MS# EDE19-0473

Deep learning-based propensity scores for confounding control in comparative effectiveness research: A large-scale, real-world data study

Authors: Janick Weberpals[a], Tim Becker[b], Jessica Davies[c], Fabian Schmich[a], Dominik Rüttinger[d], Fabian J. Theis[e,f*], Anna Bauer-Mehren[a*]

[a] Data Science, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich (RICM), Penzberg, Germany

[b] xValue GmbH, Willich, Germany, on behalf of Data Science IV, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich (RICM), Penzberg, Germany

[c] F. Hoffmann-La Roche Ltd, Welwyn Garden City, United Kingdom

[d] Early Clinical Development Oncology, Pharmaceutical Research and Early Development (pRED), Roche Innovation Center Munich (RICM), Penzberg, Germany

[e] Institute of Computational Biology, German Research Center for Environmental Health, Helmholtz Center Munich, Neuherberg, Germany

[f] Department of Mathematics, Technical University of Munich, Garching, Germany

Address for Correspondence:
* Dr. Anna Bauer-Mehren
Pharmaceutical Research and Early Development informatics (pREDi)
Roche Diagnostics GmbH
PXID....2246
Nonnenwald 2, 82377 Penzberg / Germany
Phone: +4988566010700
Mail: anna.bauer-mehren@roche.com

* Prof. Dr. Dr. Fabian J. Theis

Helmholtz Zentrum Munich
Institute of Computational Biology
Ingolstädter Landstr. 1
85764 Neuherberg, Germany
Phone: +49 89 3187-4030
Mail: fabian.theis@helmholtz-muenchen.de
* contributed equally

Type of manuscript: Original Research Article

Running title: Deep learning-based propensity score computation

Keywords: Propensity scores, deep learning, autoencoder, machine learning, electronic health records, comparative effectiveness research, causal inference

Page count: 22

Word count: 249 (abstract) & 3,999 (manuscript), 7756 total word count (incl. main text, references, figure labels, tables and figures)

Tables: 2

Figures: 7

Supplementary material:
eAppendix 1: Glossary, supplementary methods, 18 supplementary figures & 15 supplementary tables
eAppendix 2: Python code jupyter notebook for autoencoder training
eAppendix 3: Rmarkdown illustrating the simulation code

Disclaimer/Potential COI statement: Janick Weberpals, Fabian Schmich, Dominik Rüttinger, Jessica Davies and Anna-Bauer Mehren are paid employees of Roche. Janick Weberpals, Fabian Schmich, Dominik Rüttinger, Jessica Davies and Anna-Bauer Mehren report to hold shares in Roche. Tim Becker is an employee of xValue and external consultant to Roche. Fabian J. Theis reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc.

Funding: This work was funded by Roche as part of the main authors (JW) postdoctoral fellowship program (RPF-ID: 498).

Acknowledgements: The authors would like to thank all staff members of Flatiron Health for their dedicated data collection and curation.

Code availability: The computing code used in this study is available as Python Jupyter Markdown scripts (.html) as supplementary material. All of the analyses described in the manuscript were performed in R version 3.2.2. The PCA and autoencoder training was performed using sckit-learn and Keras with Tensorflow backend in Python version 3.6.0, respectively. The code that was used for the simulation is available as Rmarkdown. The data that support the findings of this study have been originated by Flatiron Health, Inc. These de-identified data may be made available upon request, and are subject to a license agreement with Flatiron Health; interested researchers should contact <DataAccess@flatiron.com> to determine licensing terms

Authors’ contributions statement

JW conceptualized the study, carried out the analysis and statistical programming and drafted the manuscript. ABM and FJT supervised the project and gave significant advice at various stages of the project. TB, FJT and FS significantly contributed to the analysis of the data. FS and TB assisted with the machine learning setups used in this study and with the programming code. DR gave valuable insights to the clinical characteristics and clinical interpretation of the data. JD significantly contributed to the curation of the Flatiron Health database, ensured various data quality measures, contributed to the conceptualization of the case study and helped with the interpretation of the data. All authors critically reviewed and edited the manuscript draft. All authors agree to the submission of the manuscript.

Ethical review: These research activities are covered in Flatiron’s parent protocol which is reviewed and approved by a central IRB.

ABSTRACT

Background: Due to the non-randomized nature of real-world data, prognostic factors need to be balanced, which is often done by propensity scores (PS). This study aimed to investigate whether autoencoders, which are unsupervised deep learning architectures, might be leveraged to compute PS.

Methods: We selected patient-level data of 128,368 first-line treated cancer patients from the Flatiron Health EHR-derived de-identified database. We trained an autoencoder architecture to learn a lower-dimensional patient representation, which we used to compute PS. To compare the performance of an autoencoder-based PS with established methods, we performed a simulation study. We assessed the balancing and adjustment performance using standardized mean differences (SMD), root-mean-square-errors (RMSE), percent bias and confidence interval (CI) coverage. To illustrate the application of the autoencoder-based PS, we emulated the PRONOUNCE trial by applying the trial's protocol elements within an observational database setting, comparing two chemotherapy regimens.

Results: All methods but the manual variable selection approach led to well-balanced cohorts with average SMDs <0.1. LASSO yielded on average the lowest deviation of resulting estimates (RMSE 0.0205) followed by the autoencoder approach (RMSE 0.0248). Altering the hyperparameter setup in sensitivity analysis, the autoencoder approach led to similar results as LASSO (RMSE 0.0203 and 0.0205, respectively). In the case study, all methods provided a similar conclusion with point estimates clustered around the null (e.g. $HR_{autoencoder}$ 1.01 [95% CI 0.80-1.27] vs. $HR_{PRONOUNCE}$ 1.07 [0.83-1.36]).

Interpretation: Autoencoder-based PS computation was a feasible approach to control for confounding but did not perform better than some established approaches like LASSO.

INTRODUCTION

Randomized controlled trials (RCTs) are the gold standard when evaluating the effects of interventions on health-related outcomes. However, the digitization of healthcare infrastructure, such as electronic health records (EHR), and a boost in computational power in the past years have led to an increase in evidence generated by routinely collected healthcare data, often termed real-world data.[1–3]

Due to the heterogeneous and non-randomized nature of these data, such analyses inherit the chance to lead to misleading conclusions when biases, such as confounding bias, are not addressed appropriately.[4] Therefore, propensity score (PS) techniques are popular analytical approaches to balance patient characteristics in observational research.[5] In general, PS are defined as an individual's (*i*) conditional probability to be assigned to a particular treatment (Z_i) given observed baseline covariates (X_i) with $Pr(Z_i = 1|X_i)$.[6] By conditioning on the PS, researchers try to create positivity; that is, if a given combination of covariate values is observed in one cohort, it should also appear in the other cohort under comparison.[7] Under the assumption of no unmeasured confounding and a correctly specified PS model, unbiased treatment effects may be estimated, e.g. via matching or weighting on the PS.

There is ongoing debate about the ideal strategy to correctly specify the PS[8,9] and in the majority of cases logistic regression models are fitted using a set of a-priori investigator-defined covariates.[10] This approach is straightforward but may be error prone when interaction terms or higher-order relationships are not appropriately modeled.[11] Moreover, as healthcare databases are getting increasingly complemented by more dimensions like genomics, selecting the correct set of covariates on a manual basis becomes infeasible and automatable data-adaptive methods are warranted.

With the ability to handle high-dimensional datasets in a non-linear and automatable fashion, deep learning models are highly attractive approaches to solve these problems.[12] We aimed to investigate if autoencoders, which are unsupervised deep learning encoder-decoder architectures that learn a latent non-linear lower-dimensional covariate representation, might be leveraged as a data-adaptive alternative to compute PS for comparative effectiveness research.

The objective of this study is twofold. First, we compare the performance of covariate balancing and confounding bias reduction with the autoencoder-based PS as compared to established adjustment strategies in a simulation framework among cancer patients with a first line (1L) systemic anti-cancer treatment. In the second part of this study, we will emulate the 2015 published PRONOUNCE trial[13] by applying the trial's major protocol elements to the observational database setting of this study in order to illustrate the application of the autoencoder-based PS to a real comparative effectiveness use case.

METHODS

Data sources and study population

For this retrospective real-world data study, we used the nationwide Flatiron Health EHR-derived de-identified database which includes data from over 280 cancer clinics including more than 2.2 million US cancer patients available for analysis. The de-identified patient-level data in the EHRs include structured data (e.g. laboratory values and prescribed drugs) in addition to unstructured data collected via technology-enabled chart abstraction from physician's notes and other unstructured documents (e.g. biomarker reports). In this study, we selected patients out of tumor-specific databases and pooled them into a single cohort. Patients were eligible to be included if they were diagnosed with any primary tumor and received a 1L systemic anti-cancer treatment (CONSORT diagram, **Figure 1**).

Data curation and covariate ascertainment

We considered covariates for modeling if they were applicable across all tumor types and for at least 20% of all patients (**eTable 1**). We imputed missing covariates or those with implausible values (as defined as being outside of 1.5 x the interquartile range from the quartiles Q1 and Q3, respectively[14]) using median imputation for continuous covariates or assigning a missing-indicator category to one-hot encoded categorical variables.[15,16] In addition, we derived empirical covariates of lab and vital sign tests. As the Flatiron Health EHR-derived de-identified database does not contain records of claims, procedure codes and outpatient diagnosis codes, these empirical covariates were derived from the frequency of clinical laboratory tests and vital sign tests (which corresponds to steps 1-3 of the high-dimensional propensity score algorithm[17]), which resulted in 123 additional covariates (**eTable 1**). All covariates were measured at or before the start of 1L therapy (= index date) with a maximum lookback window period of 90 days relative to the index date (**eFigure 1**).[18,19]

Non-linear latent variables **and propensity scores computed by autoencoder**

The following section briefly illustrates the autoencoder-based PS computation (terminology used in this paragraph is defined in eAppendix1 and in Bi et al.[20]).

Autoencoders are unsupervised neural network architectures that generally consist of an input layer, a lower-dimensional hidden “bottleneck” layer, and an output layer with the same dimensions as the input layer. Conceptually, the autoencoder-based PS computation can be described as follows (**Figure 2**). All available information about a patient may be defined as a high-dimensional covariate vector serving as the input layer. This input layer is sequentially compressed to arrive at a latent non-linear lower-dimensional covariate representation in the hidden bottleneck layer (encoding). Given the lower-dimensional information of the bottleneck layer, the actual input information can be reconstructed (decoding); the decoded information is leveraged in autoencoders in order to adjust the network parameters in each iteration by computing the loss between the actual data and the predicted reconstruction. Due to the compression and the optimization of parameters of the neural network in each encoding–decoding iteration step, the autoencoder learns essential features describing the highest variance of a dataset. This way the bottleneck layer captures the true data manifold in a much lower-dimensional representation (embedding) that can eventually be used to specify the PS.

Following the above described general setup, we developed an autoencoder architecture (details on architecture, hyperparameters and code can be found in eAppendix1). To compute the PS based on the trained embedding, we used a logistic regression as the final output layer.

Propensity score estimation methods for comparison

To investigate the performance of an autoencoder-based PS, we chose established adjustment (multivariable regression) and propensity score estimation methods (manual variable selection, principal component analysis [PCA] and LASSO) for comparison employing a simulation framework (**Table 1**). We additionally extended all machine learning models by the set of empirical covariates that were derived as described above (EC extended models 7-9). More details see **eAppendix1**.

Simulation setup

The overall objective was to simulate different realistic scenarios of confounding bias between a fictional head-to-head drug comparison and to investigate the resulting balancing and adjustment after 1:1 PS matching with PS computed using the aforementioned propensity score estimation methods (**Table 1**). We defined the outcome of interest for this simulation study as overall survival, which we computed as the time from index date to death due to any reason or censoring.

The general simulation algorithm is illustrated in **Figure 3A**. In brief, all eligible patients were equally randomized to either a Drug A or a Drug B cohort to remove any prognostic association of the covariates to the assignment probability to one of the cohorts. This resulted in a hazard ratio (HR) for overall survival of 1.00 (95% confidence interval [CI] 0.99-1.01), which served as the true estimate in this simulation (**eFigure 2**). In a next step, we grouped patients were grouped into prognostic quartiles (Q1-Q4) according to their baseline hazards towards the outcome (overall survival) with patients in Q1 having a good prognosis (lowest hazard) to patients in Q4 having a poor prognosis (**eFigure 3 & eTable 2)**. The prognostic quartiles are based on a published prognostic score for overall survival *(***eFigures 4 & 5***)* that was developed within a large pan-cancer cohort and is derived from a formula with strongly prognostic demographic, clinical, routine hematology, and blood chemistry parameters (**eTable 3**) that were modeled within a Cox proportional hazard framework to derive a multivariable prognostic risk model for overall survival.[21] The resulting prognostic score was validated in two independent phase I and III clinical studies. To simulate baseline imbalances, we exploited the correlation between prognostic score-based balance measures for propensity score models with bias in the treatment effect estimate using conditional re-sampling as described in the following.[22] Out of the Drug A cohort, we sampled 10,000 patients completely at random and independent of their assignment to the prognostic quartiles to arrive at a homogenous sample with a constant prognosis in each replication step. In contrast, we sampled 10,000 patients randomized to the Drug B cohort with a conditional sampling probability based on their assignment to a prognostic quartile (e.g. scenario 1: patients in Q1 were sampled with a probability of 40%, in Q2 with 30%, in Q3 with 20%, and in Q4 with 10%). Because quartile membership is associated with overall survival, the conditional sampling of the Drug B cohort (as compared to the random sampling of the Drug A cohort) naturally induces a spurious association, which is solely driven by the variables defining the quartiles. We applied this sampling scheme in total 27 different sampling probabilities with 100 replications each to simulate various scenarios of confounding bias yielding biased estimates with different magnitudes and directions away from the true HR of 1.00 (**Figure 3B**).

We finally assessed the comparative performance of each PS computation method as to how much each method was able to adjust for the above described induced spurious association. For this purpose, we matched the resulting cohorts without replacement in a 1:1 ratio with a caliper width of 0.2 standard deviations of the predicted PS logit[23] and HRs were estimated using Cox proportional hazards regression models with a robust variance estimator.[24] Simulations of treatment effects other than a null treatment effect were not considered to avoid complications with the collapsibility [25] and proportional hazards assumption[26] of HRs.

We assessed the overall balance in the distribution of important baseline covariates after PS matching using standardized mean differences (SMD) with a cut-off of < 0.1 indicating sufficient balance.[27] To assess the average deviation of the resulting HRs and the true HR of 1.00, we computed the root-mean-square-error (RMSE) as performance metrics. To measure the uncertainty of the point estimates, we computed the coverage probability as the proportion of times the estimated 95% confidence interval included the true HR of 1.00.[28,29] Additionally, we estimated the absolute bias (in %) as $|\frac{HR_{pooled}-HR_{True}}{HR_{True}}x100|$for each simulation scenario.[30]

Case study

To illustrate the application of the autoencoder-based PS in comparative effectiveness research, we emulated the PRONOUNCE trial by applying the major protocol design elements of this trial within the observational Flatiron Health EHR-derived de-identified database.

In brief, the PRONOUNCE trial was a randomized, open-label, phase III trial aimed at evaluating the comparative efficacy of carboplatin/pemetrexed followed by pemetrexed maintenance vs. bevacizumab/carboplatin/paclitaxel followed by bevacizumab maintenance as 1L treatment among advanced nonsquamous non-small-cell lung cancer patients.[13] In terms of overall survival, the PRONOUNCE trial did not find a difference in treatment efficacy for either of the combinations, which served as our expected outcome for the emulation of this trial.

For the implementation of the major in-/exclusion criteria and study design elements, we followed the target trial emulation framework by Hernán & Robins[31] and summarized the comparison to the original in **eTable 4** and **eFigure 6**. Instead of a random assignment to either treatment strategy in a 1:1 ratio in the original trial, we applied propensity score matching (applying the different PS computation approaches) in a 1:1 ratio (nearest neighbor without replacement as main analysis)[32] and SMR weighting (sensitivity analysis)[33]. We derived estimates for overall survival using Cox proportional hazards regression with the initiation of maintenance therapy as start of follow-up. The causal contrast of interest was analyzed as the counterfactual comparison of initiators of the two different treatment strategies as an observational equivalent of the RCT's intent-to-treat analysis. Further details are outlined in the supplementary methods (eAppendix1).

RESULTS

The characteristics of the eligible simulation population are displayed in **eTable 5**. Results of the hyperparameter selection and evaluation are illustrated in the supplementary material (**eFigures 7-11**) and computation times for the autoencoder models and simulations are summarized in **eFigures 12&13** and **eTable 6**, respectively.

Simulation – balancing properties

Figure 4 summarizes the average balancing performance of important baseline characteristics by simulation scenario. In general, most PS estimation methods led to sufficient balancing of important patient characteristics at baseline (SMD<0.1). In some scenarios, imbalances for some covariates were observed for PS computed using manual variable selection. Investigating SMDs by scenario indicated that those imbalances resulted from some of the more extreme confounded scenarios (**eFigure 14**).

Simulation – RMSE, percent bias and coverage

The overall results across all simulated scenarios and iterations are illustrated in **Table 2**. Estimates without any adjustment resulted on average in high RMSEs (0.1205) and bias (10.4% bias) and low coverage (16.41%). When covariates were manually chosen (models 2 & 3), the PS method led on average to a lower RMSE (0.0670 vs 0.0790), bias (5.73% vs 6.75%) and a higher coverage (32.81% vs. 27.67%) as compared to choosing the same covariates for direct outcome regression, respectively. Point estimates were observed to scatter broadly around the null for both methods (**eFigure 15**). Comparisons between model standard errors and empirical standard errors indicated a less reliable variance estimation for models 1-3 (**eTable 7)**. The PCA PS estimation method led to a noticeable improvement in adjustment performance as compared to selecting covariates manually with a RMSE of 0.0293 and 0.0329 for PCA and PCA EC, respectively. Employing an autoencoder-based estimation of the PS led to further improvements in RMSEs of 0.0248 and 0.0265, bias of 2.00% and 2.15% and coverage of 87.70% and 85.19% for autoencoder and autoencoder EC respectively. The best adjustment performance was observed with both LASSO approaches with around 1.7% bias and nearly 94% coverage.

We observed the same pattern when we compared the point estimates by simulated scenarios (**Figure 5**). As expected, unadjusted estimates ranged from approximately 0.8 to over 1.2. Both LASSO approaches followed by the autoencoder approaches demonstrated the best adjustment performance in most of the cases. In particular, we observed that the LASSO EC model had the best confidence interval coverage to include the true HR in at least 95% of the times in 14 out of the 27 simulated scenarios (**Figure 5 and eTables 8-10**). When the % bias was compared by simulated scenario, the results were consistent with less than 2% (LASSO and LASSO EC) and 3% (autoencoder and autoencoder EC) bias in almost all of the scenarios (**Figure 6** and **eTable 9**).

Sensitivity analyses

When we changed the autoencoder architecture from three hidden layers to one in sensitivity analysis I, the performance of the autoencoder-based models slightly improved (**eTable 11**). The overall performance remained nearly the same when the main architecture was altered to having a 128-dimensional bottleneck layer size in sensitivity analysis II (**eTable 12**). Combining the architecture alterations from sensitivity analysis I and II, results of the autoencoder approach were comparable to the ones of the LASSO approaches with an average RMSE of 0.0203 (**eTable 13**). When taking all possible PCs, instead of those describing 80% of the cumulative variance explained, the performance according to RMSE and bias decreased while the coverage improved (**eTable 14**). Increasing the number of replications to 500 did not noticeably change the results of the main analysis, indicating that 100 replications per scenario were sufficient (**eTable 15**).

Case study

There were 781 patients eligible for the case study (**eFigure 16**). The results are summarized in **Figure 7**. All analyses suggested a null association with the unadjusted point estimate being slightly below the null. All adjusted models ranged between point estimates of 1.00 to 1.09 with the autoencoder analysis being slightly closer to the null ($HR_{autoencoder}$ 1.01 [95% CI 0.80-1.27] vs. $HR_{PRONOUNCE}$ 1.07 [0.83-1.36]) as compared to the autoencoder EC model (HR 1.09 [95% CI 0.87-1.37]). SMR weighting led to very similar estimates with the exception of the LASSO approaches having much wider confidence intervals (**eFigure 17**).

DISCUSSION

In this RWD study, we developed a novel automated autoencoder-based approach and compared it with established approaches. Using a comprehensive simulation framework, we observed that in terms of confounding control, the autoencoder-based approach led to reasonable results, but did not perform substantially better than some of the established approaches such as LASSO. In an empirical case study emulating the PRONOUNCE trial using observational data, the autoencoder-based results were consistent with the conclusion of the original trial.

Propensity scores are frequently used analytical tools (**eFigure 18**) since they enable researchers to collapse many dimensions of confounding covariates into a single dimension while still maintaining sufficient precision. The advantage of deep learning-based PS is the ability to easily handle large amounts of data involving complex associations between covariates. An earlier study from 2008 investigated different techniques in PS estimation with various non-linear and non-additive associations on 10 binary/continuous covariates and concluded that even a rather simple neural network outperformed recursive partitioning algorithms in terms of providing the least numerically biased estimates.[34] This may suggest that the appropriate modeling of potentially non-linear covariate structures may be of relevant importance for confounding control. Especially analyses in EHR data may benefit from autoencoder-based PS as these usually capture routine care laboratory measurements and vital sign parameters which have been shown to be of paramount prognostic value.[21,35] This may explain why in this study the autoencoder-based PS performed better than the PCA approaches since, in case of no non-linearity, both methods should in principle lead to similar results.[36] However, given that continuous covariates are still usually rather rare in healthcare databases, we may have underestimated the abilities of the autoencoder-based approach in this study and further studies are warranted once multimodal data elements, such as medical images and sequencing data, complement contemporary databases.[37]

Application and use cases

The autoencoder-based PS can be generally used in any type of comparative effectiveness study where sufficient confounder balancing between two cohorts is required. In the here presented comparative effectiveness case study, it was possible to derive the same qualitative conclusion as in the PRONOUNCE trial by applying autoencoder-based PS. Although the primary objective in the case study was to test the use of autoencoder-based PS in a real comparative effectiveness research setup, the equal results of all methods may be explained with the fact that confounding bias was not as strong in this particular research question as compared to some of the more extreme scenarios in the simulation. This seems plausible given that due to the variety of possible treatments and sometimes lacking evidence for the most effective combination, the selective channeling of patients with higher risk (as often observed with prescription drugs like COX-2 inhibitors vs. non-selective NSAIDs) may not be apparent. This may underline an attractive feature of the autoencoder-based PS, which could be used as an automated and data-adaptive sensitivity analysis in comparative effectiveness studies with unknown extent of confounding bias.

Especially in the era of precision medicine, in which treatment decisions for specific subpopulations of patients are based on distinct molecular characteristics, comparative effectiveness research might play an increasingly important role in addressing challenges e.g. in the area of early clinical development of new therapeutics. Here, designs such as external control arms are interesting approaches which could benefit by advanced analytics like deep learning-based PS. A recent proof-of-concept study assessed how well external controls could have approximated the actual standard-of-care controls in nine lung cancer trials.[38] The authors reported that the comparison of estimates between RCTs and external controls resulted in a Pearson correlation coefficient of 0.86. This is an encouraging example suggesting that external control arms come with a sufficient validity and can play an important role in facilitating real-world data to support early clinical development and regulatory submissions.[3,39]

Strengths and limitations

Due to the nature of routinely collected health records, there is missing data. In this study, we employed median imputation and assigning a missing-indicator category to one-hot encoded categorical variables because this or similar approaches were suggested to have good performance in studies with large datasets where multiple imputation would be computationally very expensive and generally not operationalizable.[16,40] This approach is also supported by various recent prediction models trained on EHR data which reported outstanding performance.[15,41]

In addition, data-adaptive approaches always inherit the risk of including covariates that may be collider covariates (M-bias), instrumental covariates (Z-bias), or causal intermediates. Colliders are covariates that open a causal path from exposure to outcome.[42] Including such covariates in the PS computation may induce a spurious association where in fact there is none. As besides directed acyclic graphs there is no formal way to test for colliders, it may be difficult to exclude such variables prior to PS computation. However, Schneeweiss found that under realistic scenarios a collider-induced bias was negligible and outweighed by the adjustment effect for other covariates.[43] Instrumental variables (IV) are covariates that are only associated with the exposure but not with the outcome. IVs are frequently used to control for unmeasured confounding[44] but also introduce bias (Z-bias) when conditioning on them. Especially in oncology, calendar period effects are strong predictors for therapy decisions once new breakthrough treatments are approved.[45] Although there is a theoretical chance to have unintentionally included IVs, Meyers et al. showed that only in the presence of strong unmeasured confounding does Z-bias have effects worth mentioning.[46] While expert knowledge plays an important role in avoiding covariates that could mediate the association between exposure and outcome[47], the risk of adjusting for causal intermediates can also be mitigated with appropriate study designs such as an active comparator, new user design as applied in this study.[4,43]

A unique strength of this study is the novelty approach to learn patient representations for PS computation in a data-adaptive manner, which we found to have a reasonable performance and which may serve as a promising tool for the future once more data elements complement contemporary databases. Applying comprehensive sensitivity analyses, we found the methodology to be robust as all setups and scenarios resulted in a similar conclusion. The observation that the autoencoder architecture with less hidden layers and a larger bottleneck layer led to results closer to LASSO gave some concern that this may have been the consequence of overfitting of the main model. Nevertheless, differences were marginal and did not change the main conclusion while the hyperparameter setup of the main model was found to be a reasonable trade-off between compactness of the resulting embedding and sufficient reconstruction performance and generalizability.

It is further important to credit that the autoencoder approach is a pure unsupervised method, which means that the confounding control in this study has been solely achieved without optimizing the network towards the probability of patients receiving the treatment, which needs to be acknowledged when comparing to supervised approaches like LASSO. Hence, potential deep-learning architectural extensions would be of utmost interest, e.g. by jointly modeling targets and inputs using end-to-end learning architectures.

A limitation of this simulation is that due to the non-collapsibility of HRs, only a null treatment effect could be simulated which may in future research be addressed by estimating risk-differences and more sophisticated simulation techniques such as plasmode simulations.[48,49] In addition, variance estimation seemed to be less reliable for models 1-3 (**eTable 7**), limiting the ability to make final conclusions about their true CI coverage.

For this study, it was possible to use a large underlying population to train and empirically examine the comparative performance of the proposed autoencoder approach. This real-world database provided comprehensive oncology-specific data, which underwent a rigorous data quality assurance process prior to release.

Finally, it is important to acknowledge that this study primarily focused on the analytical aspects to reduce confounding. Carefully chosen analysis always needs to go along with a causal study design to avoid serious biases such as reverse causality and immortal time bias, which are known as sources for much larger bias than conventional confounding.[43,50]

Conclusion

In summary, we developed an autoencoder-based PS computation that our assessment found to be a feasible approach to reduce confounding bias, although not with a substantially stronger performance than some of the established approaches such as LASSO. As a promising tool for the future, it may be considered alongside with established approaches in non-randomized comparisons in comparative effectiveness research.

REFERENCES

1. Basch E, Schrag D. The Evolving Uses of “Real-World” Data. *JAMA*. 2019;321(14):1359-1360. doi:10.1001/jama.2019.4064

2. Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA*. 2018;320(9):867-868. doi:10.1001/jama.2018.10136

3. U.S. Food and Drug Administration. Framework for FDA’s Real-World Evidence Program. Published online December 2018. https://www.fda.gov/media/120060/download

4. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr Epidemiol Rep*. 2015;2(4):221-228. doi:10.1007/s40471-015-0053-5

5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41

6. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786

7. Westreich D, Cole SR. Invited Commentary: Positivity in Practice. *Am J Epidemiol*. 2010;171(6):674-677. doi:10.1093/aje/kwp436

8. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156. doi:10.1093/aje/kwj149

9. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable Selection for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing Healthcare Databases. *Epidemiol Camb Mass*. 2017;28(2):237-248. doi:10.1097/EDE.0000000000000581

10. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59(5):437-447. doi:10.1016/j.jclinepi.2005.07.004

11. Glynn RJ, Schneeweiss S, Stürmer T. Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98(3):253-259. doi:10.1111/j.1742-7843.2006.pto_293.x

12. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. Vol 1. MIT press Cambridge; 2016.

13. Zinner RG, Obasaju CK, Spigel DR, et al. PRONOUNCE: randomized, open-label, phase III study of first-line pemetrexed + carboplatin followed by maintenance pemetrexed versus paclitaxel + carboplatin + bevacizumab followed by maintenance bevacizumab in patients

ith advanced nonsquamous non-small-cell lung cancer. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2015;10(1):134-142. doi:10.1097/JTO.0000000000000366

14. Tukey JW. Exploratory data analysis. *Mass Addison-Wesley*. Published online 1977.

15. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116. doi:10.1038/s41586-019-1390-1

16. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun*. 2019;10(1):390. doi:10.1038/s41467-018-07931-2

17. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiol Camb Mass*. 2009;20(4):512-522. doi:10.1097/EDE.0b013e3181a663cc

18. Czwikla J, Jobski K, Schink T. The impact of the lookback period and definition of confirmatory events on the identification of incident cancer cases in administrative data. *BMC Med Res Methodol*. 2017;17(1):122. doi:10.1186/s12874-017-0407-4

19. Schneeweiss S, Rassen JA, Brown JS, et al. Graphical Depiction of Longitudinal Study Designs in Health Care Databases. *Ann Intern Med*. Published online March 12, 2019. doi:10.7326/M18-3079

20. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol*. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189

21. Becker T, Weberpals J, Jegg AM, et al. An Enhanced Prognostic Score for Overall Survival of Patients with Cancer Derived from a large Real World Cohort. *Ann Oncol*. 2020;0(0). doi:10.1016/j.annonc.2020.07.013

22. Stuart EA, Lee BK, Leacy FP. Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. *J Clin Epidemiol*. 2013;66(8 0):S84-S90.e1. doi:10.1016/j.jclinepi.2013.01.013

23. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014;33(6):1057-1069. doi:10.1002/sim.6004

24. Lin DY, Wei LJ. The Robust Inference for the Cox Proportional Hazards Model. *J Am Stat Assoc*. 1989;84(408):1074-1078. doi:10.1080/01621459.1989.10478874

25. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. *Stat Sci*. 1999;14(1):29-46. doi:10.1214/ss/1009211805

26. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? *JAMA*. Published online March 13, 2020. doi:10.1001/jama.2020.1267

27. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107. doi:10.1002/sim.3697

28. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24):4279-4292. doi:10.1002/sim.2673

29. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086

30. Desai RJ, Wyss R, Abdia Y, et al. Evaluating the use of bootstrapping in cohort studies conducted with 1:1 propensity score matching-A plasmode simulation study. *Pharmacoepidemiol Drug Saf*. 2019;28(6):879-886. doi:10.1002/pds.4784

31. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

32. Ho D, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *J Stat Softw*. 2011;42(1):1-28. doi:10.18637/jss.v042.i08

33. Greifer N. *WeightIt: Weighting for Covariate Balance in Observational Studies (R Package Version 0.5. 1)*.; 2019.

34. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546-555. doi:10.1002/pds.1555

35. Garrido-Laguna I, Janku F, Vaklavas C, et al. Validation of the Royal Marsden Hospital prognostic score in patients treated in the Phase I Clinical Trials Program at the MD Anderson Cancer Center. *Cancer*. 2012;118(5):1422-1428. doi:10.1002/cncr.26413

36. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep*. 2016;6:26094. doi:10.1038/srep26094

37. Binder H. Big Data und Deep Learning in der Onkologie. *Onkol*. 2018;24(5):361-367. doi:10.1007/s00761-018-0359-2

38. Carrigan G, Whipple S, Capra WB, et al. Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. *Clin Pharmacol Ther*. 0(ja). doi:10.1002/cpt.1586

39. Burcu M, Dreyer NA, Franklin JM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. *Pharmacoepidemiol Drug Saf*. n/a(n/a). doi:10.1002/pds.4975

40. Rodenburg FJ, Sawada Y, Hayashi N. Improving RNN Performance by Modelling Informative Missingness with Combined Indicators. *Appl Sci*. 2019;9(8):1623. doi:10.3390/app9081623

41. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digit Med*. 2018;1(1):1-10. doi:10.1038/s41746-018-0029-1

42. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. *Am J Epidemiol*. 2012;176(10):938-948. doi:10.1093/aje/kws165

43. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clin Epidemiol*. 2018;10:771-788. doi:10.2147/CLEP.S166545

44. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537-554. doi:10.1002/pds.1908

45. Mack CD, Brookhart MA, Glynn RJ, Stürmer T. Calendar Time As An Instrumental Variable In Nonexperimental Comparative Effectiveness Research Of Emerging Therapies. *Value Health*. 2013;16(3):A129-A130. doi:10.1016/j.jval.2013.03.629

46. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174(11):1213-1222. doi:10.1093/aje/kwr364

47. Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. Effect modification, interaction and mediation: an overview of theoretical insights for clinical investigators. *Clin Epidemiol*. 2017;9:331-338. doi:10.2147/CLEP.S129728

48. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219-226. doi:10.1016/j.csda.2013.10.018

49. Huitfeldt A, Stensrud MJ, Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. *Emerg Themes Epidemiol*. 2019;16(1):1. doi:10.1186/s12982-018-0083-9

50. Weberpals J, Jansen L, Herk-Sukel MPP van, et al. Immortal time bias in pharmacoepidemiological studies on cancer patient survival: empirical illustration for beta-blocker use in four cancers with different prognosis. *Eur J Epidemiol*. 2017;32(11):1019-1031. doi:10.1007/s10654-017-0304-5

FIGURES

Figure 1. Consort diagram illustrating selection of eligible patients for simulation.

Figure 2. Conceptual architecture of patient representation learning and autoencoder-based propensity score computation.

Figure 3. (A) Sampling algorithm for simulation and **(B)** overview of magnitude of induced confounding bias by simulation scenario (CI = Confidence interval, HR = Hazard ratio, Q = Quartile, ROPRO = Real world prognostic score)

Figure 4. Baseline covariate balance by propensity score computation method and simulation scenario. Average standardized mean differences (SMDs) are displayed for each of the 27 scenarios per baseline characteristic (ALP = Alkaline phosphatase, ALT = Alanine aminotransferase, AST = Aspartate aminotransferase, BMI = Body mass index, EC = Empirical covariates, ECOG = Eastern Cooperative Oncology Group (ECOG) Performance Status, LASSO = Least absolute shrinkage and selection operator, LDH = Lactate dehydrogenase, NLR = Neutrophil-to-lymphocyte ratio, PC(A) = Principal component (analysis),

Figure 5. Average hazard ratios (HRs) for each of the 27 simulated scenarios and paneled PS estimation method. Asterisk (*) indicates that the confidence interval coverage for the respective scenario included the true HR in at least 95% of the times. The red dashed line indicates the true HR that is intended to be recovered by the propensity score adjustment (EC = Empirical covariates, HR = Hazard ratio, LASSO = Least absolute shrinkage and selection operator, PCA = Principal component analysis)

Figure 6. Comparison of average absolute % bias by simulation scenario for each propensity score (PS) estimation method (EC = Empirical covariates, LASSO = Least absolute shrinkage and selection operator, PCA = Principal component analysis).

Figure 7. Forest plot illustrating hazard ratio (HRs) and 95% CIs for overall survival by propensity score (PS) estimation method (CI = Confidence interval, EC = Empirical covariates, HR = Hazard ratio, LASSO = Least absolute shrinkage and selection operator, PCA = Principal component analysis).

Table 1

Formatted: Top: 0.62"

Table 1. Models and adjustment strategies compared in simulation framework.

Model	Adjustment strategy[a]	Data-adaptive covariate selection / transformation	Covariates adjusted for or potential covariates to choose from
1	Unadjusted	-	-
2	Multivariable Regression (direct outcome model)	No	Age, cancer entity, gender, stage, histology, healthcare provider, race/ethnicity, time from initial cancer diagnosis to 1L initiation, calendar year of initial cancer diagnosis
3	Manual variable selection	No	Age, cancer entity, gender, stage, histology, healthcare provider, race/ethnicity, time from initial cancer diagnosis to 1L initiation, calendar year of initial cancer diagnosis
4	LASSO	Selection	All generally available covariates. Algorithm picks covariates according to shrinkage/regularization
5	PCA	Transformation	All generally available covariates. Algorithm computes linear transformation of all covariates in a dataset to principal components (PCs) of which the top *n* PCs, explaining 80% variance, were chosen
6	Autoencoder	Transformation	All generally available covariates. Algorithm computes lower-dimensional representation of j dimensions based on non-linear data operations into latent-space variables
7	LASSO EC	Transformation	Model 4 + 123 empirical covariates[c]
8	PCA EC	Selection	Model 5 + 123 empirical covariates[c]
9	Autoencoder EC	Transformation	Model 6 + 123 empirical covariates[c]

Abbreviations: 1L = first-line systemic cancer treatment, EC = Empirical covariates, LASSO = Least absolute shrinkage and selection operator, PC(A) = Principal component (analysis)

[a] In model 2 the estimate is directly computed from a multivariable regression while models 3-9 are based on propensity score matching

[b] Total of 318 demographic, clinical, cancer-/disease-specific covariates (see eTable 1)

[c] Total of 123 empirical frequency covariates derived, corresponds to step 1-3 of the high dimensional propensity score algorithm (see eTable 1)

Table 2

Table 2. Summary of adjustment performance across all scenarios.

Method	RMSE	Bias (%)	CI coverage (%)
Unadjusted	0.1205	10.4	16.41
Multivariable regression	0.0790	6.75	27.67
Manual variable selection	0.0670	5.73	32.81
LASSO	0.0205	1.65	93.74
PCA	0.0293	2.39	79.59
Autoencoder	0.0248	2.00	87.70
LASSO EC	0.0210	1.69	93.52
PCA EC	0.0329	2.71	74.00
Autoencoder EC	0.0265	2.15	85.19

Abbreviations: CI = Confidence interval, EC = Empirical covariates, LASSO = Least absolute shrinkage and selection operator, PC(A) = Principal component (analysis), RMSE = Root mean squared error

Figure 1

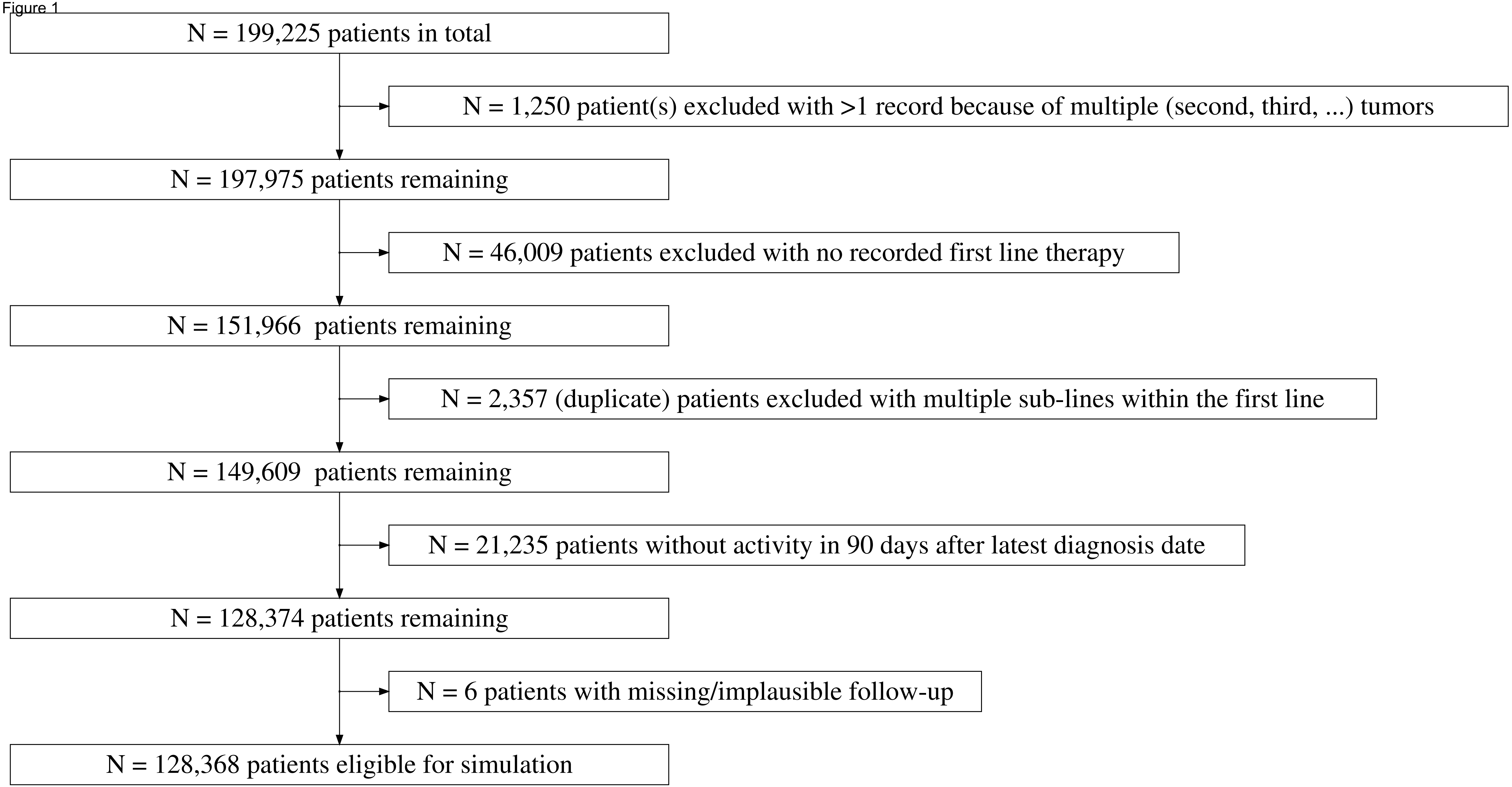
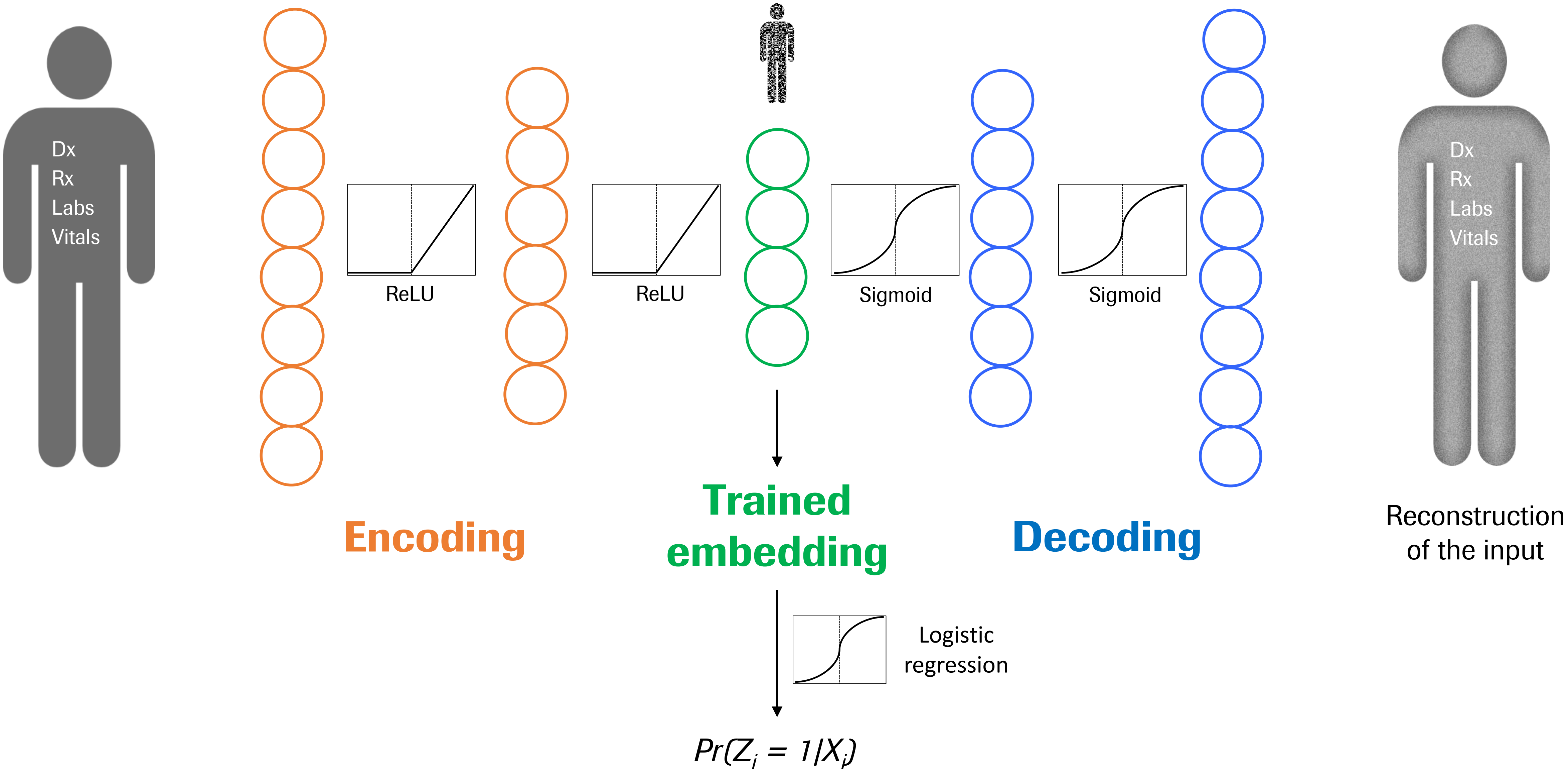

N = 199,225 patients in total
N = 1,250 patient(s) excluded with >1 record because of multiple (second, third, ...) tumors
N = 197,975 patients remaining
N = 46,009 patients excluded with no recorded first line therapy
N = 151,966 patients remaining
N = 2,357 (duplicate) patients excluded with multiple sub-lines within the first line
N = 149,609 patients remaining
N = 21,235 patients without activity in 90 days after latest diagnosis date
N = 128,374 patients remaining
N = 6 patients with missing/implausible follow-up
N = 128,368 patients eligible for simulation

Figure 2


Dx
Rx
Labs
Vitals
ReLU
ReLU
Sigmoid
Sigmoid
Dx
Rx
Labs
Vitals
Encoding
Trained embedding
Decoding
Reconstruction of the input
Logistic regression
$Pr(Z_i = 1|X_i)$

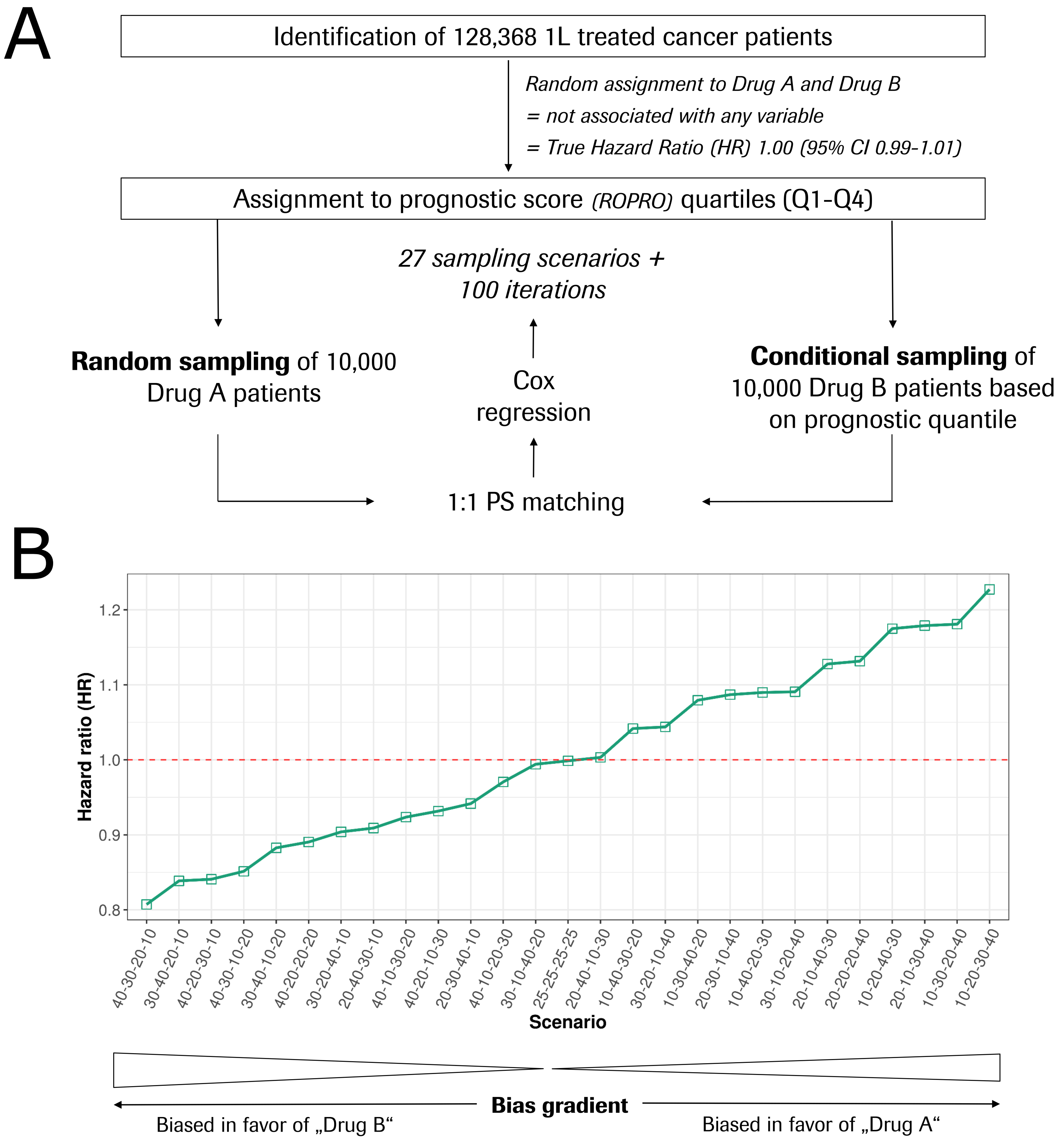

A
Identification of 128,368 1L treated cancer patients
Random assignment to Drug A and Drug B
= not associated with any variable
= True Hazard Ratio (HR) 1.00 (95% CI 0.99-1.01)
Assignment to prognostic score (ROPRO) quartiles (Q1-Q4)
27 sampling scenarios +
100 iterations
Random sampling of 10,000
Drug A patients
Cox
regression
Conditional sampling of
10,000 Drug B patients based
on prognostic quantile
1:1 PS matching
B
Hazard ratio (HR)
1.2
1.1
1.0
0.9
0.8
40-30-20-10
30-40-20-10
40-20-30-10
40-30-10-20
30-40-10-20
40-20-20-20
30-20-40-10
20-40-30-10
40-10-30-20
40-20-10-30
20-30-40-10
40-10-20-30
30-10-40-20
25-25-25-25
20-40-10-30
10-40-30-20
30-20-10-40
10-30-40-20
20-30-10-40
10-40-20-30
30-10-20-40
20-10-40-30
20-20-20-40
10-20-40-30
20-10-30-40
10-30-20-40
10-20-30-40
Scenario
Bias gradient
Biased in favor of „Drug B“
Biased in favor of „Drug A“

Figure 4

Click here to access/download;Figure;Figure4_SMD_by_method_V1_2021-01-15.tiff

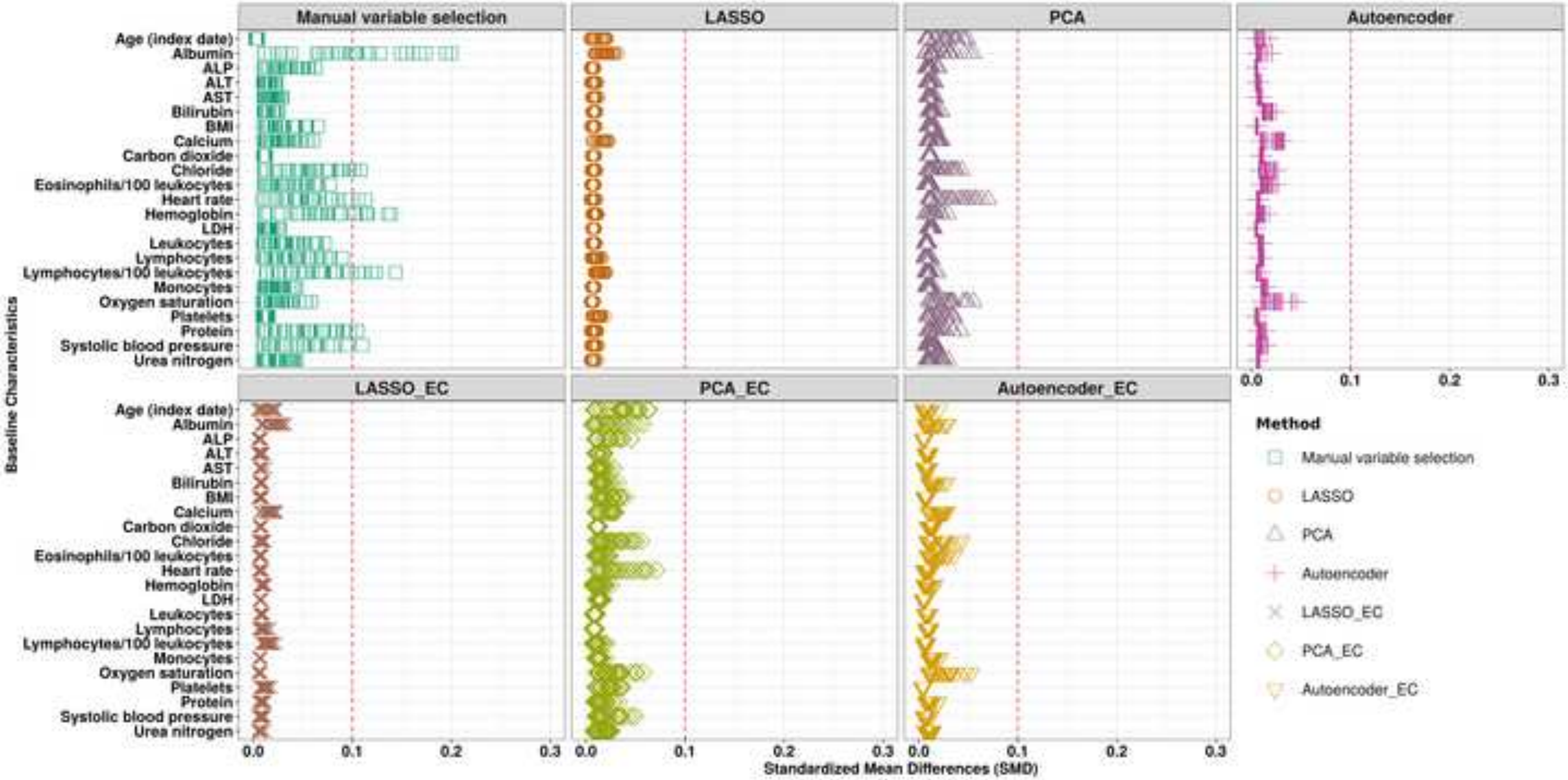
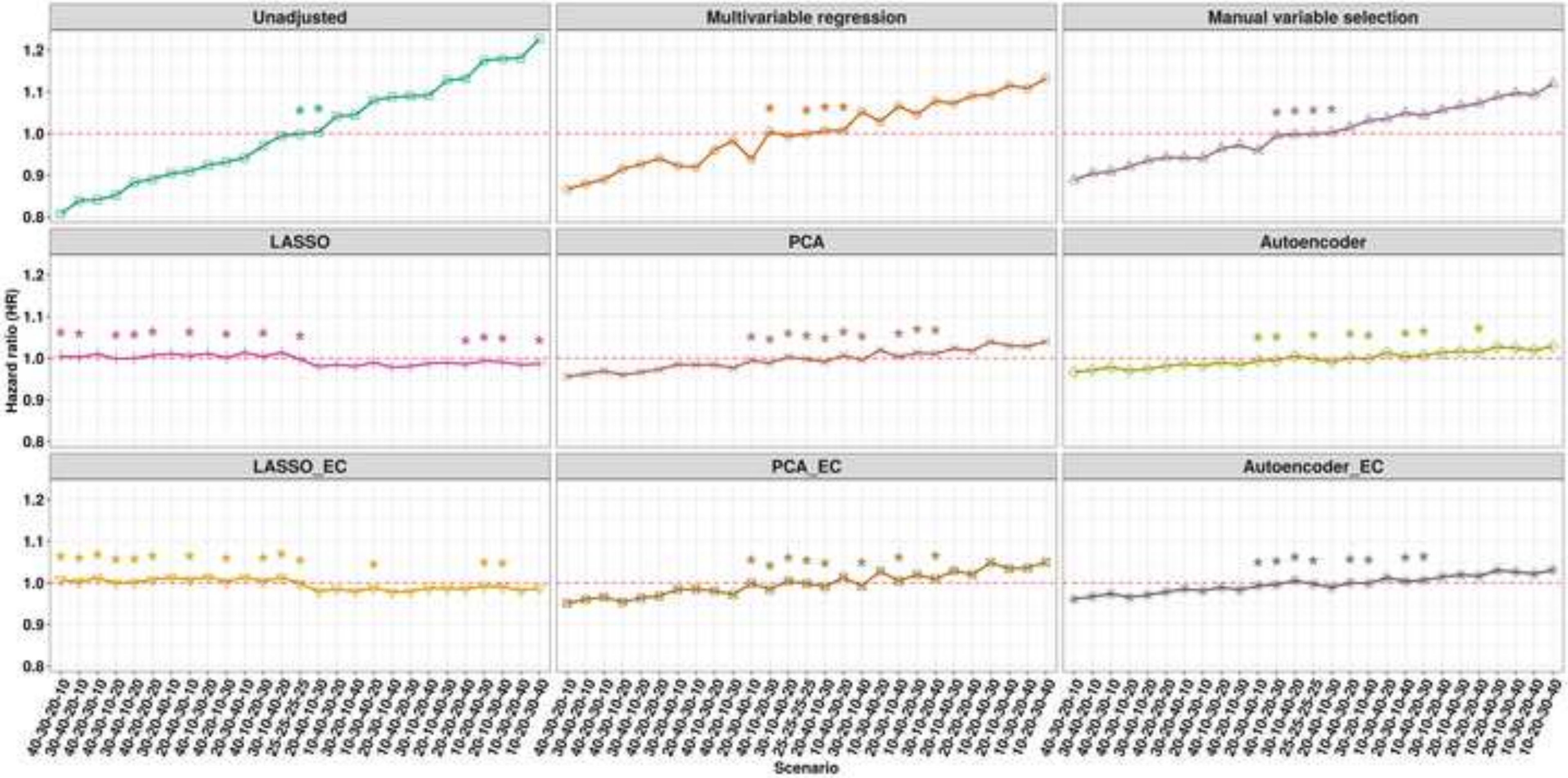


Figure 5

Click here to access/download;Figure;Figure5_HR_by_scenario_2021-01-15.tiff



Click here to access/download;Figure;Figure6_bias_2021-01-15.tiff

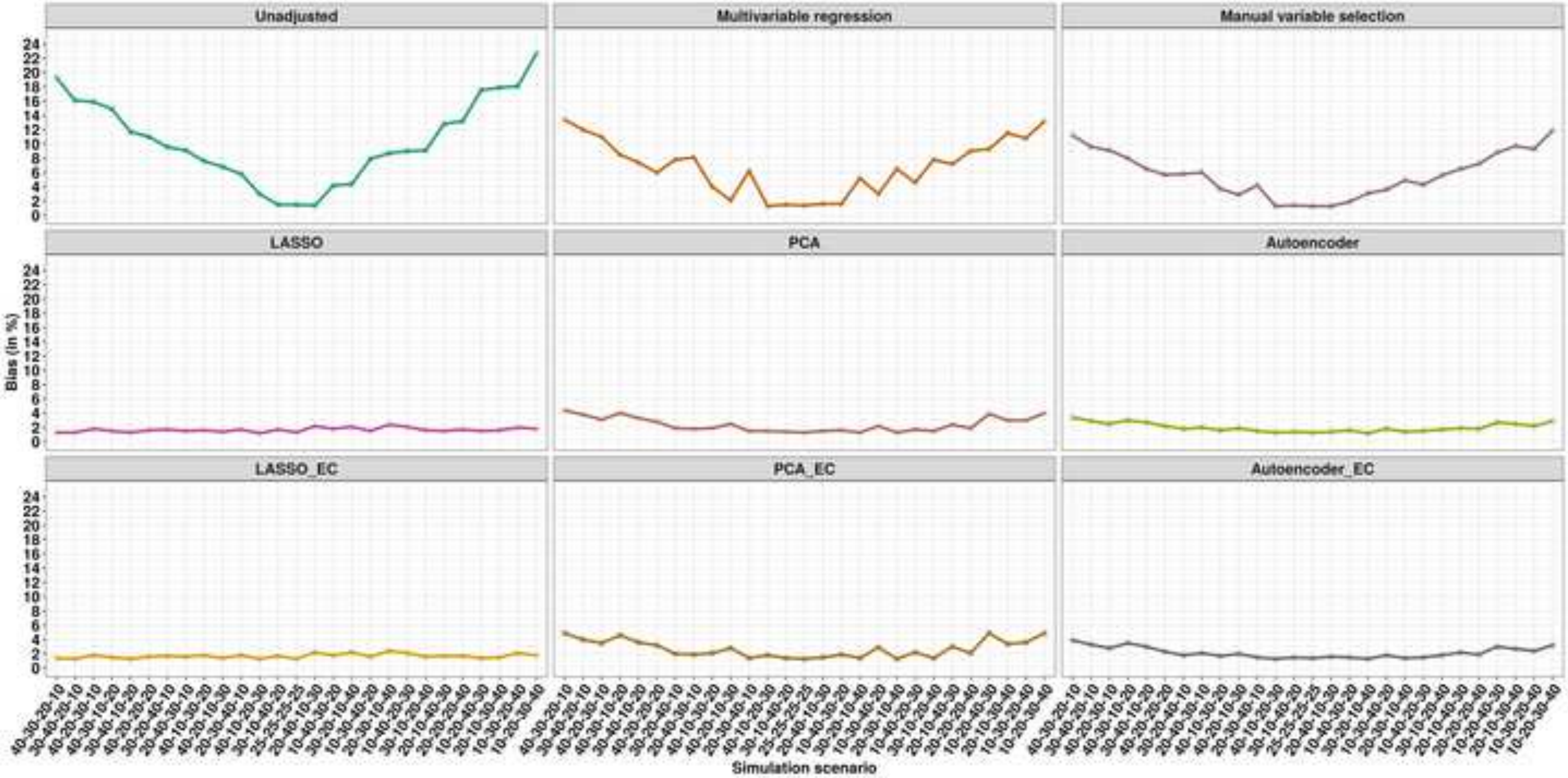
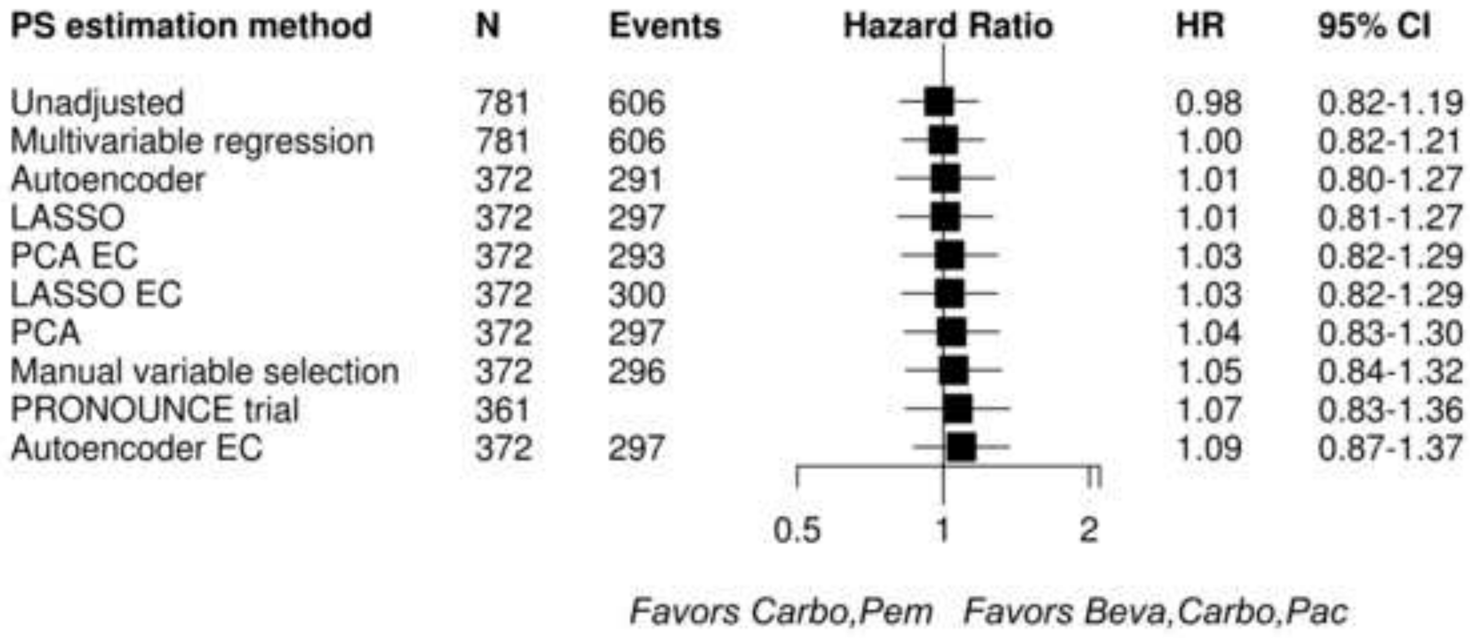


Figure 7



PS estimation method	N	Events	HR	95% CI
Unadjusted	781	606	0.98	0.82-1.19
Multivariable regression	781	606	1.00	0.82-1.21
Autoencoder	372	291	1.01	0.80-1.27
LASSO	372	297	1.01	0.81-1.27
PCA EC	372	293	1.03	0.82-1.29
LASSO EC	372	300	1.03	0.82-1.29
PCA	372	297	1.04	0.83-1.30
Manual variable selection	372	296	1.05	0.84-1.32
PRONOUNCE trial	361		1.07	0.83-1.36
Autoencoder EC	372	297	1.09	0.87-1.37

eAppendix 1

eAppendix 2

Click here to access/download

Supplemental Digital Content

eAppendix2_Jupyter Notebook_autoencoder_R2.html

eAppendix 3

Item Not Required 1

Item Not Required 2