Journal of Clinical Epidemiology 152 (2022) 269–280

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Causal analyses with target trial emulation for real-world evidence removed large self-inflicted biases: systematic bias assessment of ovarian cancer treatment effectiveness

Felicitas Kuehne[a], Marjan Arvandi[a], Lisa M. Hess[b], Douglas E. Faries[b], Raffaella Matteucci Gothe[a], Holger Gothe[a,c], Julie Beyrer[b], Alain Gustave Zeimet[d], Igor Stojkov[a], Nikolai Mühlberger[a], Willi Oberaigner[a,e], Christian Marth[d], Uwe Siebert[a,f,g,*]

[a]*Department of Public Health, Health Services Research and Health Technology Assessment, Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT TIROL – University for Health Sciences, Medical Informatics and Technology, Hall i.T., Austria*

[b]*Eli Lilly and Company, Indianapolis, IN, USA*

[c]*Chair of Health Sciences/Public Health, Medical Faculty "Carl Gustav Carus", Technical University Dresden, Dresden, Germany*

[d]*Department of Obstetrics and Gynecology, Innsbruck Medical University, Innsbruck, Austria*

[e]*Institute for Clinical Epidemiology, Cancer Registry Tyrol, Tirol Kliniken, Innsbruck, Austria*

[f]*Center for Health Decision Science and Departments of Epidemiology and Health Policy & Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA*

[g]*Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA*

Accepted 3 October 2022; Published online 15 October 2022

Abstract

Background and Objectives: Drawing causal conclusions from real-world data (RWD) poses methodological challenges and risk of bias. We aimed to systematically assess the type and impact of potential biases that may occur when analyzing RWD using the case of progressive ovarian cancer.

Methods: We retrospectively compared overall survival with and without second-line chemotherapy (LOT2) using electronic medical records. Potential biases were determined using directed acyclic graphs. We followed a stepwise analytic approach ranging from crude analysis and multivariable-adjusted Cox model up to a full causal analysis using a marginal structural Cox model with replicates emulating a reference randomized controlled trial (RCT). To assess biases, we compared effect estimates (hazard ratios [HRs]) of each approach to the HR of the reference trial.

Results: The reference trial showed an HR for second line vs. delayed therapy of 1.01 (95% confidence interval [95% CI]: 0.82–1.25). The corresponding HRs from the RWD analysis ranged from 0.51 for simple baseline adjustments to 1.41 (95% CI: 1.22–1.64) accounting for immortal time bias with time-varying covariates. Causal trial emulation yielded an HR of 1.12 (95% CI: 0.96–1.28).

Conclusion: Our study, using ovarian cancer as an example, shows the importance of a thorough causal design and analysis if one is expecting RWD to emulate clinical trial results. © 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Causal inference; Comparative effectiveness; Longitudinal data; Electronic health records; Target trial; Inverse probability weighting

Funding: This work was supported by independent research grant by Lilly Research Award Program (LRAP) and partly funded by the state Tyrol (grant number F.16731/5-2019).

Ethics Approval: This is an observational study using anonymized data. The Research Committee for Scientific and Ethical Questions (RCSEQ) of UMIT-Tirol has confirmed that no ethical approval is required.

Conflict of interest: The following authors were salaried employees of Eli Lilly and Company at the time of conducting the analyses: Lisa M Hess, Douglas E Faries, and Julie Beyrer.

Author contribution: All authors contributed to the study conception and design. Material preparation, data analysis were performed by Felicitas Kuehne, Marjan Arvandi, Lisa M Hess, Douglas E Faries, Raffaella Matteucci Gothe, Julie Beyrer, and Uwe Siebert. The first draft of the manuscript was written by Felicitas Kuehne and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

* Corresponding author. Department of Public Health, Health Services Research and Health Technology Assessment, Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT TIROL – University for Health Sciences, Medical Informatics and Technology, Eduard-Wallnoefer-Zentrum 1, A-6060 Hall i.T., Austria, Tel.: +43-0-50 8648 3930; fax: +43-0-50 678648.

E-mail address: usiebert@hsph.harvard.edu (U. Siebert).

https://doi.org/10.1016/j.jclinepi.2022.10.005

0895-4356/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

What is new?

- This conceptual paper using a real-world case example offers a comprehensive overview of potential biases that may occur in real-world data (RWD) analysis and provides a summary for causal inference tools and potential adjustment methods including causal graphs, target trial emulation, and g-methods.
- This bias assessment demonstrates that self-inflicted biases can be avoided by using causal frameworks and that residual (unmeasured) confounding may contribute much less to bias than often suspected when using real-world data. Therefore, confidence in observational studies using appropriate innovative causal analytic methods − if applied correctly and completely − may increase. This work also underlines the need for carefully carefully designing observational studies based on RWD and the importance of the target trial approach, which is now also taken up by health technology assessment agencies in Europe.

What this adds to what was known?

- We systematically assessed type, direction and magnitude of potential biases in real-world observational data analysis by applying a stepwise analytic approach ranging from simple crude analysis, over traditional adjustment methods, to full causal analyses with target trial emulation and comparing results to a reference randomized controlled trial (RCT).
- In addition to traditionally considered baseline confounding, immortal time bias, time-dependent confounding, and selection bias are driving systematic errors in the case of ovarian cancer therapy, leading to over- and underestimation of the true treatment effect depending on the imperfect adjustment method.

What is the implication and what should change now?

- It is important to increase the knowledge about causal analytic frameworks that go beyond simple regression or propensity score analyses in clinical research, clinical guideline development and health technology assessment, to ultimately make sure patients receive treatments with causally substantiated benefits that outweigh the harms.

1. Introduction

Real-world evidence (RWE) can complement evidence from randomized controlled trials (RCTs) in order to assess comparative treatment effectiveness in routine practice under real-life conditions, where the artificial settings of trials can be avoided [1,2].

However, comparative effectiveness analysis of real-world data (RWD) poses methodological challenges [1,3−7]. Traditional statistical methods attempt to control for time-independent confounding by matching techniques, stratification, weighting, or multivariable-adjusted analyses incorporating baseline variables. For studies with the potential of time-dependent confounding, further causal inference approaches have been developed, applied, and discussed during the last decades [3−5,8,9]. These approaches involve three complementary conceptual components: causal diagrams, g-methods, and the target trial approach [3−5,10−13].

The target trial approach minimizes immortal time bias, which is a key concern for ‘ever vs. never’ treatment comparisons. It can be difficult to understand whether patients live longer because they receive a particular treatment or whether patients received that treatment because they lived longer [3−5,7,14]The target trial approach is a structural approach emulating an RCT by following its structure, defining a time zero representing the time of inclusion, randomization, and treatment allocation time. This structural approach attempts to avoid immortal time biases and is very useful for comparing multiple dynamic treatment strategies [5,15,16].

An example of a dynamic research question is how to optimize treatment management in women with ovarian cancer. While first-line therapy is well defined as the surgery followed by platinum-based chemotherapy [17], second-line chemotherapy (LOT2) in women with progressive ovarian cancer is less well-defined. It is not only debated whether or not LOT2 should be provided but also when would be the best time to provide LOT2. Potential starting points may be (in timely order), at the time of progression, defined by increasing biomarker (i.e., cancer-antigen 125 [CA-125]), when a computerized tomography (CT) scan shows tumor growth, when symptoms occur, or never. Besides the dynamic treatment component, the case of assessing when to start LOT2 in women with progressive ovarian cancer using observational data bears all the problems of RWD such as unmeasured, time-independent, and time-dependent confounding, immortal time bias, and selection bias. This case was therefore chosen to demonstrate potential biases when inferring causal treatment effects from RWD.

The aim of this study was to systematically assess and demonstrate the type and impact of potential biases that may occur when deriving causal conclusions from large real-world database analyses using different methodological approaches. As a case example, we used a retrospective observational dataset linking electronic health records, hospital data, and claims data from patients treated for ovarian cancer in practices throughout the United States.

2. Methods

To estimate the impact of potential biases when analyzing RWD, we created and followed a causal

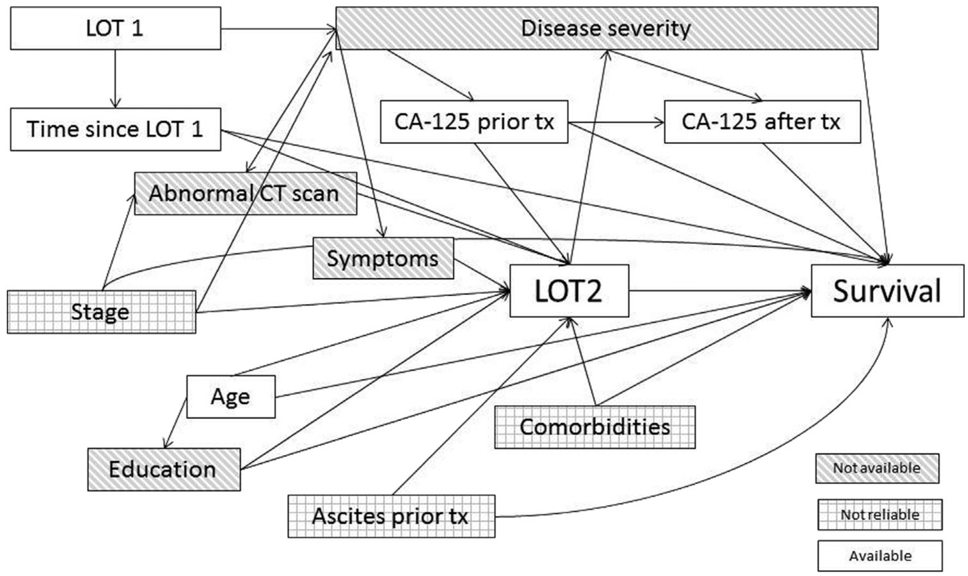

Fig. 1. Causal diagram including measured and unmeasured confounding. Ca-125: Cancer-Antigen 125; LOT1: first line chemotherapy; LOT2: second-line chemotherapy; CT: computed tomography; tx: treatment.

analytic framework prior to the data analysis. We 1) used the case of ovarian cancer, 2) identified potential biases using a causal graph (Fig. 1), 3) judged the direction of potential biases based on expert assumptions encoded in the causal graph following the techniques described by VanderWeele et al. [8] (Table 1), 4) selected a published RCT [18] as reference case (“gold standard”), 5) defined analytic approaches from crude statistical associations and traditional techniques adjusting for time-independent (baseline) confounding to more sophisticated causal inference methods adjusting for time-dependent confounding, and 6) emulated a target trial based on the study population of the reference case RCT to appropriately compare results from the observational data analysis to the RCT results. For details on steps 2 and 3 see eAppendix A.1.

The causal diagram is a simplified version of a directed acyclic graph (DAG) with time-varying variables. It shows the correlation of interest, being the effect of LOT2 on survival and variables that directly or indirectly correlate with both variables. White boxes indicate variables that are available in the dataset; variables indicated by checked boxes contain a substantial fraction of missing or not adequately measured variables; striped boxes indicate variables that are not present in the database.

2.1. Description of the case example and definition of the research question

To estimate the presence, direction, and magnitude of potential biases when analyzing RWD, we chose a dynamic treatment question: Does (LOT2) improve overall survival in patients with ovarian cancer who progressed after the initial successful surgery and first-line chemotherapy (LOT1). We expected to see time-independent, time-dependent confounding, selection bias, and immortal time bias. Furthermore, a published RCT

Table 1. Expert panel assessment of assumed bias direction

	Direction of bias (pro or contra LOT2) estimation of HR		
Bias	**HR - in favor of LOT2**	**HR ± either**	**HR + against LOT2**
Confounding			
Unmeasured(disease severity, CT scan, symptoms)			X
(Education)	X		
Time-independent(ascites, stage)			X
(Age, comorbidities, time since LOT1)		X	
Time-dependent(CA-125)			X
Immortal-time Bias	X		
Selection Bias/Confounding by indication			X

Abbreviations: HR, Hazard Ratio; HR -, underestimation of HR; HR ±, either under- or overestimation of HR; HR +, overestimation of HR; CT, computer tomography; LOT1, first line treatment.

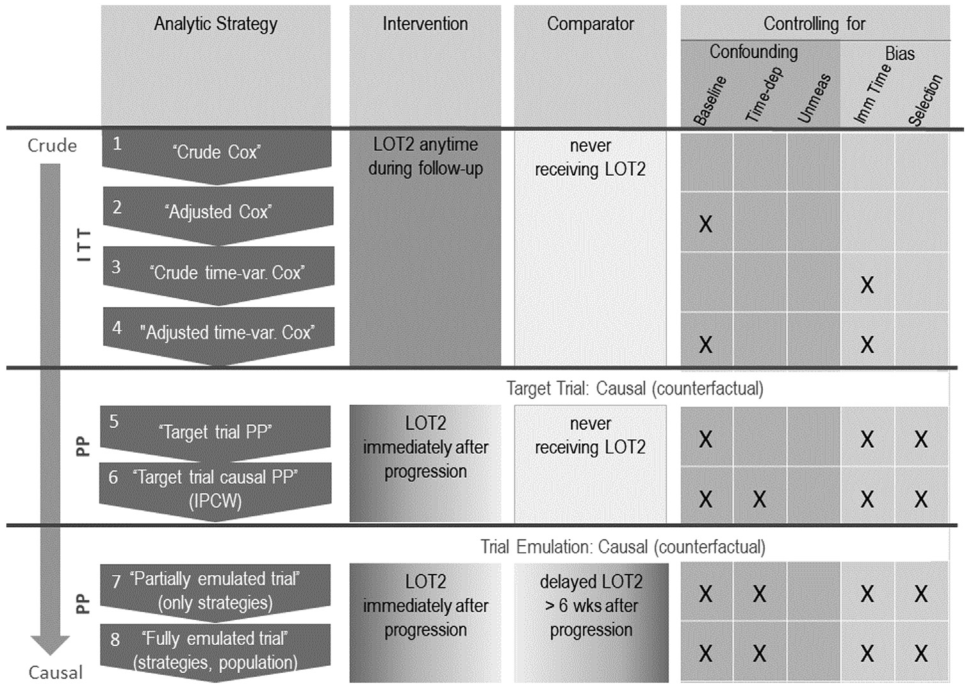

Fig. 2. Analytic strategies. Time-dep: time-dependent; Unmeas: unmeasured; Imm Time: immortal-time; LOT2: second-line therapy; Wks: weeks; time-var.: time-varying; IPCW: inverse probability of censoring weighting; ITT: intention to treat; PP: per protocol.
Strategies:
1. "Crude Cox": Univariable Cox regression without adjustment for covariates, comparing the overall survival of patients receiving LOT2 at any time after progression to the overall survival of those never receiving LOT2.
2. "Adjusted Cox": Cox regression with adjustment for baseline confounding covariates, comparing the overall survival of patients receiving LOT2 at any time after progression to the overall survival of those never receiving LOT2.
3. "Crude time-var. Cox": Cox regression including treatment as time-varying covariate to compare (treated vs. nontreated) person time rather than (ever treated vs. never treated) individuals.
4. "Adjusted time-var. Cox": Cox regression including treatment as time-varying covariate and additionally adjusted for baseline confounding to compare (treated vs. nontreated) person time rather than (ever treated vs. never treated) individuals.
5. "Target trial PP": Replication of all patients to mimic a "counterfactual" clinical trial assigning each patient to each treatment arm and performing a per protocol analysis where individuals are being censored at the time of treatment violation.
6. "Target trial causal PP" (inverse probability of censoring weighting [IPCW]): Performing a target trial as described but accounting for informative censoring by applying the IPCW.
7. "Partially emulated trial" (only strategies): Applying the target trial approach and adapting the protocol regarding treatment strategies only to the one of the gold-standard RCT described by Rustin et al., that is, comparing "immediate treatment" to "delayed treatment".
8. "Fully emulated trial" (strategies, population): Emulating the gold-standard RCT by using the same treatment strategies as in the gold-standard RCT and additionally standardizing the study population to the study population of the gold-standard RCT.

investigated the same research question and served as reference case.

For this analysis, observational study cases were selected from a cohort of more than 12,000 patients with ovarian cancer with information collected from electronic medical record of primarily medium and large community-based oncology practices in the United States from January 2000 to June 2014 (see eAppendix A.1.).

We included female patients aged 18 years or older with ovarian, fallopian tube, or primary peritoneal cancer. Eligible patients must have disease progression after standard LOT1 treatement. Progression was defined as the doubling value of CA-125 (details see eAppendix A.1.) [19].

Some variables such as the biomarker CA-125 are just sporadically measured. For example, the biomarker CA-125 is not routinely measured at each clinical visit. Hence, it is not possible to determine whether the biomarker is missing or not measured. We assumed parameters were measured as indicated by the data. The last measurement therefore reflecting the knowledge of the physician.

2.2. Selection of reference case/gold standard

We selected a reference (gold standard) study by Rustin et al [18] to compare our effect estimates. The RCT estimated the benefit of early LOT2 in women with ovarian cancer and included women with ovarian cancer who had undergone surgery and LOT1. Women were randomized to early treatment (LOT2 within 28 days after progression that was purely based on increased CA-125 concentrations, that is, twice the upper limit of normal) or delayed treatment (delaying treatment and only commencing treatment at clinical or symptomatic relapse). Survival was compared between arms. They could not find evidence for a difference in overall survival between early and delayed treatment adjusted for stratification and prognostic factors (HR 1.01, 95% CI: 0.82−1.25) [18].

2.3. Definition of analytic approaches from crude to causal

To identify the impact of different biases that may occur when estimating causal effects of LOT2 on overall survival using RWD, we followed a stepwise analytic approach (analyses 1−6), which is described in the following paragraphs. The approach ranges from crude analysis and traditional multivariate adjustments up to a full causal analysis. A crude (i.e., purely statistical association) vs. causal analysis is not only depicted by the statistical method but also by the precision of the research question and treatment allocation. The analytic strategies and the corresponding treatment allocations are illustrated in Figure 2 and described below. All statistical analyses were performed with SAS software version 9.4 (SAS Institute Inc).

Figure 2 shows the analytic strategies and their corresponding intervention and comparator to assess the type and impact of potential biases. The strategies built upon each other and increase in complexity when going from crude to a full causal analysis. The target trial follows a counterfactual approach asking for specific definition of treatment and a per protocol analysis. To allow for a comparison to our reference case, we adapted the full causal approach (analysis 6 "target trial") to emulate the reference trial by adapting the protocol as well as the trial cohort.

Typical biases occurring when analyzing real-world data are listed, and the "X" indicates whether the given strategy is controlling for that bias.

We started with a simple research question and simple treatment group allocation by comparing the crude (i.e., unadjusted) survival of progressed patients who had received LOT2 anytime during follow-up with the survival of those with progressive disease who had not received LOT2 at any time during follow-up, from here on called, "ever vs. never" comparison. In analysis 1, we applied a simple univariable Cox regression for overall survival without adjustment for covariates ("Crude Cox"). In analysis 2 ("adjusted Cox"), we controlled for baseline confounders (i.e., age, nadir, CA-125 at the time of progression, and time since first-line treatment) by including them as covariates into the Cox model. If the assumption of proportional hazards was violated, an interaction between treatment and time was included to model a time-dependent treatment effect.

To account for immortal time bias, which may occur in the "ever vs. never" treatment comparison, we compared (treated vs. non-treated) person time rather than (ever treated vs. never treated) individuals. Each patient contributed his/her person time to the treatment he/she received to the corresponding time point, from here on called, "treated vs. untreated person time" comparison [20−22]. In analysis 3, ("Crude time-var. Cox"), we included treatment as time-varying covariate in the crude Cox model in order to eliminate the immortal-time bias. In analysis 4, ("Adjusted time-var. Cox"), we additionally adjusted for baseline confounding using the same covariates as in analysis 2. Additionally, to treatment, CA-125 value was included as time-varying covariate as it changed over time.

For the more complex causal analyses, we followed the target trial approach, structuring any data analysis as if one would design a RCT as described by Hernan, Robins, Cain, and others [3,15,23−26]. We started with a well-defined research question assessing the causal effect of LOT2 on survival when provided to women with ovarian cancer immediately after progression vs. never LOT2. To account for natural time variation within RWD, we allowed for a lag time of 6 weeks ("grace period") after the diagnosis of progression. We refer to these adapted strategies as "immediate vs. never" treatment. In analysis 5, ("target trial: PP"), we followed the target trial approach [3,5,16,27,28], which estimated the per protocol effect. We replicated all patients in order to mimic a "counterfactual" clinical trial, assigning each patient to each treatment arm and censored them at the time of treatment violation. In analysis 6, ("target trial: causal PP"), we considered the fact that artificial censoring is potentially informative. Hence, we applied a marginal structural Cox model adjusting for informative censoring by IPCW [4,11,28−31]. IPCW aims at correcting for informative censoring by applying a two-step approach. First, a weight model estimates the probability of not being censored. Second, the inverse of the estimated probability is used as weight in the outcome model. This weighting procedure creates an unconfounded "pseudo-population" [32]. In sensitivity analyses, we assessed the robustness of results of the outcome model using different weight models [33−35] (Table 3).

2.4. Trial emulation using the reference case as gold standard

To be able to compare the estimated effect measures of the observational data to the gold standard, we followed the recommendations of Lodi et al. to harmonize the study protocols and study population [36,37] (analyses 7−8). In analysis 7, ("partially emulated trial" (only strategies)), we adapted the target trial protocol (only) to the treatment strategies of the protocol of the gold-standard RCT described by Rustin et al. We introduced a new strategy labeled "delayed treatment", as used in the Rustin et al. trial, and compared this strategy to "immediate treatment" [18]. The RCT protocol for the "delayed treatment" arm dictated the start of LOT2 purely based on abnormalities on the CT scan or symptoms and not on progression based on CA-125 increase. In the absence of information on CT scans or symptoms in the observational data and any initiation of LOT2, all treatment not based on biomarker increase (i.e., 6 weeks after progression defined by biomarker increase) were considered delayed treatment. Hence, patients in the "delayed treatment" arm were

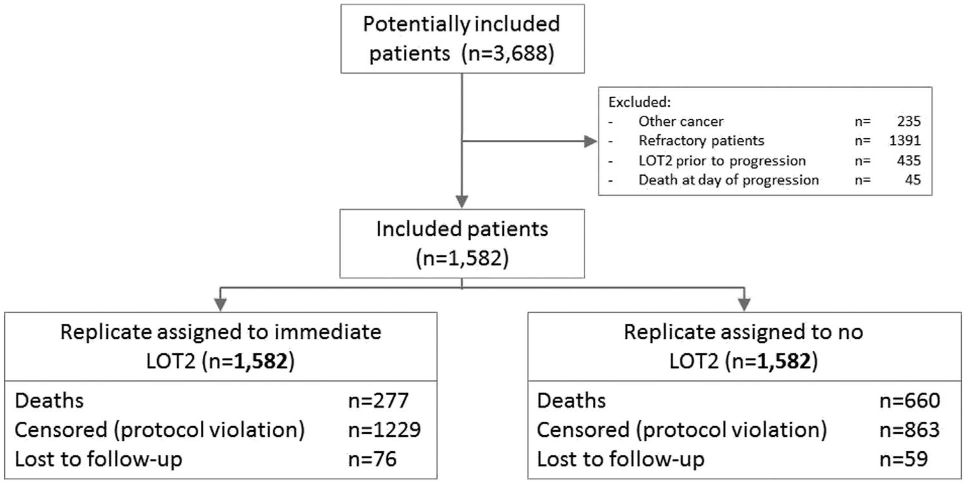

Fig. 3. Flowchart of the included cohort. Time-var.: time-varying; IPCW: inverse probability of censoring weighting; ITT: intention to treat; PP: per protocol; HR: hazard ratio; 95% CI: 95% confidence interval.
LOT2: second-line chemotherapy.
This flowchart shows the included patients. 3,688 of patients in the database showed a diagnosis of peritoneal cancer. Those not meeting the full inclusion criteria were excluded. Most excluded patients were not successfully receiving LOT1. All included patient data were replicated and allocated to each treatment arm. The chart shows the number of deaths, censoring due to protocol violation, and those who lost to follow-up.

artificially censored only during the first 6 weeks after progression if they started treatment.

In analysis 8, (“fully emulated trial” [strategies, population]), we emulated the gold-standard RCT by not only using the treatment strategies as defined in the Rustin et al. trial, but also by standardizing our study population to the study population of the RCT. In other words, we used proportional weights in our analysis to create a similar baseline cohort as the cohort of the gold-standard RCT with regard to the baseline distributions of age distribution, first-line treatment, and progression-free survival.

In sensitivity analyses, we tested the robustness of the treatment effect estimate by changing assumptions around time functions, duration of follow-up, population age, grace period, definition of delayed treatment, and weight models.

2.5. Bias estimation

We estimated the size of the bias in each analytic strategy by comparing the estimated HR to the HR from the reference case. We did that visually and calculated the proportional difference of the treatment effect. The effect of potential unmeasured confounding bias was assessed using the techniques described by VanderWeele et al. [8] (see eAppendix A.2.2.).

3. Results

3.1. Descriptive results of the ovarian cancer data

Out of a total of 3,688 patients meeting the inclusion criteria, 1,582 remained in our observational cohort study after applying the exclusion criteria (Figure 3). The mean age was 67 years with a standard deviation of 11 years.

3.2. Effect estimates of analytic approaches from simple to complex and comparison to reference case

Comparing the groups of women who never received LOT2 to women who did receive LOT2 at any time in the database (analysis 1) provided us with an estimated crude HR of 0.56 (95% CI: 0.49–0.64) assuming a constant HR. When adjusting for baseline confounding (analysis 2), the HR was 0.51 (95% CI: 0.44–0.59). Adjusting for immortal time bias by including time-varying covariates into the Cox model yielded estimated crude (analysis 3) and adjusted (analysis 4) HRs of 1.41 (95% CI: 1.22–1.64) and 1.37 (95% CI: 1.18–1.59), respectively. Applying the target trial concept and adapting the compared strategies, provided an HR of 1.35 (95% CI: 1.17–1.55), when not accounting for informative censoring (analysis 5) and of 1.38 (95% CI: 1.22–1.63) when accounting for informative censoring by applying IPCW (analysis 6). Results for all these analyses visualizing the directions and magnitudes of different biases are shown in Figure 4 and in Table 2.

Figure 4 shows the HR and its 95% confidence interval for each analytic strategy and puts it into comparison to the HR of the reference case indicated by the dotted line and the gray area indicating the 95% CI. A HR of 1 suggests no treatment effect; a HR below 1 suggests a beneficial treatment effect, while a hazard ratio above 1 indicates a harmful treatment effect.

3.3. Trial emulation using the reference case as gold standard

Partially emulating the trial by adapting the compared strategies to the reference case and comparing immediate

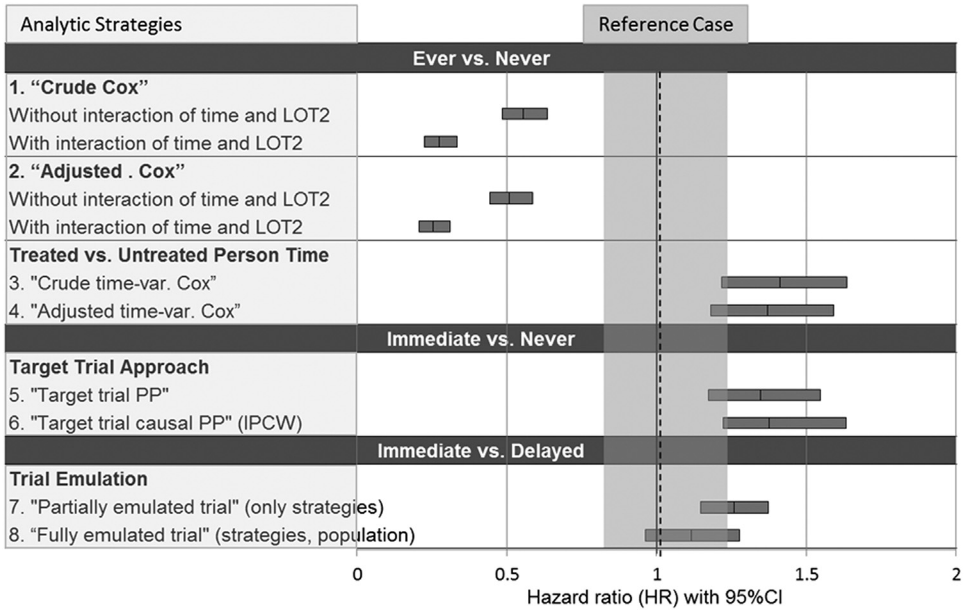

Fig. 4. Base-case results.
Treatment allocation:
1. Ever vs. never: Comparing the survival of progressed patients who had received LOT2 anytime during follow-up to the survival of those who had not received LOT2 at any time during follow-up.
2. Immediate vs. never: Comparing the survival of progressed patients who had received LOT2 within 6 weeks after progression to the survival of those who had not received LOT2 at any time during follow-up.
3. Immediate vs. delayed: Comparing the survival of progressed patients who had received LOT2 within 6 weeks after progression to the survival of those who had received LOT2 later than 6 weeks after progression or never.
Analytic strategies:
4. ''Crude Cox'': Univariable Cox regression without adjustment for covariates.*
5. ''Adjusted Cox'': Cox regression with adjustment for baseline confounding covariates.*
6. ''Crude time-var. Cox'': Univariable Cox regression including treatment as time-varying covariate to compare (treated vs. non-treated) person time rather than (ever treated vs. never treated) individuals.
7. ''Adjusted time-var. Cox'': Cox regression including treatment as time-varying covariate and additionally adjusted for baseline confounding to compare (treated vs. nontreated) person time rather than (ever treated vs. never treated) individuals.
8. ''Target trial PP'': Replication of all patients to mimic a ''counterfactual'' clinical trial assigning each patient to each treatment arm and performing a per protocol analysis where individuals are being censored at the time of treatment violation.
9. ''Target trial causal PP'' (IPCW): Performing a target trial as described but accounting for informative censoring by applying the IPCW.
10. ''Partially emulated trial'' (only strategies): Applying the target trial approach and adapting the protocol regarding treatment strategies only to the one of the gold-standard RCT described by Rustin et al., that is, comparing ''immediate treatment'' to ''delayed treatment''.
11. ''Fully emulated trial'' (strategies, population): Emulating the gold-standard RCT by using the same treatment strategies as in the gold-standard RCT and additionally standardizing the study population to the study population of the gold-standard RCT.
*In analytic strategies 1 and 2, the proportional hazards assumption was violated. In these cases, in addition to the ''average'' HR, the initial HR of a model with interaction between linear time and treatment is reported.

LOT2 to delayed LOT2 (analysis 7), the estimated HR was 1.26 (95% CI: 1.15−1.37). The HR was 1.12 (95% CI: 0.96−1.28) when fully emulating the trial by adjusting the trial cohort from the observational study to trial cohort of the RCT as described by Rustin et al. (analysis 8).

3.4. Sensitivity analyses

To test the robustness of our results, we conducted several sensitivity analyses (see Table 3). We changed the time horizon from the base case (tailored) to 5 years, and 7 years, used different assumptions when modeling time (base case as spline to linear time), looked at patients older than 65 years, changed the grace period from 6 weeks to 4 weeks, defined delayed treatment not only by a treatment start later than 6 weeks after progression but also by a minimum biomarker of 3 times the nadir, and applied a weight function modeling time as linear function. All those changes changed the point estimate by less than 5%.

4. Discussion

We used the case of LOT2 in women with ovarian cancer to investigate the potential biases that may occur when using observational RWD for comparative effectiveness

Table 2. Base case results with bias estimation

Estimation method	HR	95% Conf. Int.	Bias
Ever vs. Never			
1. ‘‘Crude Cox’’			
Without interaction of time and LOT2	0.56	0.49-0.64	45%
With interaction of time and LOT2[a]	0.27	0.22-0.34	73%
2. ‘‘Adjusted Cox’’			
Without interaction of time and LOT2	0.51	0.44-0.59	50%
With interaction of time and LOT2[a]	0.25	0.21-0.31	75%
‘‘Treated vs. Untreated Person Time’’			
3. ‘‘Crude time-var. Cox’’	1.41	1.22-1.64	−40%
4. ‘‘Adjusted time-var. Cox’’	1.37	1.18-1.59	−36%
Immediate vs. Never			
Target trial approach			
5. ‘‘Target trial PP’’	1.35	1.17-1.55	−33%
6. ‘‘Target trial causal PP’’ (IPCW)	1.38	1.22-1.63	−36%
Immediate vs. Delayed			
Trial emulation			
7. ‘‘Partially emulated trial’’ (IPCW)	1.26	1.15-1.37	−25%
8. ‘‘Fully emulated trial’’ (IPCW)	1.12	0.96-1.28	−10%

Abbreviations: HR, hazard ratio; 95% Conf. Int., 95% confidence interval; LOT2, second-line therapy; Time-var., time-varying; vs., versus; PP, per protocol; IPCW, inverse probability of censoring weighting; Partially Emulated, partially emulating the Rustin trial by emulating the treatment strategies as described by Rustin et al.; Fully Emulated, fully emulating the Rustin trial by emulating the trial cohort described by Rustin et al. in addition to emulating the treatment strategies.

Bias is estimated as proportional difference to the reference case point estimate [18], where a positive number indicates bias in favor of the treatment and a negative number indicates bias against the treatment.

[a] In analytic strategies 1 and 2, the proportional hazards assumption was violated. In these cases, in addition to the ‘‘average’’ HR, the initial HR of a model with interaction between linear time and treatment is reported.

research. At times when RWD are widely available, it is extensively debated how such data can be used for assessing comparative effectiveness outside of the artificial setting of RCTs [38–41]. To assess potential biases that may occur in RWD analysis, we conducted several analyses assessing the effect of LOT2 on survival. We started with crude, purely associative analyses and added more and more complexity to result in a full causal assessment. We learned that RWD have potential for several biases that may go in different directions. In the presented case-example, immortal time bias plays a major role, typically biasing results in favor of treatment. However, time-independent, time-dependent, and unmeasured confounding may bias the results in different directions (see eAppendix C). We can confirm that the estimated treatment effect most closely matched the RCT treatment effect when applying all causal features and emulating the trial by matching the trial design as well as the trial study population.

We started with a crude Cox regression model comparing treated patients with those that never received second-line therapy and found that women receiving LOT2 after progression had a longer life expectancy than those who did not receive LOT2. These results are purely associative. Causal interpretations as well as transferring the results to other situations and populations need to be handled with caution. During our analyses, we changed the compared strategies to contrast the simple approaches including ill-defined (but still frequently used) comparisons with the causal target trial approach and the trial emulation reflecting the increasing complexity of the analyses (details see eAppendix D).

The potential for biases such as immortal time bias in observational data is known and several studies exist that provide insight in techniques to correct for them [3,5,7,16,24,25,29]. Those techniques include visual, structural, and statistical approaches, which are validated in several study designs and therapeutic areas [3,5,8,12,15,23–26,29–31,34,35,42–58]. Also, studies exist applying and comparing several analytic strategies to observational data to assess potential biases [59–62]. One study compared results for patients eligible for a trial to those not eligible for that trial [63]. In our study, we emulated a trial with IPCW and compared it to results of other analytical methods. We compared analytic strategies with increasing complexity, applying visual, structural, and analytical causal methods, and comparing it to the results of an RCT by emulating that trial. By the estimation of bias direction, combination of methods, and the increasing complexity, we offer a novel approach for understanding each type of bias and each methodological approach. Being able to closely reproduce the findings of

Table 3. Sensitivity analyses

Sensitivity analyses	% Change in effect estimate (HR)
Study time horizon	
5 years	4.9%
3 years	1.8%
Modeling time as linear covariate	1.3%
Study population only >65 years	1.4%
Grace period modeled as 4 weeks	2.1%
Delayed tx defined as minimum 3 times nadir and >6 weeks after progression	0.3%
Weight function with linear time	0.1%

Abbreviations: HR, hazard ratio; tx, treatment.

our reference RCT when thoroughly justifying and applying causal methods provided us with trust in such methods.

Our study has several limitations. First, our data have limitations typical for RWD. Some variables necessary for an unbiased causal analysis according to our DAG were not available (e.g., imaging, symptoms). Our assessment of the direction of bias due to unmeasured confounders based on our DAG indicated a bias overestimating the HR. This is confirmed by the comparison of our causal analysis results to the findings of the Rustin trial, which reported a slightly lower HR. Some other variables available in our dataset are just sporadically measured; for example, the biomarker CA-125 is not routinely measured at each clinical visit. In this case, we assumed the last measurement available reflects the knowledge of the physician.

Second, we used progression as indicated by the marker CA-125 as the decision criterion. However, the time between progression as defined by the biomarker and clinical onset may vary widely [64]. In our dataset, we did not have any information on progression indicated by CT scans or symptoms. Hence, not receiving any LOT2 in our study may reflect either no treatment despite progression or no treatment because of absence of clinical symptoms. Clinically, the comparative effect estimates of analyses 1−6 should therefore be interpreted with caution but likely this issue does not affect the overall qualitative picture of bias assessment.

Third, we did not consider any genetic proxies such as family history as potential confounders. Such prognostic factors may introduce potential confounding, for example, because they may influence either physicians' prescription or patient awareness and preference for starting LOT2.

Fourth, we call analysis 6 a causal per protocol analysis despite residual unmeasured confounding. Using the DAG, showed that all residual confounding is likely to overestimate the estimated HR comparing treated women to not treated women.

Fifth, our study population reflects patients in medium/large oncological practices, and therefore, may not be generalizable to all patients.

Sixth, the delayed treatment strategy is likely a more relevant comparative strategy than the never treatment strategy. However, it is not fully compliant with a well-defined target trial approach as it does not define the treatment strategy explicitly. We would have liked to include concrete strategies of starting LOT2 based on clinical onset of progression. However, Rustin et al. show that even an RCT may not define a treatment strategy explicitly. He defined the delayed LOT2 strategy more broadly which is matched by our approach more closely [18].

Seventh, it must be noted that comparing conditional with marginal HRs is comparing apples with oranges, as HRs are not collapsible [65,66]. We therefore used the conditional results of the Rustin trial as a reference to be compared with the results of our conditional analyses.

Eighth, we did not apply alternative g-methods, such as the parametric g-formula [29,67] or g-estimation, with structural nested models [10,29,68,69]. However, the g-formula fits best if there are natural intervals (e.g., visits) [67,70]. For example, the first application of the parametric g-formula was performed in 2002 in the Framingham Offspring Study with scheduled 4- and 8-year intervals [71] which is not the case in our study. Another causal inference approach, g-estimation using structural nested models, relies on the assumption of a common treatment effect across all patients, which is unlikely to be true in second-line ovarian cancer chemotherapy, where some women may benefit and others may not.

Lastly, we used the Rustin trial as the reference case and emulated the trial by mapping the structure and study population (e.g., inclusion criteria) of the Rustin trial. However, some differences to the Rustin trial persist. Patients in the Rustin trial were closely monitored after the LOT1 (every 3 months), which may have led to an earlier detection of disease progression than in the cohort of our analysis. Also, the allowed time to start therapy after detection of progression was shorter in the Rustin trial (28 days) than in our study (42 days). However, we felt that our assumption was reasonable for an observational study as the physician did not have the information of the grace period prior to their treatment decision. Additionally, our clinical experts supported the application of a 42-day period as it is considered realistic in the real-world setting. A sensitivity analysis changing the grace period to 28 days showed robust results. For more details on differences, see eAppendix B3. We were able to identify and quantify several biases that may occur when analyzing observational data using an RCT as the comparative gold standard. Further, we assessed the comparative effectiveness of LOT2 in women with progressive ovarian cancer when applying complex causal methods combining visual, structural, and statistical approaches. However, a comprehensive assessment of any treatment should explicitly consider the real-world setting and patient values. This means that the final results should represent the real-world population rather than the artificial trial population. In addition, any patient-shared decision making on whether or not LOT2 should be provided must involve the entire spectrum of benefits and harms related to chemotherapy and cancer, such as anxiety, side effects, symptoms, effectiveness, comorbidities, time on treatment, time of treatment, etc. Also, the personal and economic value of all those components needs to be considered when deciding on the provision of chemotherapy. An appropriate method for the synthesis of such evidence is decision-analytic modeling, which requires causal input parameters and follows a counterfactual approach predicting and synthesizing the outcomes in a world with and without the intervention [72].

In the time of digitalization of health care data and "big" RWD, further educational efforts on structural and statistical methods aiming for causal inference from RWD to inform health care decision-making should be expanded to a broader audience, including those who plan the data collection. Current frameworks and recommendations on planning, conducting, reporting, and assessing observational studies [1,73–75] should add additional emphasis on the risk of typical biases, such as immortal time bias and time-dependent confounding and their adjustment methods. An increased knowledge on potentials and limits of RWE can serve as basis for evidence synthesis and decision analysis in medicine and public health.

5. Conclusion

We used the case example of LOT2 in women with progressive ovarian cancer to identify potential biases that may occur when applying different noncausal and causal analytic approaches to real-word data. We identified several biases resulting in considerable variation of the effect measure in different directions, with immortal time bias leading to larger biases than confounding. When emulating the reference randomized target trial, we were able to replicate the effect estimates of the RCT very well. Studies such as ours are important to demonstrate the need for causal analyses, to increase the trust and confidence in RWE, and to help in collecting appropriate data and selecting appropriate analysis methods. Although RWE should not substitute well-conducted clinical trials due to the substantial potential for bias in RWE, we do believe that RWE based on appropriate methods is a valuable addition to clinical trials.

Acknowledgments

We thank Zbigniew Kadziola for his continuing advice regarding statistical methods and SAS coding including the coding for the bootstrap analysis.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.10.005.

References [76–80] were cited in the Supplementary Appendix.

References

[1] Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goettsch W, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Pharmacoepidemiol Drug Saf 2017;26(9):1033–9. https://doi.org/10.1002/pds.4297.

[2] Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106(1):1–9. https://doi.org/10.1007/s00392-016-1025-6.

[3] Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016;183:758–64. https://doi.org/10.1093/aje/kwv254.

[4] Hernan MA, Robins JM. Per-protocol analyses of pragmatic trials. N Engl J Med 2017;377:1391–8. https://doi.org/10.1056/NEJMsm1605385.

[5] Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016;79:70–5. https://doi.org/10.1016/j.jclinepi.2016.04.014.

[6] Avoiding potential biases in real world data analysis by emulating a clinical trial. Berlin, Germany: BEMC (Berlin Epidemiological Methods Colloquium); 2020.

[7] Maringe C, Benitez Majano S, Exarchakou A, Smith M, Rachet B, Belot A, et al. Reflections on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. Int J Epidemiol 2020;49:1719–29. https://doi.org/10.1093/ije/dyaa057.

[8] VanderWeele TJ, Hernan MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias. Epidemiology 2008;19(5):720–8. https://doi.org/10.1097/EDE.0b013e3181810e29.

[9] Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013;177:292–8. https://doi.org/10.1093/aje/kws412.

[10] Robins JM. Marginal structural models versus structural nested models as tools for causal inference. Statistical models in epidemiology, the environment, and clinical trials. New York, NY: Springer; 2000:95–133.

[11] Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000;56:779–88.

[12] Pearl J. Causality: models, reasoning, and inference. Cambridge, UK: Cambridge University Press; 2009.

[13] Vanderweele TJ, Tan Z. Directed acyclic graphs with edge-specific bounds. Biometrika 2012;99(1):115–26. https://doi.org/10.1093/biomet/asr059.

[14] Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167:492–9. https://doi.org/10.1093/aje/kwm324.

[15] Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernan MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. Int J Biostat 2010;6(2). Article 18.

[16] Kuehne F, Jahn B, Conrads-Frank A, Bundo M, Arvandi M, Endel F, et al. Guidance for a causal comparative effectiveness analysis emulating a target trial based on big real world evidence: when to start statin treatment. J Comp Eff Res 2019;8(12):1013–25. https://doi.org/10.2217/cer-2018-0103.

[17] NCCN. In: Clinical practice guideline in oncology (NCCN Guidelines): Ovarian Cancer/Fallopian Tube Cancer/Primary Peritoneal

Cancer. Version: 5.2022. https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1453 Accessed November 30, 2022.

[18] Rustin GJ, van der Burg ME, Griffin CL, Guthrie D, Lamont A, Jayson GC, et al. Early versus delayed treatment of relapsed ovarian cancer (MRC OV05/EORTC 55955): a randomised trial. L ancet 2010;376(9747):1155−63. https://doi.org/10.1016/s0140-6736(10)61268-8.

[19] Rustin GJ, Vergote I, Eisenhauer E, Pujade-Lauraine E, Quinn M, Thigpen T, et al. Definitions for response and progression in ovarian cancer clinical trials incorporating RECIST 1.1 and CA 125 agreed by the Gynecological Cancer Intergroup (GCIG). Int J Gynecol Cancer 2011;21(2):419−23. https://doi.org/10.1097/IGC.0b013e3182070f17.

[20] Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health 1999;20:145−57. https://doi.org/10.1146/annurev.publhealth.20.1.145.

[21] Kleinbaum DG, Klein M. Survival analysis: a self-learning text. In: Statistics for biology and health. 3th ed. New York, NY: Springer-Verlag; 2012.

[22] Powell TM, Bagnell ME. Your ''survival'' guide to using time-dependent covariates. In: Providers PaHC. Orlando, FL: SAS Global Forum 2012; 2012:168−77.

[23] Cain LE, Caniglia EC, Phillips A, Olson A, Muga R, Perez-Hoyos S, et al. Efavirenz versus boosted atazanavir-containing regimens and immunologic, virologic, and clinical outcomes: a prospective study of HIV-positive individuals. Medicine 2016;95(41):e5133. https://doi.org/10.1097/md.0000000000005133.

[24] Cain LE, Saag MS, Petersen M, May MT, Ingle SM, Logan R, et al. Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. Int J Epidemiol 2016;45:2038−49. https://doi.org/10.1093/ije/dyv295.

[25] Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006;60:578−86. https://doi.org/10.1136/jech.2004.029496.

[26] Lodi S, Sharma S, Lundgren JD, Phillips AN, Cole SR, Logan R, et al. The per-protocol effect of immediate versus deferred antiretroviral therapy initiation. AIDS 2016;30(17):2659−63. https://doi.org/10.1097/qad.0000000000001243.

[27] Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15(5):615−25.

[28] Kuehne F, Siebert U, Faries DE. A target trial approach with dynamic treatment regimes and replicates analyses. In: Faries D, Zhang Z, Kadziola ZA, Siebert U, Kuehne F, Obenchain RL, et al, editors. Real World Health Care Data Analysis: Causal Methods and Implementation Using SAS. Cary, NC: SAS Institute; 2020:321−52.

[29] Hernan M, Robins J. Causal inference: what if. London, UK: Chapman & Hall/CRC; 2020.

[30] Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11:561−70.

[31] Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials–an economic evaluation context: methods, limitations, and recommendations. Med Decis Making 2014;34(3):387−402. https://doi.org/10.1177/0272989x13520192.

[32] Robins JM. Marginal structural models. In: Proceedings of the section on bayesian statistical science. Alexandria, VA: American Statistical Association, 1998; 1997:1−10.

[33] Latimer N, White I, Tilling K, Siebert U. Improved two-stage estimation to adjust for treatment switching in randomised trials: g-estimation to address time-dependent confounding. Stat Methods Med Res 2020;29(10):2900−18. https://doi.org/10.1177/0962280220912524.

[34] Latimer NR, Henshall C, Siebert U, Bell H. Treatment switching: statistical and decision-making challenges and approaches. Int J Technol Assess Health Care 2016;32(3):160−6. https://doi.org/10.1017/S026646231600026X.

[35] Robins JM. Correction for non-compliance in equivalence trials. Stat Med 1998;17:269−302. discussion 387-9.

[36] Lodi S, Phillips A, Lundgren J, Logan R, Sharma S, Cole SR, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. Am J Epidemiol 2;188:1569−77. https://doi.org/10.1093/aje/kwz100.

[37] Schomaker M, Kühne F, Siebert U. RE: ''effect estimates in randomized trials and observational studies: comparing apples with apples''. Am J Epidemiol 2020;189:77−8. https://doi.org/10.1093/aje/kwz194.

[38] Crown WH, Bierer BE. Real-world evidence: understanding sources of variability through empirical analysis. Value Health 2021;24(1):116−7.

[39] Thompson D. Replication of randomized, controlled trials using real-world data: what could go wrong? Value Health 2021;24:112−5.

[40] FDA U. Framework for FDA's real-world evidence program. Silver Spring, MD: US Department of Health and Human Services Food and Drug Administration; 2018.

[41] Sheffield KM, Dreyer NA, Murray JF, Faries DE, Klopchin MN. Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions. J Comp Eff Res 2020; 9(15):1043−50. https://doi.org/10.2217/cer-2020-0161.

[42] Bell H, Latimer N, Amonkar M, Swann S. Adjusting for treatment switching in rcts - identifying, analysing and justifying appropriate methods: a case study in metastatic melanoma. Value Health 2015; 18(7):A338. https://doi.org/10.1016/j.jval.2015.09.128.

[43] Latimer NR, Abrams KR, Amonkar MM, Stapelkamp C, Swann RS. Adjusting for the confounding effects of treatment switching-the BREAK-3 trial: dabrafenib versus dacarbazine. Oncologist 2015; 20(7):798−805. https://doi.org/10.1634/theoncologist.2014-0429.

[44] Latimer NR, Abrams KR. NICE DSU Technical Support Document 16: Adjusting Survival Time Estimates in the Presence of Treatment Switching. London, UK: National Institute for Health and Care Excellence (NICE); 2014.

[45] Henshall C, Latimer NR, Sansom L, Ward RL. Treatment switching in cancer trials: issues and proposals. Int J Technol Assess Health Care 2016;32(3):167−74. https://doi.org/10.1017/s026646231600009x.

[46] Robins JM, Hernán MA, Siebert U. Effects of multiple interventions. Semantic Scholar 2004;1:2191−230.

[47] Robins JM, Hernan MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, editors. Longitudinal data analysis. London, UK: Chapman & Hall/CRC; 2009:553−99. chap 23. Handbooks of Modern Statistical Methods.

[48] Murray EJ, Hernan MA. Adherence adjustment in the Coronary Drug Project: a call for better per-protocol effect estimates in randomized trials. Clin Trials 2016;13(4):372−8. https://doi.org/10.1177/1740774516634335.

[49] Jain P, Danaei G, Robins JM, Manson JE, Hernan MA. Smoking cessation and long-term weight gain in the Framingham Heart Study: an application of the parametric g-formula for a continuous outcome. Eur J Epidemiol 2016;31:1223−9. https://doi.org/10.1007/s10654-016-0200-4.

[50] Hernan MA. A good deal of humility: cochran on observational studies. Observational Stud 2015;2015(7):194−5.

[51] Caniglia EC, Sabin C, Robins JM, Logan R, Cain LE, Abgrall S, et al. When to monitor CD4 cell count and HIV RNA to reduce mortality and AIDS-defining illness in virologically suppressed HIV-positive persons on antiretroviral therapy in high-income countries: a prospective observational study. J Acquir Immune Defic Syndr 2016;72(2):214−21. https://doi.org/10.1097/qai.0000000000000956.

[52] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11(5):550−60.

[53] Griffiths C, Hay N, Sutcliffe F, Stevens A. NICE guidance on pazopanib for first-line treatment of advanced renal-cell carcinoma. Lancet Oncol 2011;12(3):221−2.
[54] Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. Stat Med 2013;32:1584−618. https://doi.org/10.1002/sim.5686.
[55] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10(1):37−48.
[56] Greenland S, Brumback B. An overview of relations among causal modelling methods. Int J Epidemiol 2002;31:1030−7.
[57] Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. Clin Trials 2008;5(1):5−13. https://doi.org/10.1177/1740774507087703.
[58] Hernan MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. Basic Clin Pharmacol Toxicol 2006;98(3):237−42. https://doi.org/10.1111/j.1742-7843.2006.pto_329.x.
[59] Kaiser P, Arnold AM, Benkeser D, Zeki Al Hazzouri A, Hirsch CH, Psaty BM, et al. Comparing methods to address bias in observational data: statin use and cardiovascular events in a US cohort. Int J Epidemiol 2018;47:246−54. https://doi.org/10.1093/ije/dyx179.
[60] Parast L, Griffin BA. Quantifying the bias due to observed individual confounders in causal treatment effect estimates. Stat Med 2020;39:2447−76. https://doi.org/10.1002/sim.8549.
[61] García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol 2017;32(6):495−500. https://doi.org/10.1007/s10654-017-0287-2.
[62] Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining bias in studies of statin treatment and survival in patients with cancer. JAMA Oncol 2018;4(1):63−70. https://doi.org/10.1001/jamaoncol.2017.2752.
[63] Lee H-J, Wong JB, Jia B, Qi X, DeLong ER. Empirical use of causal inference methods to evaluate survival differences in a real-world registry vs. those found in randomized clinical trials. Stat Med 2020;39:3003−21. https://doi.org/10.1002/sim.8581.
[64] van der Burg ME, Lammes FB, Verweij J. The role of CA 125 in the early diagnosis of progressive disease in ovarian cancer. Ann Oncol 1990;1(4):301−2. https://doi.org/10.1093/oxfordjournals.annonc.a057754.
[65] Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J 2021;63(3):528−57. https://doi.org/10.1002/bimj.201900297.
[66] Didelez V, Stensrud MJ. On the logic of collapsibility for causal effect measures. Biom J 2022;64(2):235−42. https://doi.org/10.1002/bimj.202000305.
[67] Robins JM, Hernán MA, Siebert U. Estimations of the effects of multiple interventions. In: Ezzati M, Lopez AD, Rodgers A, Murray CJL, editors. Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors. Geneva, Switzerland: World Health Organization; 2004:2191−230.
[68] Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol 2017;46:756−62. https://doi.org/10.1093/ije/dyw323.
[69] Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 1992;3(4):319−36. https://doi.org/10.1097/00001648-199207000-00007.
[70] Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med 2012;31:2000−9. https://doi.org/10.1002/sim.5316.
[71] Siebert U, Hernan M, Robins J. Monte Carlo simulation of the direct and indirect impact of risk factor interventions on coronary heart disease. An application of the g-formula (Abstract). Sicily, Italy: Society for Medical Decision Making; 2002.
[72] Kuehne F, Schomaker M, Stojkov I, Jahn B, Conrads-Frank A, Siebert S, et al. Scoping review: causal evidence in health decision making: methodological approaches of causal inference and health decision science. HTA report vol. 509-1. GMS German Med Sci 2022;20:1−25.
[73] Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;64:380−2.
[74] Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR good research practices for retrospective database analysis task force report–part III. Value Health 2009;12(8):1062−73. https://doi.org/10.1111/j.1524-4733.2009.00602.x.
[75] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
[76] Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB. Introduction to causal diagrams for confounder selection. Respirology 2014;19(3):303−11. https://doi.org/10.1111/resp.12238.
[77] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Burlington, MA: Morgan Kaufmann Publishers Inc.; 1988.
[78] Sylvestre MP, Huszti E, Hanley JA. Do OSCAR winners live longer than less successful peers? A reanalysis of the evidence. Ann Intern Med 5 2006;145:361−3. https://doi.org/10.7326/0003-4819-145-5-200609050-00009. discussion 392.
[79] Crowley J, Hu M. Covariance analysis of heart transplant: survival data. J Am Stat Assoc 1977;72:27−36.
[80] Gail MH. Does cardiac transplantation prolong life? A reassessment. Ann Intern Med 1972;76:815−7.