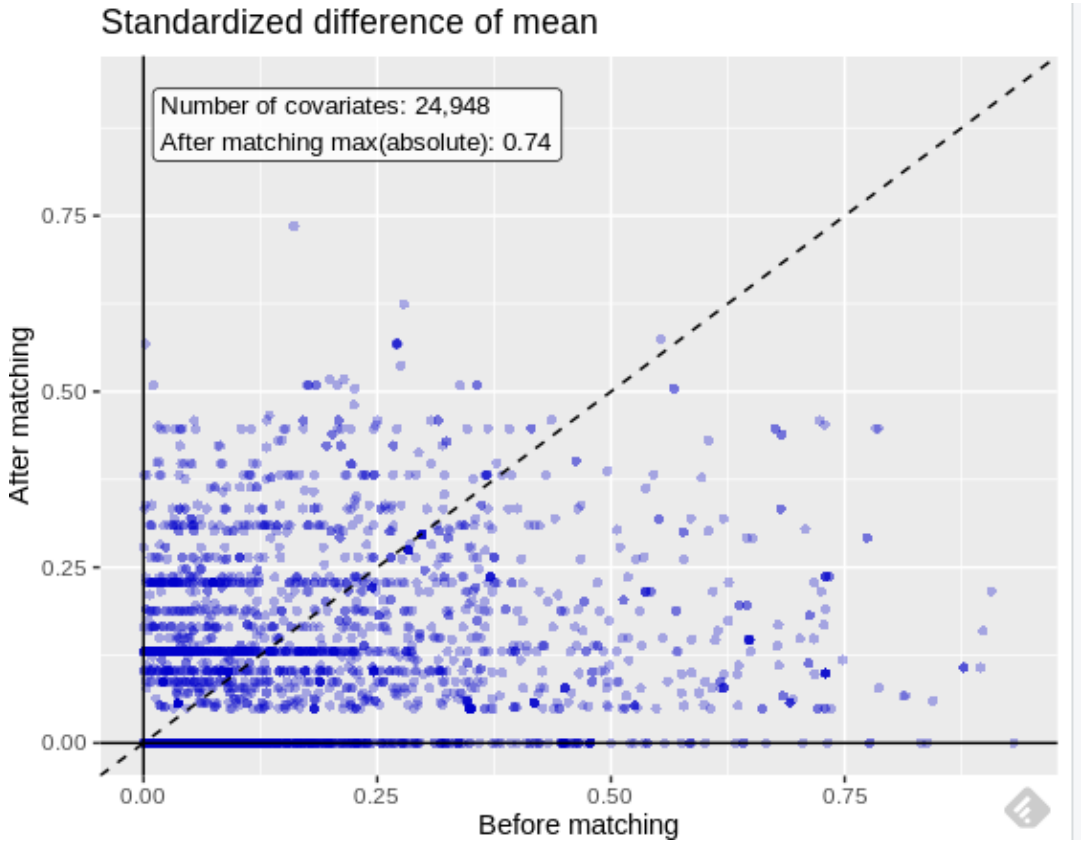
- Balance scatter plot

LSPS



Number of covariates with Abs Std. Diff <= 0.1 : 146
Number of covariates with Abs Std. Diff > 0.1 : 302

RLPS

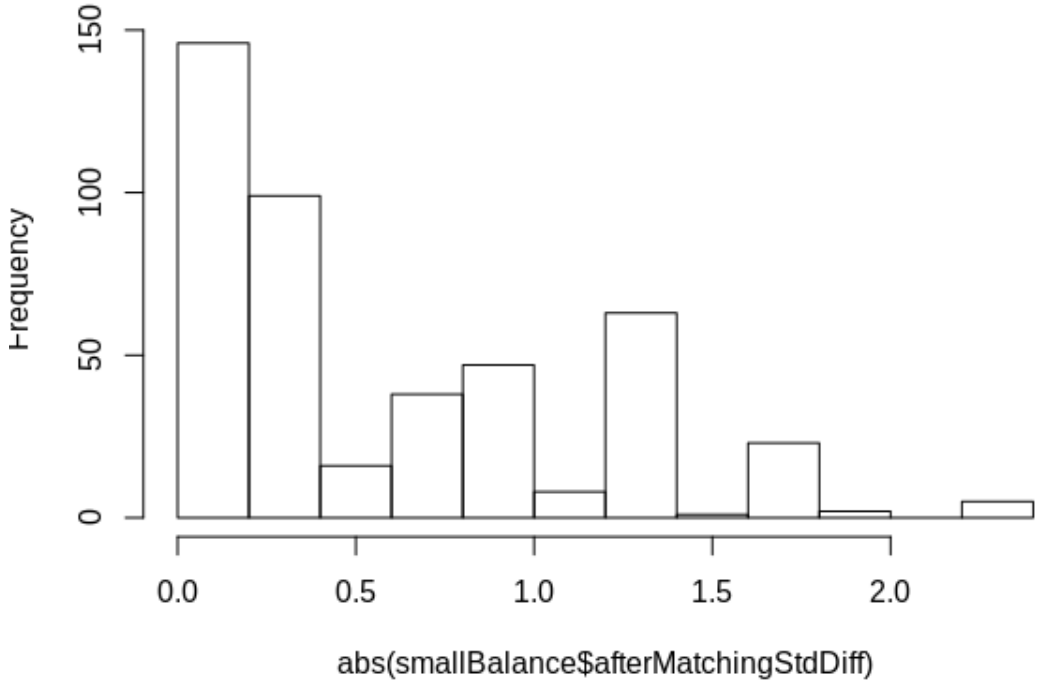


Number of covariates with Abs Std. Diff <= 0.1 : 1772
Number of covariates with Abs Std. Diff > 0.1 : 1855

Experiment 1) Large-scale PS matching vs RLPS in small population (n=500)

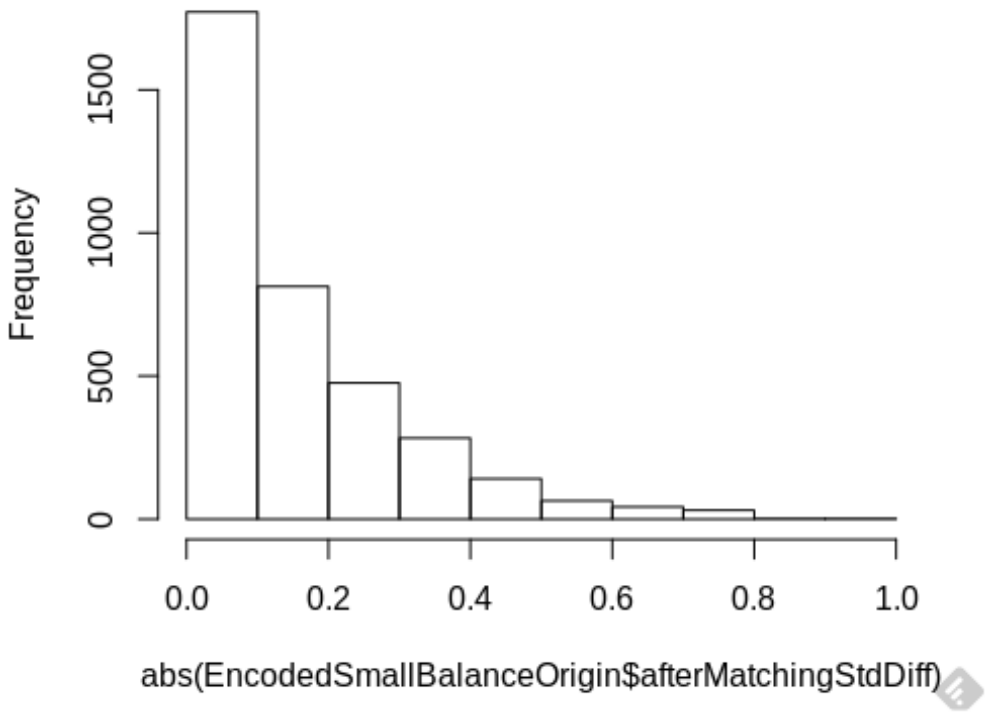
- Distribution of absolute mean difference of covariates after matching

LSPS



Number of covariates with Abs Std. Diff <= 0.1 : 146
Number of covariates with Abs Std. Diff > 0.1 : 302

RLPS



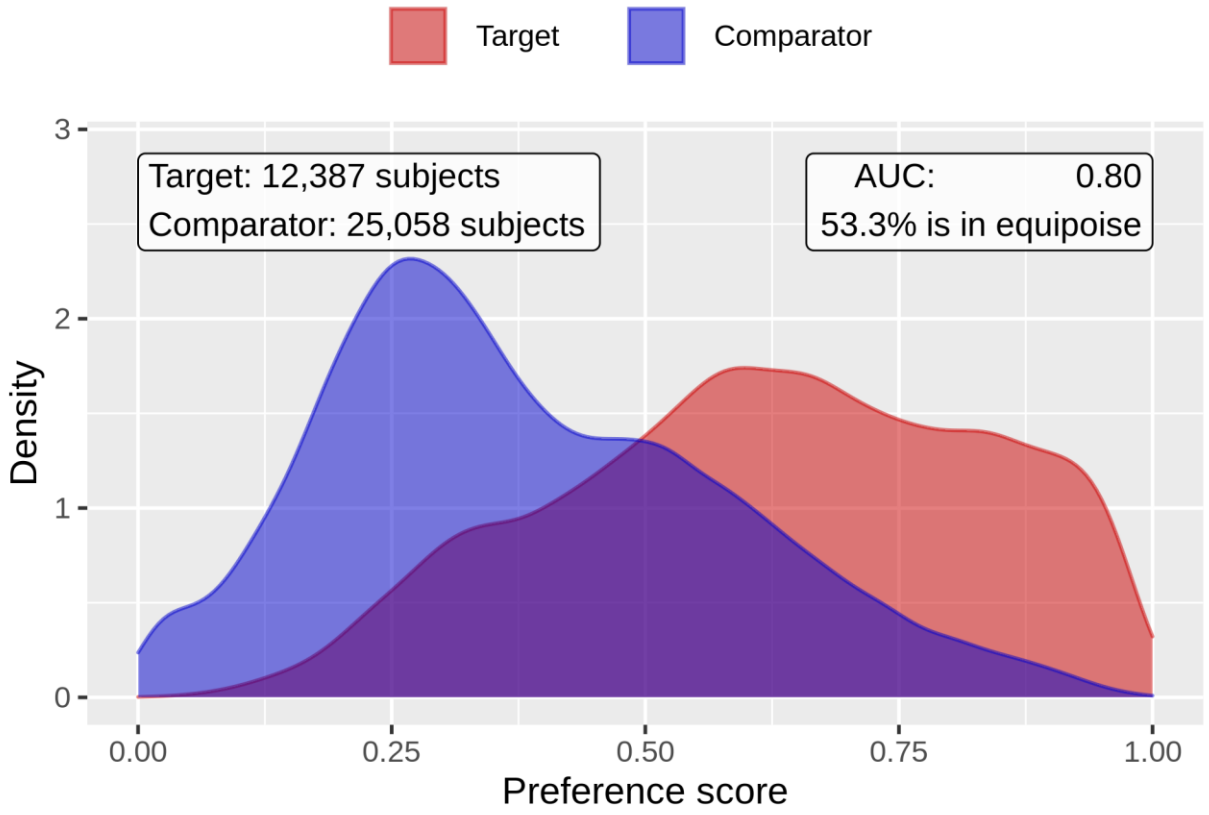
Number of covariates with Abs Std. Diff <= 0.1 : 1772
Number of covariates with Abs Std. Diff > 0.1 : 1855

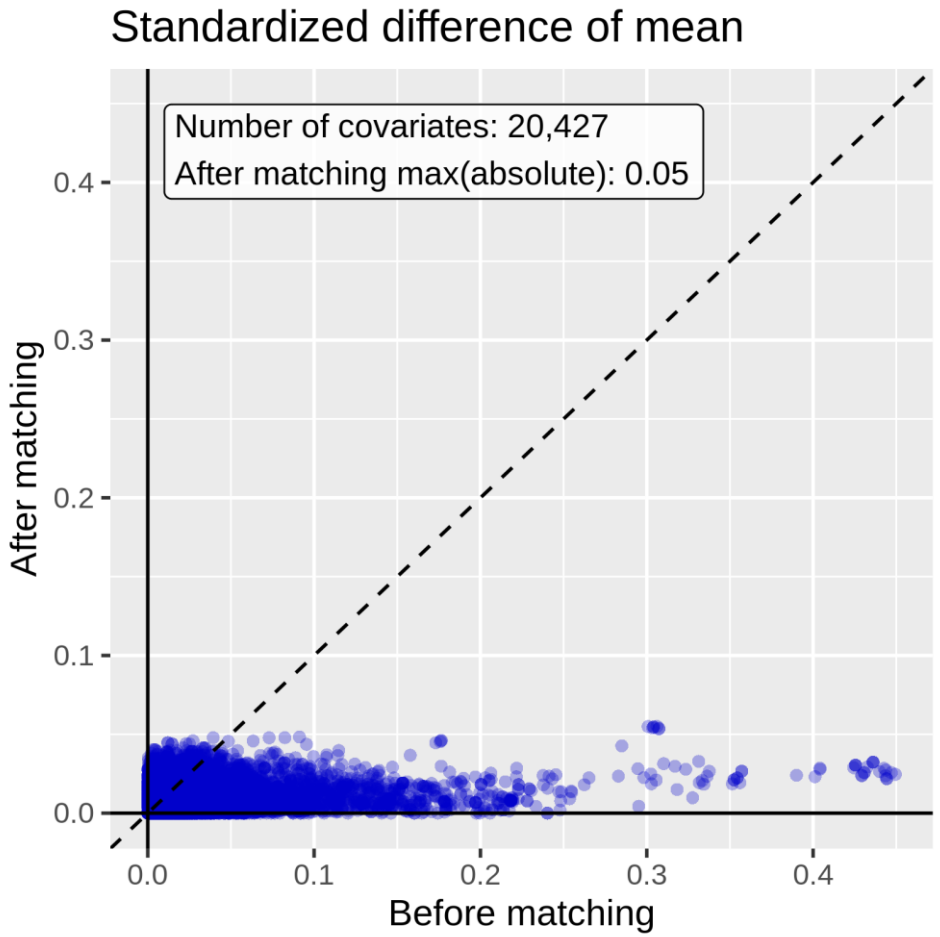
Experiment 2) ARB vs CCB

- According to the Book of OHDSI and LEGEND-HTN
- I generated ARB and CCB user for hypertension from claim database

Experiment 2) 1:1 Large-scale PS matching using Full population

- Number of study population = 37,445

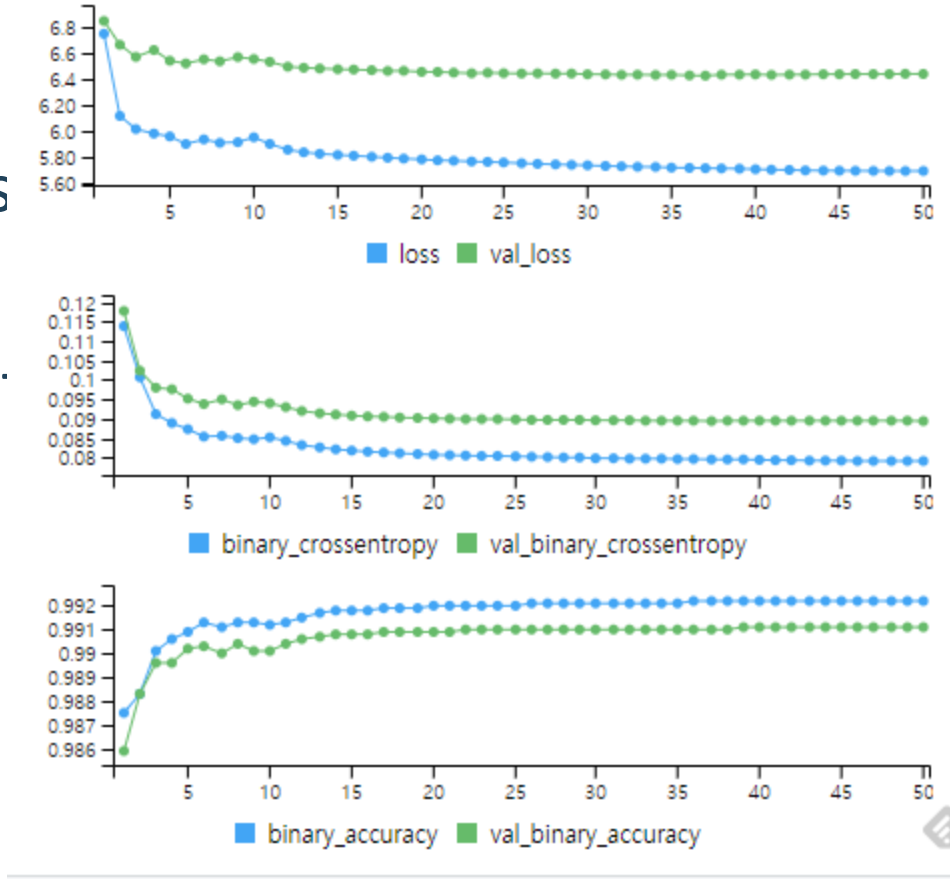




Study population were in empirical equipoise. Large-scale PS model with full study population balanced the baseline characteristics between the groups well

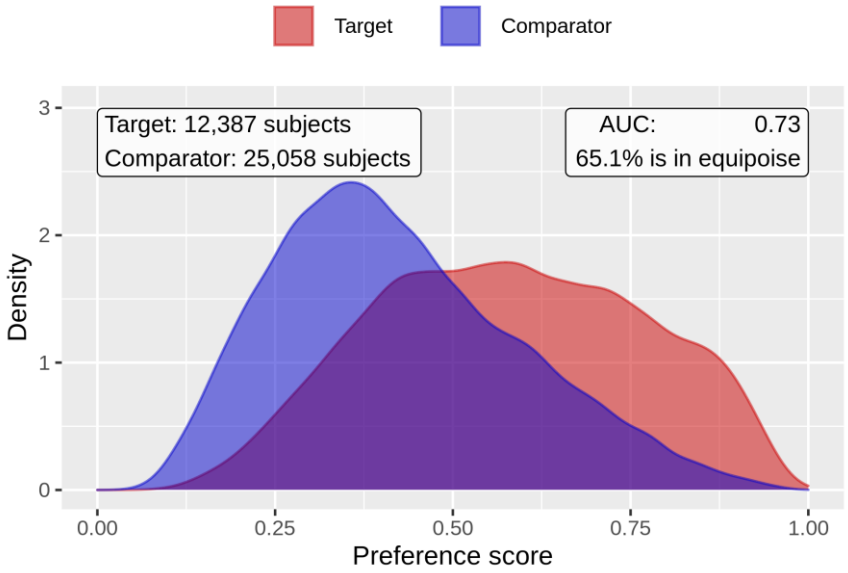
Dimension reduction of covariates using autoencoder

- Simple autoencoder with 1 layer
  - Using L1 regularization (to avoid over-fitting)
  - **Custom loss function** for weighted binary cross entropy
  - Reduce the dimension of covariates from about 10,000 (after tidying covariates) to **100,** by using 37,445 population data

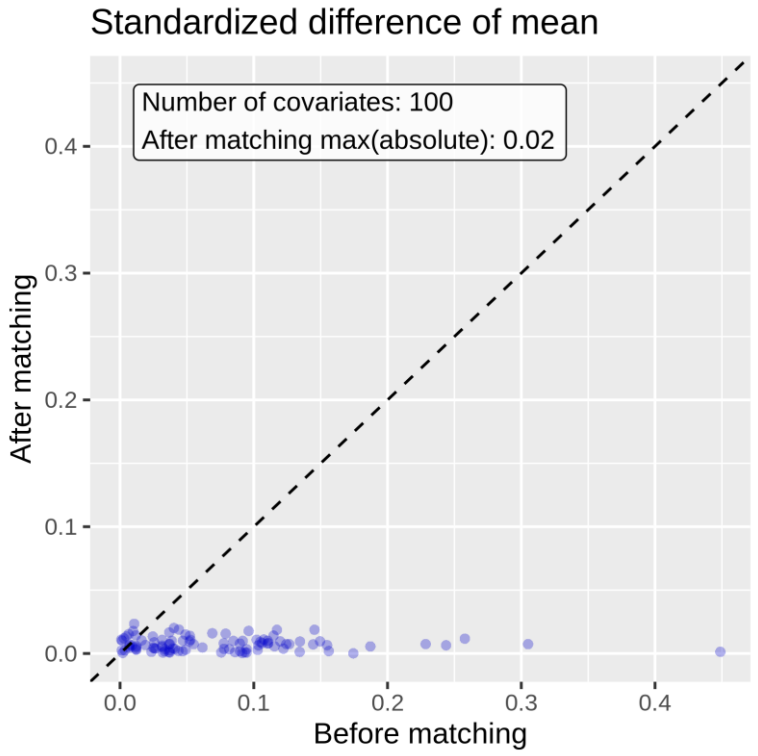


Experiment 2) Representation learning-PS matching using Full population (with autoencoder)

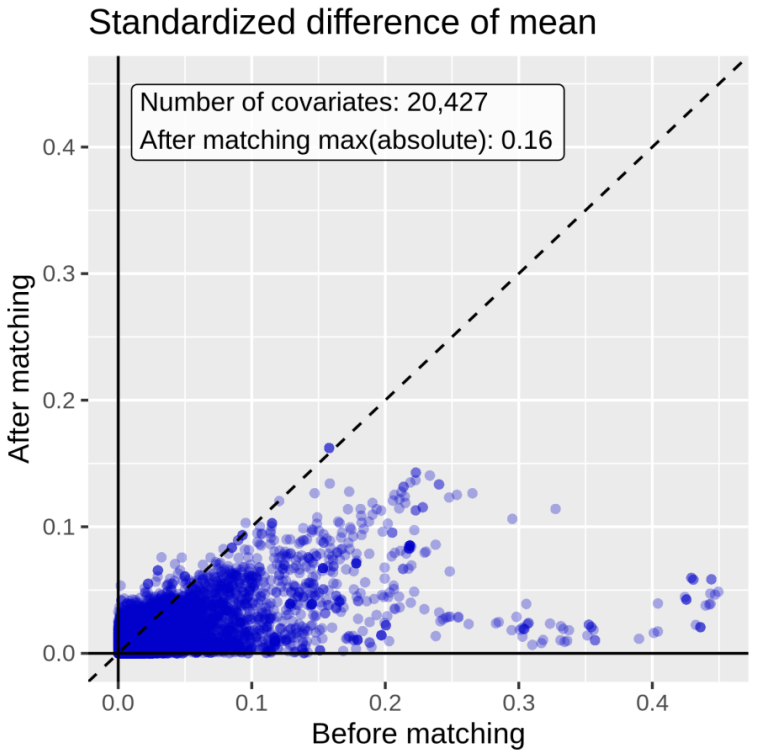
- Number of study population = 37,445



Balance of latent variables



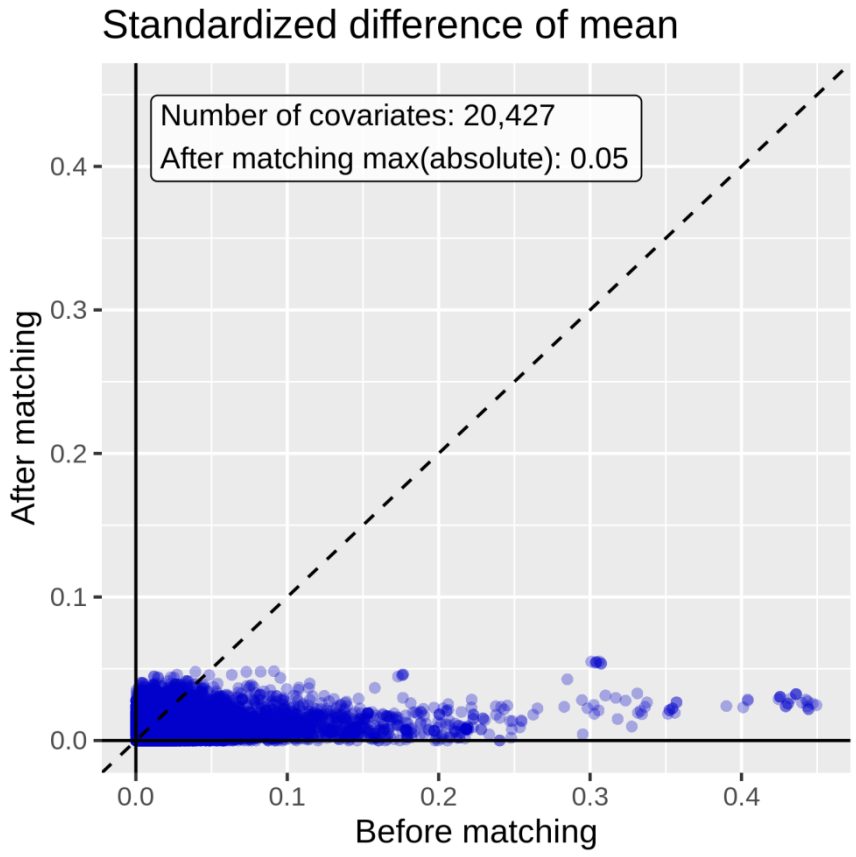
Balance of baseline characteristics



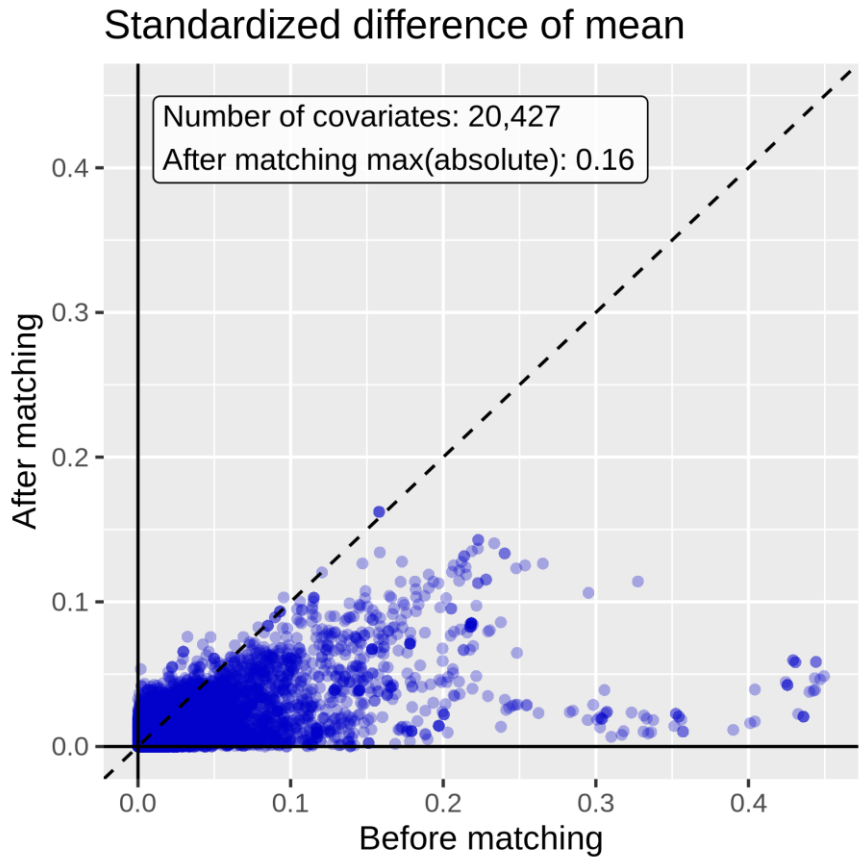
Experiment 2) Large-scale PS matching vs RLPS in full population

- Number of study population = 37,445

Balance in LSPS

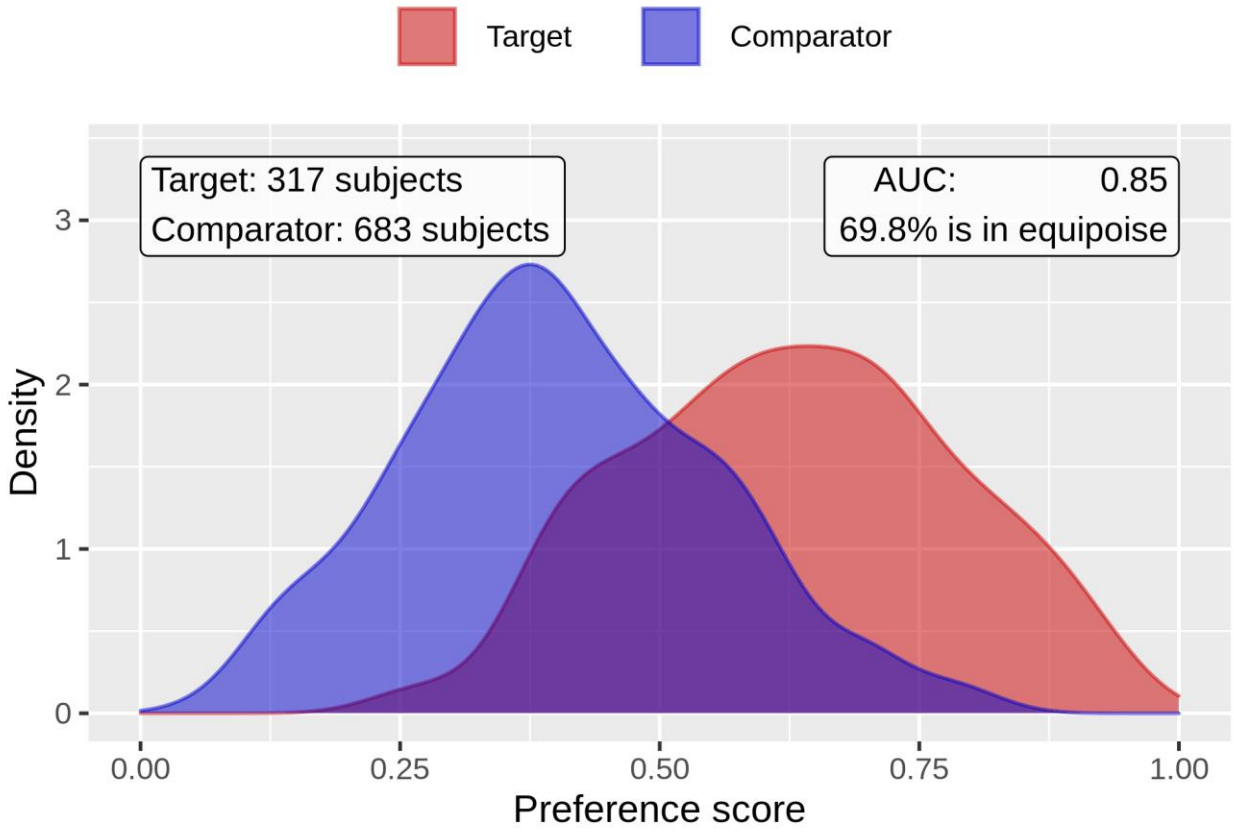


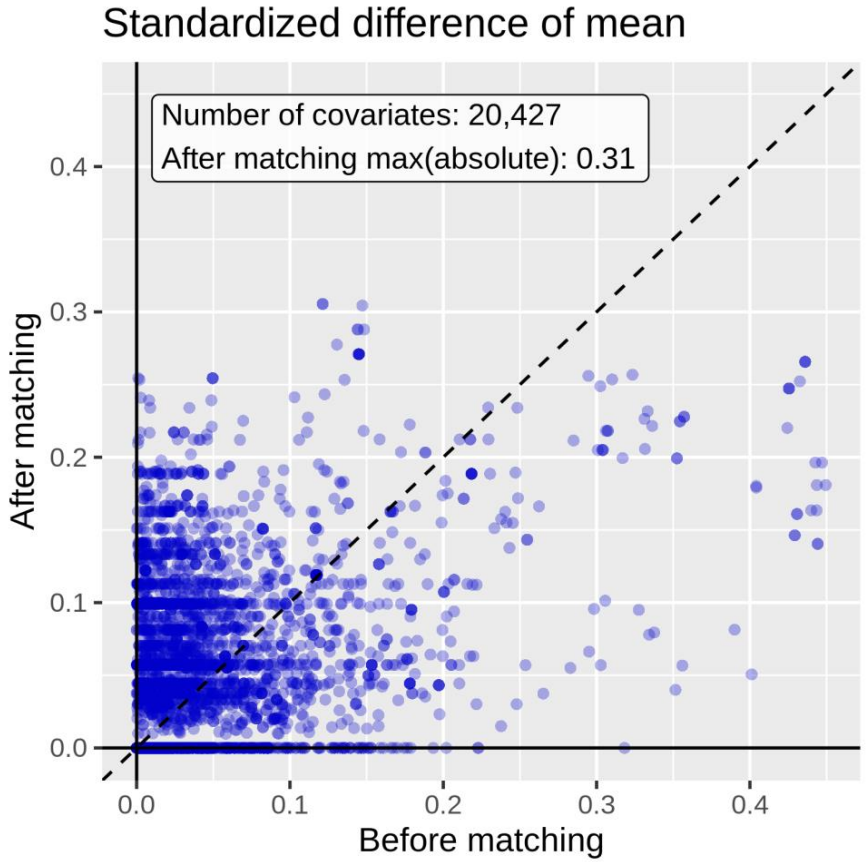
Balance in RLPS



Experiment 2) Large-scale PS matching in small study population

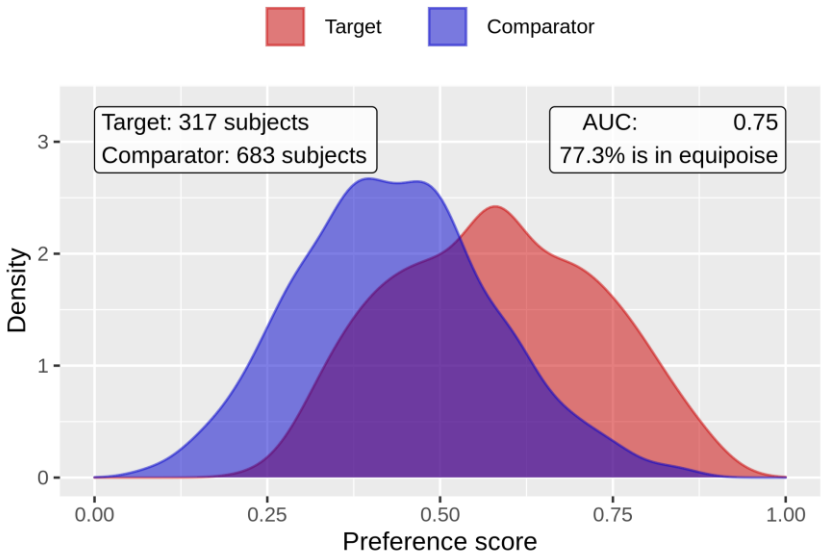
- Number of study population = 1000



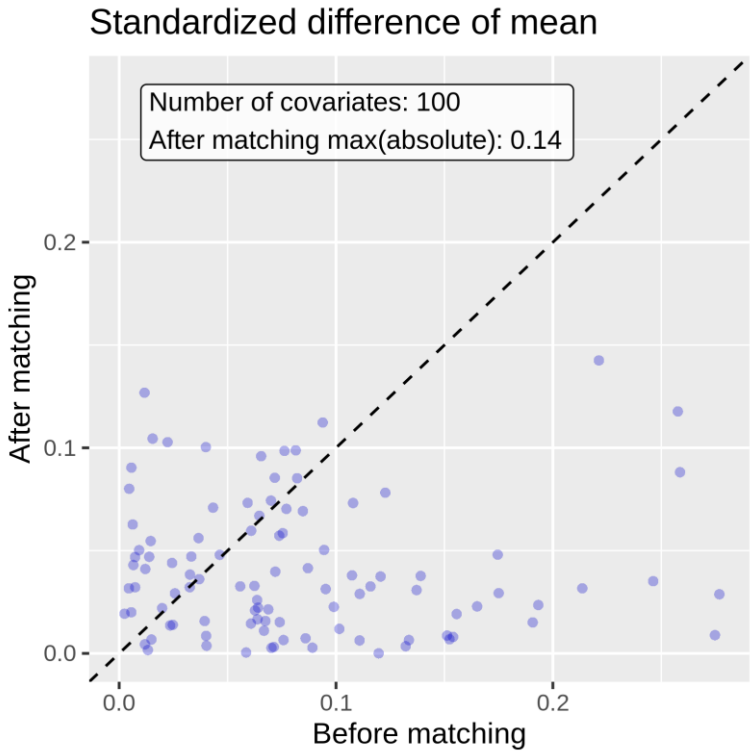


Experiment 2) Representation learning-PS matching in small study population (with autoencoder)

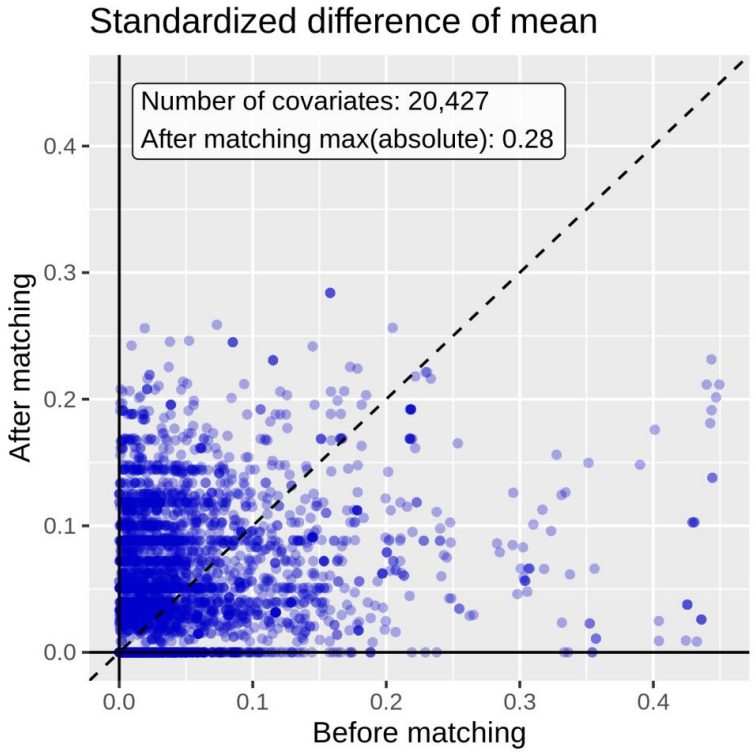
- Number of study population = 1000



Balance of latent variables



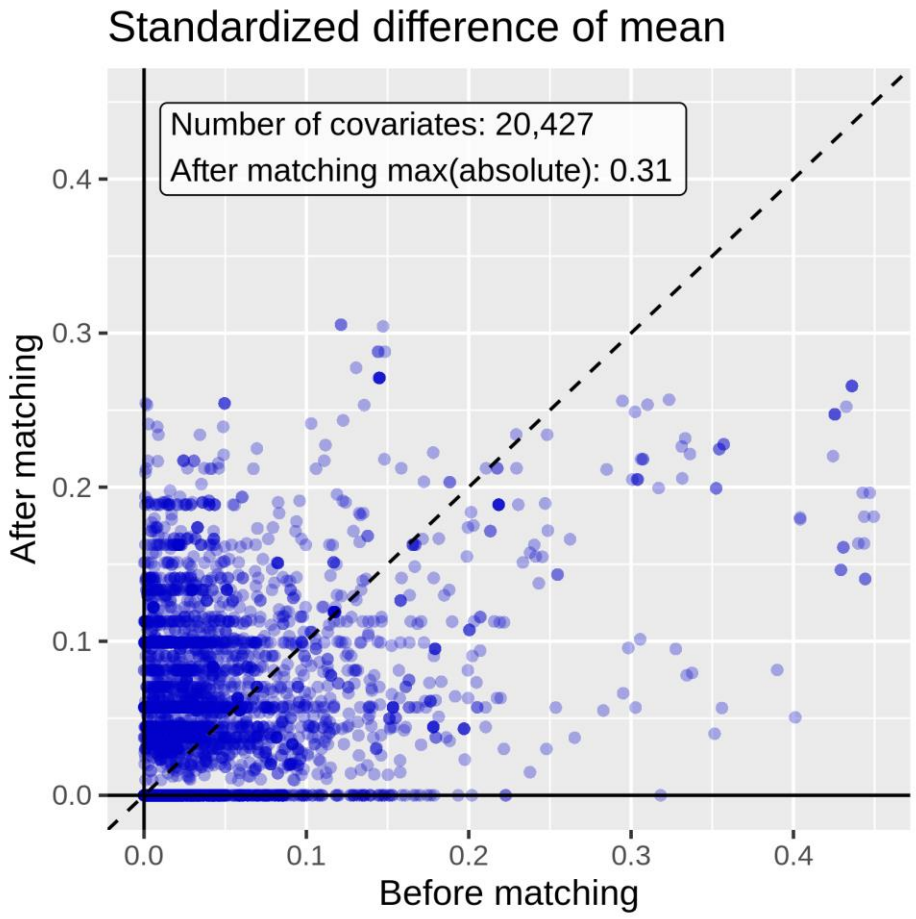
Balance of baseline characteristics



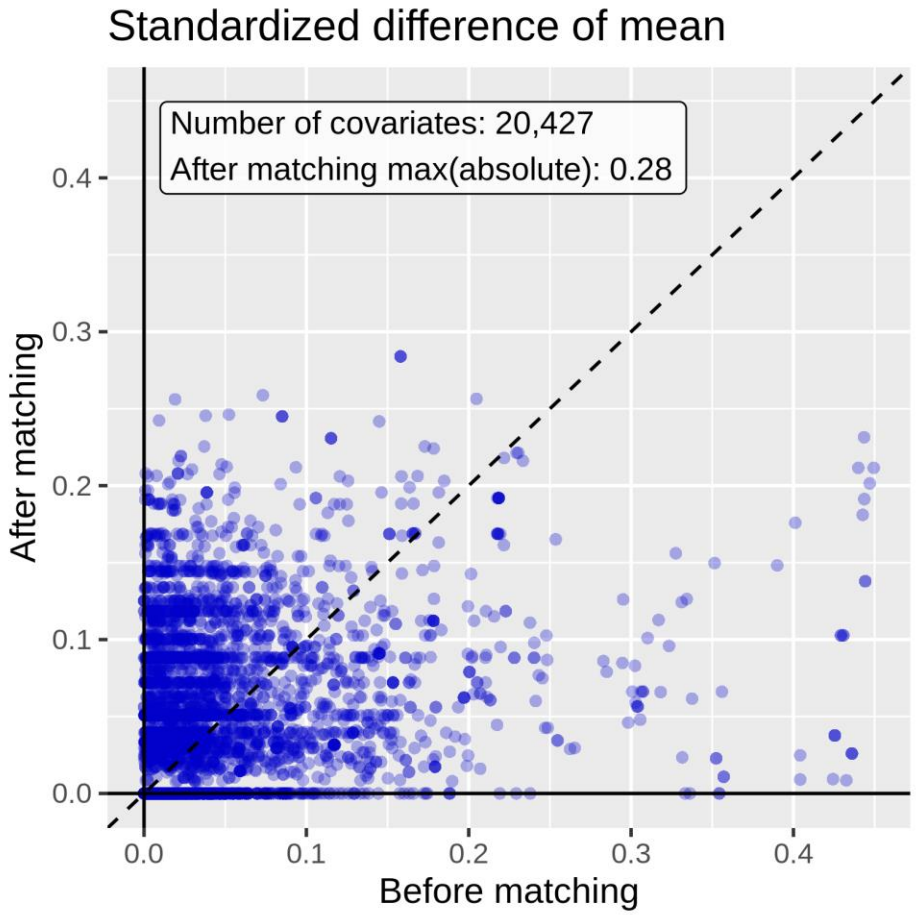
Experiment 2) Large-scale PS matching vs RLPS in small population (n=1000)

- Balance scatter plot

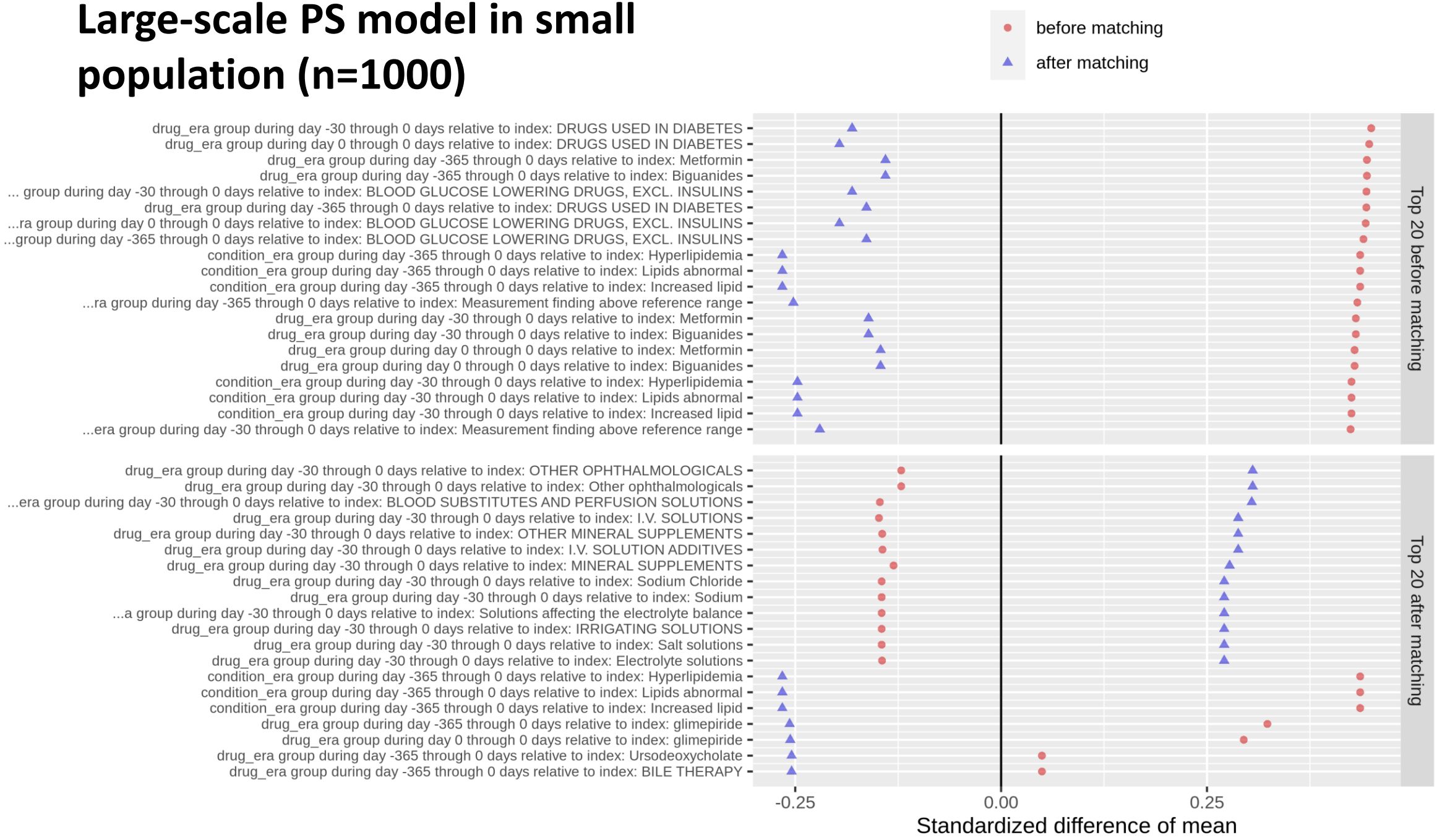
LSPS



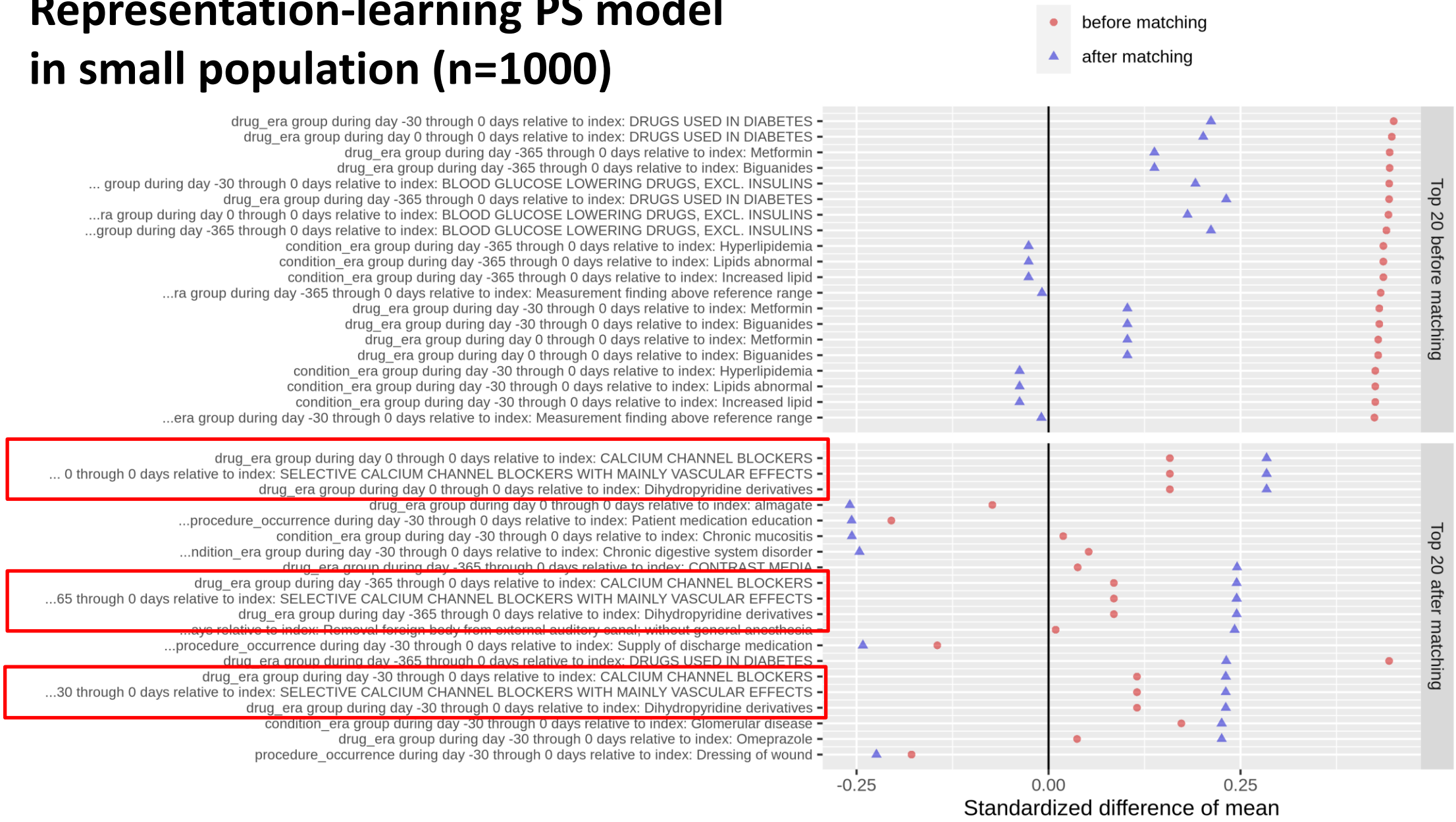
RLPS



Large-scale PS model in small population (n=1000)



Representation-learning PS model in small population (n=1000)



Implications

- Representation Learning + large-scale PS model might be more robust than large-scale PS model in small study population
- It was not easy to develop highly-efficient auto-encoder using historical population
- Current PS model need to exclude the treatment variables, but I think we do not need to exclude the treatment variables to train auto-encoder.
- The improvement of performance in auto-encoder may lead to increase overall robustness of this method

Future plan

- Empirical evaluation of Representation-learning PS model
  - 1. Evaluation using negative-controls in LEGEND-HTN
  - 2. Evaluation using OHDSI Benchmark framework
  - I am modifying the CohortMethod package to support multiple analyses using encoders, now.
- [Grandiose plan] Developing OHDSI universal encoder
  - Google developed universal language representation model (BERT) using 3.3B corpus
  - Recent OHDSI's progress, concept prevalence study and implementation of Andromeda, enables to build large vector space to cover available concepts across the network
  - Once we developed universal encoder, we can fine-tune this encoder for cohorts of interest, and then apply it to any kind of studies we do (including PLP)

Flight of ideas

- Challenge
  - high-dimensional and very sparse covariates
  - There is no single / simple standards making clinician choice for a treatment (Propensity).
    - Sometimes history of hypertension / diabetes; sometimes renal function; sometimes hepatic function; sometimes patients' economic status

- Universal autoencoder
  - Let's treat estimation problem as down-stream task with transfer learning
  - What kind of model? Can we apply BERT for this? (instead of masking the word, let the model predict the novel treatment)
  - What should the model predict?
  - More fundamentally, can we feed this data to the computer through abstraction?

OHDSI is an ocean of observational health care data across the world

- OHDSI is composed of myriad of small-to-big health care databases across the world
- Every database joins OHDSI after a long journey just like the way a river joins the sea

Current challenge in OHDSI

- Our best practice recommends to use large-scale propensity score matching for new-user population-level estimation
- It is hardly possible for data partners with small-to-medium sized database to join OHDSI network studies
- This challenge becomes so apparent for COVID-19 research

METIS: Methods to Enable Transferring Information across OHDSI

- Greek word *metis* meant a quality that combined wisdom and cunning, Odysseus being the embodiment of it.
- In myth, METIS is one of 6,000 Oceanids (river), the daughters of Oceanus (ocean) and Tethys, which implies **circulation** or **samsara 輪廻** of water.
- METIS is the first wife of Zeus and the mother of Athena. She empowers Zeus to think wise and deep with discipline after being engulfed by him.

Les Oceanides Les Naiades de la mer. Gustave Doré, 1860s

https://www.wattpad.com/86490100-the-tale-of-athena%27s-birth

METIS enables us to overcome current challenge

- METIS (Methods to Enable Transferring Information across OHDSI) may let small-to-medium sized database join the OHDSI network studies and let us analyze effects of emerging treatments

Mission, Vision, and Values of OHDSI

- Our Mission

To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

- Our Vision

A world in which observational research produces a comprehensive understanding of health and disease.