Data-Adaptive Estimation for Double-Robust Methods in Population-Based Cancer Epidemiology: Risk differences for lung cancer mortality by emergency presentation

Miguel Angel Luque-Fernandez, Aurélien Belot, Linda Valeri, Giovanni Ceruli, Camille Maringe, and Bernard Rachet

Correspondence to Dr. Miguel Angel Luque-Fernandez, Department of Non-Communicable Disease Epidemiology, Cancer Survival Group, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT (e-mail: Miguel-angel.luque@lshtm.ac.uk)

Authors affiliations: Faculty of Epidemiology and Population Health, Department of Non-Communicable Disease Epidemiology, Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, U.K. (Miguel Angel Luque-Fernandez, Aurélien Belot, Camille Maringe, and Bernard Rachet); Psychiatric Biostatistics Laboratory, MacLean Hospital, Belmont, U.S. (Linda Valeri); Harvard Medical School, Harvard University, Boston, U.S. (Linda Valeri) National Research Council of Italy, Institute for Research on Economic Sustainable Growth, Rome, Italy (Giovanni Ceruli).

This work was funded by Cancer Research U.K. grant number C7923/A18525.

Conflict of interest: none declared.

Running head: Double-Robust Estimation in Observational Population-Based Cancer Epidemiology

Abbreviations list
AIPTW: Augmented inverse-probability of treatment weighting.
ATE: Average treatment effect.
DAG: Directed acyclic graph.
IPTW-RA: Inverse-probability treatment weighted regression-adjustment.
RMSE: Root mean squared error.
TMLE: Targeted maximum likelihood estimation.

ABSTRACT

We propose a structural framework for population-based cancer epidemiology and evaluate the performance of double-robust estimators for a binary exposure in cancer mortality. We performed numerical analyses to study the bias and efficiency of these estimators. Furthermore, we compared two different model selection strategies based on i) the Akaike and Bayesian Information Criteria and ii) machine-learning algorithms, and illustrated double-robust estimators' performance in a real setting. In simulations with correctly specified models and near-positivity violations, all but the naïve estimators presented relatively good performance. However, the augmented inverse-probability treatment weighting estimator showed the largest relative bias. Under dual model misspecification and near-positivity violations, all double-robust estimators were biased. Nevertheless, the targeted maximum likelihood estimator showed the best bias-variance trade-off, more precise estimates, and appropriate 95% confidence interval coverage, supporting the use of the data-adaptive model selection strategies based on machine-learning algorithms. We applied these methods to estimate adjusted one-year mortality risk differences in 183,426 lung cancer patients diagnosed after admittance to an emergency department versus non-emergency cancer diagnosis in England, 2006-2013. The adjusted mortality risk (for patients diagnosed with lung cancer after admittance to an emergency department) was 16% higher in men and 18% higher in women, suggesting the importance of interventions targeting early detection of lung cancer signs and symptoms.

Keywords: causality; cancer epidemiology; population-based data; statistics; machine learning; targeted maximum likelihood estimation

Data from population-based cancer registries are critical for cancer control and policy.[1-3] However, the scope of the information from cancer registries refers to cancer characteristics and basic socio-demographic factors.[1, 2, 4] Recently, linkage strategies of population-based data sets from different sources have been implemented. This has allowed for more advanced modelling scenarios regarding applications in cancer policy and control.[5-10] For instance, comparative effectiveness approaches using medical records and linked population-based databases are used to evaluate the effectiveness of treatment or exposures concerning cancer mortality and survival.[6-10] Nevertheless, the evaluation of the effectiveness of treatments or exposures in a large population-based cancer epidemiology requires well-defined structural frameworks and modern statistical methods in order to overcome confounding.[9]

The use of the Neyman-Rubin potential outcomes framework[11] allows researchers to make explicit the assumptions under which an observed association from observational studies can be interpreted causally. For a given factor to be considered causal, researchers must consider a set of additional assumptions (i.e., conditional exchangeability, positivity and consistency).[12] Directed acyclic graphs (DAGs) help to evaluate whether, under a given causal model, the counterfactual outcome is independent of the observed exposure given some sets of covariates (conditional exchangeability) selected on the basis of subject matter knowledge.[12-14]

The average treatment effect (ATE) or risk difference is a commonly used parameter of interest.[12, 15, 16] Correct model specification is crucial to obtain unbiased estimates of the true ATE. Many estimators of the ATE (but not all) rely on parametric modeling assumptions, thereby introducing bias when the model is incorrect.[15] Researchers have developed double-robust estimation procedures to reduce bias due to misspecification.[17, 18] More recently, van der Laan has developed a targeted maximum likelihood estimation using machine learning algorithms to minimize the risk of model misspecification.[15, 19, 20] Simulations studies using targeted maximum likelihood estimation (TMLE) in finite samples

provide evidence of its double-robust properties and gains in performance when combined with machine learning algorithms.[15, 21, 22]

However, there is no evidence evaluating the performance of TMLE compared with other double-robust methods in the setting of population-based cancer epidemiology. We sought to compare the performance of three different double-robust causal estimators of the ATE for cancer mortality in a simulated scenario with forced near-positivity violations (i.e., certain subgroups in the sample rarely or never receive treatment) and model misspecification. Furthermore, we studied the efficiency and bias of double-robust estimators and compared two different model selection strategies based on i) a combination of Akaike-Bayesian information criteria (AIC-BIC) and ii) machine learning algorithms and TMLE. Finally, these methods are illustrated with real population-based data on lung cancer patients in England.

METHODS

Counterfactual framework

Based on background knowledge, we used a DAG to depict our general counterfactual framework (Figure 1). We considered one-year cancer mortality as a binary outcome Y and a generic binary exposure or treatment A, and we assumed that the following measured covariates were sufficient to ensure conditional exchangeability: patients' socioeconomic status (W_1), age (W_2), cancer stage (W_3), and comorbidities at diagnosis (W_4) (Figure 1). Afterward, based on our DAG, we generated data to explore the effects of near-positivity violations and dual misspecification (outcome and treatment models). The set of covariates included in W is critical for cancer treatment decision-making.[3, 23, 24] However, cancer stage and patients' comorbidities at diagnosis play a crucial role in clinical treatment choice and have been cited as the most important explanatory factors for cancer mortality and survival. [3, 23, 24] As depicted in our DAG, we highlighted the importance of patients' cancer stage, socioeconomic status, and comorbidities as the minimum set needed to assume conditional exchangeability based on the back-door criterion. Our targeted

parameter was the one-year risk differences on cancer mortality for patients exposed to a generic exposure (A) versus non-exposed patients.

Data generation process and Monte Carlo simulations

We generated data based on the structural framework represented in Figure 1 by a DAG. The covariates (W) were drawn using a set of random uniform and binomial variables. The propensity score for the binary exposure (A) and the outcome variable (Y) were derived from a binomial logit model that included the interaction between age (W_2) and comorbidities (W_4) for the generation of Y.

Afterward, we drew 1,000 replications from the data-generation process with sample sizes of 1,000 and 10,000. In each replication, we estimated the binary ATE and recorded the point estimates and standard errors based on the influence curve in order to calculate the ATE standard deviations, bias, 95% confidence interval coverage and root mean squared error (RMSE).[25]

Model estimation scenarios and performance evaluation

We set two different modeling scenarios aiming to assess the performance of double-robust estimators of the ATE using: i) correctly specified models for the treatment and the outcome, and ii) misspecified models for both treatment and outcome. Correctly specified models for the treatment and outcome models included socioeconomic status (W_1), age (W_2), cancer stage (W_3), and comorbidities (W_4) as covariates. Model misspecification for the treatment and the outcome was forced omitting the interaction between comorbidities (W_4) and age (W_2). Data-adaptive approaches were used to estimate the treatment and outcome for misspecified models (Web Appendix 1 describes in more detail the model specifications for the data generation). For both scenarios, we included near-positivity violations that forced some values of the propensity score distribution close to zero. Near-positivity violations were evaluated visually based on the summary of the propensity score distribution. Figure 2

illustrates the overlap of the distribution of the potential outcomes for one simulated sample in the first scenario (Figure 2A), and second scenario (Figure 2B).

In the first scenario, which uses correctly specified models, we evaluated the performance of a classical multivariate regression adjustment with treatment (A) and covariates (W_1-W_4) as predictors of the outcome (Y), namely the naïve approach, and of three different double-robust estimators of the ATE: i) inverse-probability treatment weighted regression-adjustment (IPTW-RA),[26] ii) augmented inverse-probability treatment weighting (AIPTW) [17, 27, 28] and iii) TMLE.[15, 29] IPTW-RA is a regression model weighted by the inverse probability of treatment whereas AIPTW is a two-step procedure with two estimating equations for the treatment and mean outcome, respectively.[27]

For the second scenario, using misspecified models, we evaluated two different data-adaptive model selection strategies in combination with the above described double-robust estimators. Models for the treatment and outcome included the above-described covariates for the first scenario but omitted the interaction between comorbidities and age used to generate the data in the second scenario. (Web Appendix 1 describes in more detail the model specifications for the data generation.) As data-adaptive strategies, we used AIC-BIC approaches for the IPTW-RA and AIPTW estimators, and ensemble-learning for the TMLE estimator. For the IPTW-RA, we used the AIC-BIC based approach implemented in the Stata user-written command "bfit" (best fit).[30] The bfit algorithm sorts a set of fitted candidate regression models using the Akaike and Bayesian Information Criteria and displays a table showing the ranking of the models. Each linear predictor of the candidate models is defined as a linear combination of functional forms of the variables. The smallest of the candidate models includes only one variable. The largest of the candidate models includes all the variables in a fully interacted polynomial of the order prespecified by the user. We set the order to "2" for comparative purposes with TMLE. For simulations and analysis of the IPTW-RA and AIPTW estimators, we used Stata v.14.1 (StataCorp, College Station, TX, U.S.) and the teffects ipwra and teffects aipw commands.[26]

The TMLE estimator has not been implemented in Stata statistical software yet, so we used the package tmle (version 1.2.0-4) [29] from the statistical software R version 3.0.2 (R Foundation for Statistical Computing, Vienna, Austria). The implementation of TMLE in R calls the Super-Learner package. The Super-Learner uses V-fold (10-fold by default) cross-validation to assess the performance of the prediction of the outcome and the propensity score models as weighted averages (ensemble-learning) of a set of machine learning algorithms.[29, 31] We used the default specifications of the tmle package, which included the following machine-learning algorithms: i) stepwise forward and backward selection; ii) generalized linear modeling (glm) with the covariates (W) and the treatment (A) as main terms; iii) a glm variant that included second order polynomials and two-by-two interactions of the main terms included in the model. In Web Appendix 2, we provide a basic implementation of the TMLE algorithm in both Stata and R statistical software as well as the link to a testing version of TMLE implemented in Stata.

Monte Carlo simulation results

First Scenario: correctly specified models and near-positivity violation.

The true risk difference of the ATE estimate from the 1,000 simulation repetitions was -18%. The naïve approach showed a biased estimate of the ATE with an overestimation of the treatment effect by 23% (relative bias). All double-robust estimators were nearly unbiased showing smaller RMSE with increasing sample size, but the TMLE presented higher precision (based on the difference in variances between estimators), the smallest RMSE, and the best coverage (95%) (Table 1, first scenario: correctly specified models).

Second scenario: misspecification, near-positivity violation and adaptive model selection.

The true risk difference of the ATE from the 1,000 simulation repetitions was -12%. The naïve approach was heavily biased, showing the highest RMSE with an underestimation of the treatment effect by approximately 90% (Table 1, second scenario: adaptive estimation

approach). The model selection strategy based on AIC-BIC did not show either bias reduction or coverage improvement. The double-robust TMLE estimator presented the best performance with more precise estimates (1% bias for a sample size of 1,000 and less than 0.5% for a sample size of 10,000 patients) and the highest coverage. By contrast, the relative bias increased with larger sample size for the AIPTW estimator using the AIC-BIC approach. The relative bias ranged from 1.5% (n= 1,000) to 11.7% (n= 10,000). (Table 1, second scenario: adaptive estimation approach)

ILLUSTRATION

Under the structural framework (DAG, Figure 1) described above for population-based cancer epidemiology, we estimated one-year adjusted mortality risk differences for cancer diagnosed after admittance to a hospital emergency department versus a non-emergency cancer diagnosis. The high proportion of lung cancer diagnosed after admittance to an emergency department observed in the UK (emergency presentation) has been hypothesized to be mainly due to multiple steps that patients undergo between the identification of the first symptoms and the final diagnosis by the healthcare system.

In addition to age and socioeconomic status, we included comorbidities and cancer stage as confounders. Evidence shows that the presence of patient comorbidity increases the odds of being diagnosed with distant metastases (advanced cancer stage), and it does not lead to an earlier cancer diagnosis.[32] Socioeconomic status was measured using quintiles of the income domain of the Index of Multiple Deprivation in England,[33] comorbidities were measured based on the Charlson comorbidity index, [34] and stage was based on the tumor, node, and metastases classification of malignant tumors.[35] In England, a cancer diagnosis after emergency presentation correlates closely with poor one-year survival. However, the strength of the evidence comes from observational data and is weak, owing to confounding.[36]

It is of public health interest to estimate the one-year adjusted mortality risk differences of cancer diagnosed after an emergency presentation, given the potential impact of a preventive intervention aiming to improve earlier cancer diagnosis. Quantifying the gender-specific adjusted risk differences for one-year mortality for lung cancer patients will reinforce the current evidence and help to promote the policy actions required for improving early cancer diagnoses.

To illustrate the estimation of the adjusted risk differences for one-year mortality, we extracted the data from the National Cancer Data Repository for 183,426 incident cases of lung cancer diagnosed between 2006 and 2013 in England, which consisted of 102,535 men and 80,891 women. All patients had a minimum potential follow-up of one year since the vital status was not assessed until December 31st, 2014. Data were restricted to cases with complete information on sex, age at diagnosis, comorbidities, cancer stage, socioeconomic deprivation, and type of cancer diagnosis. The strategy for the assessment of cancer diagnosis after an emergency presentation has been previously described elsewhere.[37] Overall, more than 80% of the patients who died within one year after a cancer diagnosis had been diagnosed after an emergency presentation, and only 96 (representing 0.05%) were lost to follow-up before one year (Web Table 1). The average age at diagnosis was 72 years in men and 73 in women. One-year mortality after diagnosis presented a balanced distribution across the different age and socioeconomic groups and by quartiles of the Charlson comorbidity index.[34] However, stages IV and III presented with 4- and 3-fold higher probabilities for one-year mortality, respectively, than stage I (Table 2).

To estimate the adjusted mortality risk difference, we used the same approaches and commands used for the simulation study. We provide commented code for the illustration in Web Appendix 2. Overall, based on double-robust estimators, we estimated that the adjusted risk of one-year mortality between cancer diagnosed after admittance to an emergency department versus non-emergency diagnosis based on double-robust estimators

was 16% higher in men and 18% higher in women than it was after non-emergency diagnosis. However, the naïve approach showed the largest risk difference with 29% and 32% adjusted risk differences for women and men, respectively (Figure 3: 3A women; 3B men).

We also used the observed covariates from the illustration to run 100 Monte Carlo simulations to estimate the adjusted mortality risk difference for one-year cancer mortality after admittance to an emergency department. Using the information on baseline covariates from the observed data, we simulated only the outcome and treatment models. To evaluate the performance of the different estimators under strong near-positivity violations, we forced some values of the propensity scores close to zero (Web Figure 1). However, the estimation models for the treatment and outcome were correctly specified during simulations to include the interaction between age and comorbidities (we provide the model specifications and the variables included for the simulations in Web Appendix 1). The propensity score distributions among the exposed and unexposed overlapped considerably in the real setting (Web Figure 1A) while the overlap in the simulated scenario was poor given the strong near-positivity violation (Web Figure 1B). Table 3 presents the results of the simulations, which validate the previous results with similar findings, but with a larger sample size and fixed covariates coming from a real scenario, thus reproducing reality much better. TMLE presented the best precision and coverage and outperformed all other double-robust estimators. By contrast, AIPTW showed high sensitivity to the violation of the positivity assumption with a relative bias of 8% (Table 3).

DISCUSSION

Given the increasing availability of a different range and variety of data in population-based cancer epidemiology, the proposed structural framework (DAG, Figure 1) constitutes a basis for further development of comparative effectiveness research in population-based cancer

epidemiology. Developed for a binary treatment and outcome, the framework can be easily extended to handle time-to-event outcomes, and might be adapted to specific comparative effectiveness scenarios. For instance, we considered cancer patients' comorbidities and stage as confounders, but it might not be the case in other comparative effectiveness research questions. We have recently published an article where we argue that multivariate adjustment for cancer-related comorbidities (those with onset date close before or after the date of cancer diagnosis), to evaluate the effectiveness of cancer treatment, might be inappropriate as it could induce collider stratification bias.[38]

We also applied the proposed structural framework (DAG, Figure 1) to a real data scenario and highlighted the critical importance of considering cancer stage and patients' comorbidities in the structural framework to satisfy the conditional exchangeability assumption in population-based cancer epidemiology. Conventional methods control for confounding by assuming that the effect measure of the exposure of interest is constant across all levels of the covariates included in the model.[39] We provided evidence of highly imprecise estimates of ATE in the classical naïve regression method, underestimating the effect of the treatment, particularly for the misspecified model in the simulation setting.

Model misspecification with parametric modelling is always a concern in epidemiologic research. ATE estimators based on the propensity score or regression adjustment are unbiased only if estimation models are correctly specified.[17, 27, 40] Double-robust estimation combines these two approaches so that only one of the two models needs to be correctly specified to obtain an unbiased estimate of the ATE.[17, 27, 40] Previous simulation studies have shown that double-robust methods, including TMLE, consistently provide almost unbiased estimates when either the propensity score or the outcome model is misspecified but the other is correct.[41-43] However, more evidence is needed to evaluate TMLE statistical properties under different modeling scenarios.

TMLE is a general algorithm that can estimate the g-formula[44] as a generalization of standardization defining the parameters of interest semi-parametrically as a function of the data-generating distribution. TMLE evaluates the target parameter (ATE) by using a double-robust semi-parametric substitution estimation based on machine learning algorithms to avoid misspecification and reduce bias.[22]

Our results showed that, when the models were correctly specified, standardization implemented through the IPTW-RA, AIPTW, and TMLE provided nearly unbiased estimates of ATE, despite near-positivity violations. TMLE, however, was the most efficient estimator. Nevertheless, dual misspecification is the likely scenario in population-based cancer epidemiology; thus, attempting to obtain the best possible estimates is paramount for policy recommendations. Under dual misspecification and near-positivity violations, both in simulations and a real-life illustration, AIPTW showed poorer performance than IPTW-RA and TMLE, illustrating the instability of the AIPTW to estimate values of the propensity score close to zero (near-positivity violations) as previously reported by Kang and Shafer.[27] However, basic machine-learning algorithms and ensemble-learning techniques implemented in the tmle and Super-Learner R-packages avoid misspecification of the models (either for the treatment or the outcome) used to estimate the ATE.

To the best of our knowledge, the performance of double-robust methods using different model selection strategies has not been evaluated in the context of adverse estimation situations with a near-violations of the positivity assumption and misspecified models. Based on a simulated scenario, we compared the Stata user-written program bfit,[30] with machine and ensemble-learning algorithms implemented in the R package tmle based on the Super-Learner.[29, 45] TMLE outperformed model selection strategies based on AIC-BIC for the IPTW-RA and the AIPTW estimators. By default, TMLE implementation in R sets a bounded distribution of the propensity score to 0.025 and 0.975, and the adaptive estimation respects the limits of the possible range of the targeted parameter, but AIPTW does not. So, AIPTW

could, for instance, produce estimates that are outside the range of the targeted parameter. Moreover, the default AIPTW implementation in Stata will not converge for very small values of the propensity score with a tolerance set by default to 10^{-5}. We had to increase the tolerance of the weights for the propensity score to 10^{-8} when using the AIC-BIC adaptive approach (Stata bfit) for the AIPTW estimator, given convergence problems associated with the near-positivity violations. The relative bias using an adaptive approach based on AIC-BIC for the estimation of the AIPTW under difficult scenarios increases with a larger sample size (from 1,000 to 10,000 in our simulations setting). Hence, using AIC-BIC for the AIPTW estimator might not be a good option when there is a strong suspicion of model misspecification and near-violation of the positivity assumption. Further evidence is needed to evaluate our findings.

However, AIPTW performance is similar to IPTW-RA and TMLE under certain scenarios (correct specification and without near-positivity violations). TMLE is computationally demanding, manifesting in slow run times for large cancer population data (e.g. using a computer with 4 cores and 16 GB of memory, the R-package tmle took 5.4 minutes to estimate the ATE for 10,000 patients using more advanced machine learning algorithms such as generalized additive models, random forests, and boosting).

Under an adverse estimation scenario, with near-positivity violations and dual misspecification, the TMLE estimator of the ATE for a binary treatment and outcome performs better than other double-robust estimators. Its reductions in bias and gains in efficiency supporting the use of TMLE for a binary treatment and outcome in population-based cancer epidemiology research. Results from the illustration provide quantitative evidence of an increased one-year mortality risk in patients diagnosed with lung cancer after attending a hospital emergency department, which should boost calls for policy interventions such as the implementation of the multidisciplinary diagnosis centers to improve early cancer diagnosis and management.

ACKNOWLEDGMENTS

We would like to thank Dr. Mark van der Laan and Dr. Michael Schomaker for their insightful comments and suggestions.

Portions of this work were presented at the U.K. Causal Inference meeting, April 2016, London, U.K.

Conflict of interest: none declared.

Figure 1. Directed acyclic graph for the proposed structural causal framework in population-based cancer research.Conditional exchangeability of the treatment effect or exposure (A) on one-year cancer mortality (Y) is obtained through conditioning on a set of available covariates (Y1,Y0 ⊥ A|W). The minimum sufficient set, based on the back-door criterion, is obtained through conditioning on only W_1, W_3, and W_4. The average treatment effect for the structural framework is estimated as the average risk difference between the expected effect of the treatment conditional on W among those treated (E(Y|A=1; W)) and the expected effect of the treatment conditional on W among those untreated (E(Y|A=0; W)). W: W_1: socioeconomic status; W_2: age; W_3: cancer stage; W_4: comorbidities

Figure 2. Overlap of the propensity score for correctly specified (first scenario) and misspecified models (second scenario).

Figure 3. Gender-specific adjusted risk difference of one-year lung cancer mortality by different double-robust estimators between 2006 and 2013 in England. Risk difference in 183,426 lung cancer patients diagnosed after admittance to an emergency department versus non-emergency cancer diagnosis in England, 2006-2013. A: Women; B: Men; A-IPTW: Augmented inverse-probability of treatment weighting; BF-AIPTW: Best fit augmented inverse-probability treatment weighting (data-adaptive estimation based on AIC-BIC); IPTW-RA: Inverse-probability treatment weighted regression-adjustment; BF-IPTW-RA: Best fit inverse-probability treatment weighted regression-adjustment (data-adaptive estimation based on AIC-BIC); TMLE: Targeted maximum likelihood estimation (data adaptive estimation based on ensemble learning and k-fold cross-validation)

REFERENCES

1. Allemani, C., et al., *Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2).* Lancet, 2015. 385(9972): p. 977-1010.
2. Parkin, D.M., *The role of cancer registries in cancer control.* Int J Clin Oncol, 2008. 13(2): p. 102-111.
3. Rachet, B., et al., *Socioeconomic inequalities in cancer survival in England after the NHS cancer plan.* Br J Cancer, 2010. 103(4): p. 446-453.
4. Siesling, S., et al., *Uses of cancer registries for public health and clinical research in Europe: Results of the European Network of Cancer Registries survey among 161 population-based cancer registries during 2010-2012.* Eur J Cancer, 2015. 51(9): p. 1039-1049.
5. Andersson, K., et al., *The interface of population-based cancer registries and biobanks in etiological and clinical research--current and future perspectives.* Acta Oncol, 2010. 49(8): p. 1227-1234.
6. Giordano, S.H., *Comparative effectiveness research in cancer with observational data.* Am Soc Clin Oncol Educ Book, 2015: p. e330-5.
7. Chen, V.W., et al., *Enhancing cancer registry data for comparative effectiveness research (CER) project: overview and methodology.* J Registry Manag, 2014. 41(3): p. 103-112.
8. Mack, C.D., et al., *Calendar time-specific propensity scores and comparative effectiveness research for stage III colon cancer chemotherapy.* Pharmacoepidemiol Drug Saf, 2013. 22(8): p. 810-818.
9. Carpenter, W.R., et al., *A framework for understanding cancer comparative effectiveness research data needs.* J Clin Epidemiol, 2012. 65(11): p. 1150-1158.
10. Glasgow, R.E., *Commentary: Electronic health records for comparative effectiveness research.* Med Care, 2012. 50 Suppl: p. S19-20.

11. Little, R.J. and D.B. Rubin, *Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches.* Annu Rev Public Health, 2000. 21: p. 121-145.

12. Imbens, G. and D.B. Rubin, *Causal inference for statistics, social, and biomedical sciences : an introduction*. 2015, New York, NY: Cambridge University Press. xix, 625 p.

13. Greenland, S. and J.M. Robins, *Identifiability, exchangeability, and epidemiological confounding.* International journal of epidemiology, 1986. 15(3): p. 413--419.

14. Pearl, J., *Causality : models, reasoning, and inference*. 2nd ed. 2009, New York, NY: Cambridge University Press. xix, 464 p.

15. Laan, M.J.v.d. and S. Rose, *Targeted learning : causal inference for observational and experimental data*. 2011, New York, NY: Springer Verlag. lxxi, 626 p.

16. Imai, K., G. King, and E.A. Stuart, *Misunderstandings between experimentalists and observationalists about causal inference.* Journal of the Royal Statistical Society Series a-Statistics in Society, 2008. 171: p. 481-502.

17. Bang, H. and J.M. Robins, *Doubly robust estimation in missing data and causal inference models.* Biometrics, 2005. 61(4): p. 962-973.

18. *Robins J.M., Rotnitzky A. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". Stat Sinica, 11(4): 920–936, 2001.*

19. van der Laan, M.J., *Targeted Maximum Likelihood Based Causal Inference: Part II.* International Journal of Biostatistics, 2010. 6(2).

20. van der Laan, M.J., *Targeted Maximum Likelihood Based Causal Inference: Part I.* International Journal of Biostatistics, 2010. 6(2).

21. Porter, K.E., et al., *The Relative Performance of Targeted Maximum Likelihood Estimators.* International Journal of Biostatistics, 2011. 7(1).

22. Schuler, M.S. and S. Rose, *Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies.* Am J Epidemiol, 2017. 185(1): p. 65-73.

23. Sarfati, D., B. Koczwara, and C. Jackson, *The impact of comorbidity on cancer and its treatment.* CA Cancer J Clin, 2016. 66(4): p. 337-350.

24. Woods, L.M., B. Rachet, and M.P. Coleman, *Origins of socio-economic inequalities in cancer survival: a review.* Ann Oncol, 2006. 17(1): p. 5-19.

25. Burton, A., et al., *The design of simulation studies in medical statistics.* Stat Med, 2006. 25(24): p. 4279-4292.

26. StataCorp. Stata 13 Treatment Effects Manual: Potential Outcomes/Counterfactual Outcomes. Stata Press: College Station, Texas. U.S.A, 2014.

27. Kang, J.D.Y. and J.L. Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.* Statistical Science, 2007. 22(4): p. 523-539.

28. Robins, J.M., Rotnitzky A., Zhao L.P. *Estimation of regression coefficients when some regressors are not always observed.* Journal of the American statistical Association 89.427 (1994): 846-866.

29. Gruber, S. and M.J. van der Laan, *tmle: An R Package for Targeted Maximum Likelihood Estimation.* Journal of Statistical Software, 2012. 51(13): p. 1-35.

30. Cattaneo, M.D., D.M. Drukker, and A.D. Holland, *Estimation of multivalued treatment effects under conditional independence.* Stata Journal, 2013. 13(3): p. 407-450.

31. Pirracchio, R., et al., *Fitting Icu Data Complexity Need for Innovative Prediction Tools Mortalilty Prediction Using Superlearner.* Intensive Care Medicine, 2013. 39: p. S222-S222.

32. Gurney, J., D. Sarfati, and J. Stanley, *The impact of patient comorbidity on cancer stage at diagnosis.* Br J Cancer, 2015. 113(9): p. 1375-1380.

33. Department for Communities and Local Government Publications. *The English Indices of Deprivation, 2007.* London. Communities and Local Government Publications. 2008 (publication no. 07 NRAD 05137)

34. Charlson, M.E., et al., *A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.* J Chronic Dis, 1987. 40(5): p. 373-383.

35. Sobin, L.H., et al., *TNM classification of malignant tumours*. 7th ed. 2010, Chichester, West Sussex, UK ; Hoboken, NJ: Wiley-Blackwell. xx, 310 p.

36. Tataru, D., et al., *The effect of emergency presentation on surgery and survival in lung cancer patients in England, 2006-2008.* Cancer Epidemiol, 2015. 39(4): p. 612-616.

37. Elliss-Brookes, L., et al., *Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets.* Br J Cancer, 2012. 107(8): p. 1220-1226.

38. Maringe, C., et al., *Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities.* PLoS One, 2017. 12(3): p. e0172814.

39. Keil, A.P., et al., *The parametric g-formula for time-to-event data: intuition and a worked example.* Epidemiology, 2014. 25(6): p. 889-897.

40. Emsley, R., et al., *Implementing double-robust estimators of causal effects.* Stata Journal, 2008. 8(3): p. 334-353.

41. Kreif, N., et al., *Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching.* Stat Methods Med Res, 2014;25(5): 2315-2336

42. Leon, S., A.A. Tsiatis, and M. Davidian, *Semiparametric estimation of treatment effect in a pretest-posttest study.* Biometrics, 2003. 59(4): p. 1046-1055.

43. Lunceford, J.K. and M. Davidian, *Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study.* Stat Med, 2004. 23(19): p. 2937-2960.

44. Robins, J., *A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period - Application to Control of the Healthy Worker Survivor Effect.* Mathematical Modelling, 1986. 7(9-12): p. 1393-1512.

45. van der Laan, Mark J., Eric C Polley and Alan E. Hubbard. *Super Learner.* Statistical Applications in Genetics and Molecular Biology, 2007; 6: Article25. Epub 2007.

Table 1. Monte Carlo simulations (10,000) of the ATE for correctly specified models (first scenario) and misspecified models using adaptive approaches (second scenario) for different double-robust estimators.

	ATE		Absolute BIAS		Relative BIAS (%)		RMSE		SD-ATE		95%CI coverage (%)	
First scenario[a]	N=1,000	N=10,000	N=1,000	N=10,000	N=1,000	N=10,000	N=1,000	N=10,000	N=1,000	N=10,000	N=1,000	N=10,000
True ATE	-0.1813											
Naïve	-0.2234	-0.2218	0.0421	0.0405	23.2	22.3	0.0575	0.0423	0.0391	0.0123	77	89
AIPTW	-0.1843	-0.1848	0.0030	0.0035	1.6	1.9	0.0534	0.0180	0.0533	0.0177	93	94
IPTW-RA	-0.1831	-0.1838	0.0018	0.0025	1.0	1.4	0.0500	0.0174	0.0500	0.0172	91	95
TMLE	-0.1832	-0.1821	0.0019	0.0008	1.0	0.4	0.0482	0.0158	0.0482	0.0158	95	95
Second scenario[b]												
True ATE	-0.1172											
Naïve	-0.0127	-0.0121	0.1045	0.1051	89.2	89.7	0.1470	0.1100	0.1034	0.0326	0	0
BFit AIPTW	-0.1155	-0.0920	0.0017	0.0252	1.5	11.7	0.0928	0.0773	0.0928	0.0731	65	65
BFit IPTW-RA	-0.1268	-0.1192	0.0096	0.0020	8.2	1.7	0.0442	0.0305	0.0431	0.0305	52	73
TMLE	-0.1181	-0.1177	0.0009	0.0005	0.8	0.4	0.0281	0.0107	0.0281	0.0107	93	95

a: First scenario: correctly specified models and near-positivity violation

b: Second scenario: misspecification, near-positivity violation and adaptive model selection

AIPTW: Augmented Inverse-Probability Treatment Weights; ATE: Average treatment effect across 1,000 simulated data sets; BFit IPTW-RA: Best fit based on AIC and BIC criteria inverse-Probability treatment weighted regression-adjustment; BFit AIPTW: Best fit based on AIC and BIC criteria augmented Inverse-Probability Treatment Weights; IPTW-RA: Inverse-Probability treatment weighted regression-adjustment; RMSE: Root mean square error; SD: Standard deviation; TMLE: Targeted Maximum Likelihood Estimation calling basic Super-Learner libraries (SL): SL. Step; SL.glm; SL.glm.interaction

Table 2. One-year mortality in lung cancer patients (incident cases) by stage, comorbidities, age, socioeconomic status, and cancer diagnosis after admittance to an emergency department versus non-emergency in England between 2006 and 2013, n = 183,426 (males: 102,535; females: 80,891).

	Mortality one year after diagnosis	
Variables	Female, deaths (%)	Male, deaths (%)
Emergency presentation		
No	53.4	59.9
Yes	83.7	86.4
Stage		
I	18.1	24.2
II	35.1	37.6
III	58.6	62.4
IV	82.2	85.8
Quartiles Charlson index		
Q1	62.8	67.6
Q2	64.1	68.3
Q3	67.2	71.4
Q4	72.4	75.5
Socioeconomic Status		
Q1	62.6	66.7
Q2	63.3	68.1
Q3	64	69.5
Q4	64.2	69.6
Q5	64.1	68.2
Age at diagnosis (mean, sd)	73.0 (10.8)	72.6 (10.3)

Table 3. Monte Carlo simulation of the risk differences of one-year mortality in lung cancer patients (incident cases) diagnosed after admittance to an emergency department between 2006 and 2013 in England, n = 183,426.

Estimators	ATE	Absolute BIAS	Relative BIAS (%)	RMSE	SD-ATE	95%CI Coverage (%)
True ATE	0.1621					
AIPTW	0.1493	0.0128	7.9	0.0165	0.0104	79
IPTW-RA	0.1587	0.0034	2.1	0.0072	0.0063	92
TMLE	0.1620	0.0001	0.1	0.0034	0.0034	92

AIPTW: Augmented Inverse-Probability Treatment Weights; ATE: Average treatment effect across 1,000 simulated data sets; IPTW-RA: Inverse-Probability treatment weighted regression-adjustment; RMSE: Root mean square error; SD: Standard deviation; TMLE: Targeted Maximum Likelihood Estimation calling basic Super-Learner libraries (SL): SL. Step; SL.glm; SL.glm.interaction