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, MSc, Marjan Arvandi, PhD, Lisa M. Hess, PhD, Douglas E. Faries, PhD, Raffaella Matteucci Gothe, Dipl.-Stat., Holger Gothe, MD, Julie Beyrer, MPH, MTSC, Alain Gustave Zeimet, MD, Igor Stojkov, MPharm, Nikolai Mühlberger, DVM, MPH, Willi Oberaigner, PhD, Christian Marth, MD, Uwe Siebert, MD, MPH, MSc, ScD

PII: S0895-4356(22)00246-3
DOI: https://doi.org/10.1016/j.jclinepi.2022.10.005
Reference: JCE 10928

To appear in: Journal of Clinical Epidemiology

Received Date: 13 June 2022
Revised Date: 17 August 2022
Accepted Date: 3 October 2022


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting, and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc.
Causal analyses with target trial emulation for real-world evidence removed large self-inflicted biases: Systematic bias assessment of ovarian cancer treatment

Felicitas Kuehne, MSc¹, Marjan Arvandi, PhD¹, Lisa M Hess, PhD², Douglas E Faries, PhD², Raffaella Matteucci Gothe, Dipl.-Stat.¹, Holger Gothe, MD¹⁻³, Julie Beyrer, MPH, MTSc², Alain Gustave Zeimet, MD⁴, Igor Stojkov, MPHm⁵, Nikolai Mühlberger, DVM, MPH⁶, Willi Oberaigner, PhD¹⁵, Christian Marth, MD⁴ and Uwe Siebert, MD, MPH, MSc, ScD¹⁶,⁷

Affiliations:
¹ Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT – University for Health Sciences, Medical Informatics and Technology, Hall i.T., Austria
² Eli Lilly and Company, Indianapolis, IN, USA
³ Chair of Health Sciences / Public Health, Medical Faculty “Carl Gustav Carus”, Technical University Dresden, Dresden, Germany
⁴ Department of Obstetrics and Gynecology, Innsbruck Medical University, Innsbruck, Austria
⁵ Institute for Clinical Epidemiology, Cancer Registry Tyrol, Tirol Kliniken, Innsbruck, Austria
⁶ Center for Health Decision Science and Departments of Epidemiology and Health Policy & Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA
⁷ Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Email Addresses:
Felicitas Kuehne: felicitas.kuehne@umit-tirol.at
Marjan Arvandi: marjan.arvandi@umit-tirol.at
Lisa M Hess: hess_lisa_m@lilly.com
Douglas E Faries: faries_douglas_e@lilly.com
Raffaella Matteucci Gothe: raffaella.gothe@umit-tirol.at
Holger Gothe: holger.gothe@umit-tirol.at
Julie Beyrer: beyrer_julie_kay@lilly.com
Alain Gustave Zeimet: alain.zeimet@tirol-kliniken.at
Igor Stojkov: igor.stojkov@umit-tirol.at
Nikolai Mühlberger: nikolai.muehlberger@umit-tirol.at
Willi Oberaigner: wilhelm.oberaigner@umit-tirol.at
Christian Marth: christian.marth@tirol-kliniken.at
Uwe Siebert: usiebert@hsph.harvard.edu

Corresponding Author:
Uwe Siebert, MD, MPH, MSc, ScD
Adjunct Professor of Epidemiology and Health Policy & Management, Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA.
Professor of Public Health, Medical Decision Making and Health Technology Assessment, Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT – University for Health Sciences, Medical Informatics and Technology, Eduard-Wallnoefer-Zentrum 1, A-6060 Hall i.T., Austria.
Phone: +43 (0)50 8648 3930, Fax: +43 (0)50 678648
Email: usiebert@hsph.harvard.edu

Acknowledgment: We thank Zbigniew Kadziola for his continuing advice regarding statistical methods and SAS coding including the coding for the bootstrap analysis.
Independent research grant by Lilly Research Award Program (LRAP) and partly funded by the state Tyrol
Keywords: causal inference; comparative effectiveness; longitudinal data; electronic health records; target trial; inverse probability weighting

1. Abstract

Background: 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 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 (MSCM) with replicates emulating a reference randomized controlled trial. 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 a HR for second-line versus 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 a HR of 1.12 (95%CI: 0.96-1.28).

Conclusions: 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.

Keywords: causal inference; comparative effectiveness; longitudinal data; electronic health records; target trial; inverse probability weighting
2. 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 versus never’ treatment comparisons. It can be difficult to understand whether patients live longer because they receive a particular treatment or whether patients receive that treatment because they lived longer.\(^3-5,7,14\) The target trial approach is a structural approach emulating a 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 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.

3. Methods

To estimate the impact of potential biases when analyzing RWD, we created and followed a causal analytic framework prior to the data analysis. We 1) used the case of ovarian cancer, 2) identified potential biases using a causal graph (Figure 1), 3) judged the direction of potential biases based on expert assumptions encoded in the causal graph following the techniques described by VanderWeele and colleagues 8(Table 1), 4) selected a published randomized clinical trial 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.
Figure 1. Causal Diagram including measured and unmeasured confounding
Ca-125: Cancer-Antigen 125; LOT1: first line of chemotherapy; LOT2: second-line of chemotherapy; CT: computer tomography; tx: treatment

The causal diagram is a simplified version of a 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 be adequately measured variables; striped boxes indicate variables that are not present in the database.

Table 1. Expert panel assessment of assumed bias direction

<table>
<thead>
<tr>
<th>Bias</th>
<th>Direction of bias (pro or contra LOT2)</th>
<th>Estimation of HR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HR -</td>
</tr>
<tr>
<td>Confounding</td>
<td></td>
<td>in favor of LOT2</td>
</tr>
<tr>
<td>- Unmeasured</td>
<td>(disease severity, CT scan, symptoms)</td>
<td>X</td>
</tr>
<tr>
<td>- (education)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Time-independent</td>
<td>(ascites, stage)</td>
<td>X</td>
</tr>
<tr>
<td>- (age, comorbidities, time since LOT1)</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>- Time-dependent</td>
<td>(CA-125)</td>
<td>X</td>
</tr>
<tr>
<td>Immortal-time Bias</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Selection Bias / Confounding by indication</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

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

3.1. 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 second-line of therapy (LOT2) improve overall survival in patients with ovarian cancer who progressed after initially successful surgery and first-line chemotherapy. We expected to see time-independent, time-dependent confounding, selection bias, and immortal time bias. Furthermore, a published randomized controlled trial 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 US from January 2000 to June 2014 (see eAppendix A.1.).
We included female patients ages 18 years or older with ovarian, fallopian tube or primary peritoneal cancer.

Eligible patients must have disease progression after standard first-line chemotherapy treatment (LOT1).

Progression was defined as doubling value of CA-125 (details see eAppendix A.1).\(^\text{19}\)

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.

### 3.2. Selection of Reference Case/Gold Standard

We selected a reference (gold standard) study by Rustin and colleagues\(^\text{18}\) to compare our effect estimates to. The RCT estimated the benefit of early second-line chemotherapy in women with ovarian cancer and included women with ovarian cancer who had undergone surgery and first-line chemotherapy. Women were randomized to early treatment (second-line chemotherapy 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)\(^\text{18}\).

### 3.3. 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) versus 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. Analytic strategies

<table>
<thead>
<tr>
<th>Analytic Strategy</th>
<th>Intervention</th>
<th>Comparator</th>
<th>Controlling for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Confounding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Baseline</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time-Up</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unmeas</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Initial Time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Selection</td>
</tr>
<tr>
<td><strong>Crude</strong></td>
<td><strong>LOT2 anytime during follow-up</strong></td>
<td><strong>never receiving LOT2</strong></td>
<td></td>
</tr>
<tr>
<td>1 <em>Crude Cox</em></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>2 <em>Adjusted Cox</em></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>3 <em>Crude time-var. Cox</em></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>4 <em>Adjusted time-var. Cox</em></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td><strong>ITT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 <em>Target trial PP</em></td>
<td><strong>LOT2 immediately after progression</strong></td>
<td><strong>never receiving LOT2</strong></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>6 <em>Target trial causal PP</em></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td><strong>PP</strong></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td><strong>Causal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 <em>Partially emulated trial</em></td>
<td><strong>LOT2 immediately after progression</strong></td>
<td><strong>delayed LOT2 &gt; 6 wks after progression</strong></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>8 <em>Fully emulated trial</em></td>
<td></td>
<td></td>
<td><strong>X</strong></td>
</tr>
</tbody>
</table>
Time-dep: time-dependent; Unmeas: unmeasured; Imm Time: immortal-time; LOT2: second-line of 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 second-line chemotherapy at any time after progression to the overall survival of those never receiving second-line chemotherapy.
2. “Adjusted Cox”: Cox regression with adjustment for baseline confounding covariates, comparing the overall survival of patients receiving second-line chemotherapy at any time after progression to the overall survival of those never receiving second-line chemotherapy.
3. “Crude time-var. Cox”: Cox regression including treatment as time-varying covariate to compare (treated versus non-treated) person time rather than (ever treated versus 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 versus non-treated) person time rather than (ever treated versus 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” (IPCW): Performing a target trial as described but accounting for informative censoring by applying the inverse probability of censoring weighting (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 and colleagues, 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.

This figure 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 a 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 versus 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 time of progression, 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 versus never” treatment comparison, we compared (treated versus non-treated) person time rather than (ever treated versus never treated) individuals.
Each patient contributed her person time to the treatment she received to the corresponding time point, from here on called “treated versus untreated person time” comparison. 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 randomized controlled trial as described by Hernan, Robins, Cain, and others. 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 versus never LOT2. To account for natural time variation within RWD, we allowed for a lag time of six weeks (“grace period”) after diagnosis of progression. We refer to these adapted strategies as to “immediate versus never” treatment. In analysis 5 (“Target trial: PP”), we followed the target trial approach, 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 inverse probability of censoring weighting (IPCW). 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”. In sensitivity analyses, we assessed the robustness of results of the outcome model using different weight models (Table 3).

### 3.4. 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 and colleagues to harmonize the study protocols and study population (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”\textsuperscript{18}. The RCT protocol for the “delayed treatment” arm dictated the start of second-line chemotherapy 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, any initiation of LOT2all treatment not based on biomarker increase (i.e., six weeks after progression defined by biomarker increase) was considered delayed treatment. Hence, patients in the “delayed treatment” arm were artificially censored only during the first six 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.

3.5. 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 and colleagues\textsuperscript{8} (see Appendix A.2.2.).

4. Results

4.1. 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.
Figure 3. Flowchart of the included cohort

LOT2: second-line of 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 first-line chemotherapy. 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 lost to follow-up.

<table>
<thead>
<tr>
<th>Potentially included patients (n=3,688)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded:</td>
</tr>
<tr>
<td>- Other cancer</td>
</tr>
<tr>
<td>- Refractory patients</td>
</tr>
<tr>
<td>- LOT2 prior to progression</td>
</tr>
<tr>
<td>- Death at day of progression</td>
</tr>
<tr>
<td>Included patients (n=1,582)</td>
</tr>
<tr>
<td>Replicate assigned to immediate LOT2 (n=1,582)</td>
</tr>
<tr>
<td>Deaths</td>
</tr>
<tr>
<td>Censored (protocol violation)</td>
</tr>
<tr>
<td>Lost to follow-up</td>
</tr>
<tr>
<td>Replicate assigned to no LOT2 (n=1,582)</td>
</tr>
<tr>
<td>Deaths</td>
</tr>
<tr>
<td>Censored (protocol violation)</td>
</tr>
<tr>
<td>Lost to follow-up</td>
</tr>
</tbody>
</table>

4.2. 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 a 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 the Table 2.
Figure 4. Base-case results

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

Treatment allocation:
1. Ever versus 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 versus never: Comparing the survival of progressed patients who had received LOT2 within six weeks after progression to the survival of those who had not received LOT2 at any time during follow-up
3. Immediate versus delayed: Comparing the survival of progressed patients who had received LOT2 within six weeks after progression to the survival of those who had received LOT2 later than six weeks after progression or never

Analytic strategies:
1. “Crude Cox”: Univariable Cox regression without adjustment for covariates.*
2. “Adjusted Cox”: Cox regression with adjustment for baseline confounding covariates.*
3. “Crude time-var. Cox”: Univariable Cox regression including treatment as time-varying covariate to compare (treated versus non-treated) person time rather than (ever treated versus 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 versus non-treated) person time rather than (ever treated versus 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” (IPCW): Performing a target trial as described but accounting for informative censoring by applying the inverse probability of censoring weighting (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 and colleagues, 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.
* In analytic strategies 1 and 2, the proportional hazards assumption was violated. In these cases, in addition to the “average” hazard ratio, the initial hazard ratio of a model with interaction between linear time and treatment is reported.

The graph shows the hazard ratio and its 95% confidence interval for each analytic strategy and puts it into comparison to the hazard ratio of the reference case indicated by the dotted line and the grey area indicating the 95% confidence interval. A hazard ratio of one suggests no treatment effect; a hazard ratio below one suggests a beneficial treatment effect while a hazard ratio above one indicates a harmful treatment effect.

Table 2. Base Case Results with Bias Estimation

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>HR</th>
<th>95% Conf. Int.</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ever vs. Never</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. “Crude Cox”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without interaction of time and LOT2</td>
<td>0.56</td>
<td>0.49 - 0.64</td>
<td>45%</td>
</tr>
<tr>
<td>With interaction of time and LOT2*</td>
<td>0.27</td>
<td>0.22 - 0.34</td>
<td>73%</td>
</tr>
<tr>
<td>2. “Adjusted Cox”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without interaction of time and LOT2</td>
<td>0.51</td>
<td>0.44 - 0.59</td>
<td>50%</td>
</tr>
<tr>
<td>With interaction of time and LOT2*</td>
<td>0.25</td>
<td>0.21 - 0.31</td>
<td>75%</td>
</tr>
<tr>
<td><strong>“Treated vs. Untreated Person Time”</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. &quot;Crude time-var. Cox&quot;</td>
<td>1.41</td>
<td>1.22 - 1.64</td>
<td>-40%</td>
</tr>
<tr>
<td>4. &quot;Adjusted time-var. Cox&quot;</td>
<td>1.37</td>
<td>1.18 - 1.59</td>
<td>-36%</td>
</tr>
<tr>
<td><strong>Immediate vs. Never</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Target Trial Approach</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. &quot;Target trial PP&quot;</td>
<td>1.35</td>
<td>1.17 - 1.55</td>
<td>-33%</td>
</tr>
<tr>
<td>6. &quot;Target trial causal PP&quot; (IPCW)</td>
<td>1.38</td>
<td>1.22 - 1.63</td>
<td>-36%</td>
</tr>
<tr>
<td><strong>Immediate vs. Delayed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trial Emulation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. &quot;Partially emulated trial&quot; (IPCW)</td>
<td>1.26</td>
<td>1.15 - 1.37</td>
<td>-25%</td>
</tr>
<tr>
<td>8. &quot;Fully emulated trial&quot; (IPCW)</td>
<td>1.12</td>
<td>0.96 - 1.28</td>
<td>-10%</td>
</tr>
</tbody>
</table>
HR: hazard ratio; 95% Conf. Int.: 95% confidence interval; LOT2: second-line of 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.

* In analytic strategies 1 and 2, the proportional hazards assumption was violated. In these cases, in addition to the “average” hazard ratio, the initial hazard ratio of a model with interaction between linear time and treatment is reported.

4.3. 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 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 and colleagues (analysis 8).

4.4. 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 modelling time (base case as spline to linear time), looked at patients older than 65 years, changed the grace period from six 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%.

Table 3. Sensitivity Analyses

<table>
<thead>
<tr>
<th>Sensitivity Analyses</th>
<th>% change in effect estimate (HR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study time horizon</strong></td>
<td></td>
</tr>
<tr>
<td>5 years</td>
<td>4.9%</td>
</tr>
<tr>
<td>3 years</td>
<td>1.8%</td>
</tr>
<tr>
<td>modelling time as linear covariate</td>
<td>1.3%</td>
</tr>
<tr>
<td><strong>Study population only &gt;65 years</strong></td>
<td>1.4%</td>
</tr>
<tr>
<td>Grace period modeled as 4 weeks</td>
<td>2.1%</td>
</tr>
<tr>
<td>Delayed tx defined as minimum 3 times nadir and &gt;6 weeks after progression</td>
<td>0.3%</td>
</tr>
<tr>
<td>Weight function with linear time</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

HR: hazard ratio; tx: treatment
5. Discussion

We used the case of second-line chemotherapy in women with ovarian cancer to investigate potential biases that may occur when using observational real-world data for comparative effectiveness research. At times when real-world data are widely available, it is extensively debated how such data can be used for assessing comparative effectiveness outside of the artificial setting of randomized controlled trials. 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 real-world data 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 second-line chemotherapy after progression had a longer life expectancy than those that 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. Those techniques include visual, structural, and statistical approaches, which are validated in several study designs and therapeutic areas. Also, studies exist applying and comparing several analytic strategies to observational data to assess potential biases. One study compared results for patients eligible for a trial to those not eligible for that trial. 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 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 real-world data. 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. In our dataset, we did not have any information on progression indicated by CT scan 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 a RCT may not define a treatment strategy explicitly. He defined the delayed second-line chemotherapy strategy more broadly which is matched by our approach more closely 18.

Seventh, it must be noted that comparing conditional with marginal hazard ratios is comparing apples with oranges, as hazard ratios 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 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 2nd-line ovarian cancer chemotherapy, where some women may benefit and others do 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 first-line chemotherapy (every three 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 a RCT as the comparative gold standard. Further, we assessed the comparative effectiveness of second-line chemotherapy 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 second-line chemotherapy 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.

In the time of digitalization of health care data and “big” real-world data, further educational efforts on structural and statistical methods aiming for causal inference from RWD to inform healthcare 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 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.

6. Conclusion

We used the case example of second-line chemotherapy in women with progressive ovarian cancer to identify potential biases that may occur when applying different noncausal and causal analytic approaches to real-world 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 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.
7. References


44. Latimer NR, Abrams KR. Adjusting Survival Time Estimates in the Presence of Treatment Switching. unless otherwise stated; 2014.


8. Statements & Declarations

Highlights

- To assess potential biases in real-world evidence (RWE), this paper compares the HR of the reference trial to the estimated hazard ratios (HRs) following different analytic approaches.

- Biases resulting from the different analytic approaches varied in size and direction, ranging from 75% underestimating the HR to 36% overestimating the HR.

- The full causal analysis (including the target trial emulation using a marginal-structural-Cox-model) yielded the smallest bias, overestimating the HR by 10%.

- In RWE, a thorough causal-inference-based design and analysis is important.
What is new

- We systematically assessed type and impact of potential biases in real-world observational database analysis by applying a stepwise analytic approach ranging from simple crude to full causal analyses with trial emulation and comparing it to a reference randomized controlled trial.

- 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 adjustment method.

- This conceptual paper using a real-world case example offers a good overview of potential biases that may occur in real-world data analysis and provides a summary for causal inference tools and potential adjustment methods including causal graphs, target trial emulation and g-methods.

- The findings of this paper provide confidence in causal analytic methods that are currently discussed – if applied correctly and completely – and underline the need of carefully designing observational studies based on real world evidence.

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

Author Contributions

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.

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.

9. Supplementary Appendix

A. Methods

A.1. Data

We used data from the EMR database in the EHI DataMart as purchased from IMS Health. These data consist primarily of medium and large community-based oncology practices in the US. Each practice utilizes an Electronic Medical Record system capturing detailed, patient-level clinical data, which was then de-identified, assigned a synthetic ID, and integrated into the warehouse that consists of several linked hospital and claims databases. Each patient can be identified through a patented, HIPAA-compliant patient identifier.

For this analysis, we created a new database with all information that was needed from the different datasets to perform the analyses. Some information that was needed for our analyses was not available in a directly
compatible format. In these cases, we created new variables. The dataset for this analysis included data from more than 740,000 patient with cancer from January 2000 to June 2014, representing 737 facilities and patients from all 50 states.

We included female patients ages 18 years or older with ovarian, fallopian tube or primary peritoneal cancer. Eligible patients must have disease progression after standard first-line chemotherapy treatment (LOT1) that managed to lower the CA-125 value to a level within normal ranges (<35 U/mL) for 45 days after the last LOT1 dose. We defined progression via the biomarker CA-125 and followed the definition of the Gynecologic Cancer Inter Group (GCIG) as twice the nadir, that is, lowest CA-125 value during LOT1, or twice the upper limit of response range, that is, 35 U/mL. Progression defined according to these conditions by the biomarker CA-125 is from hereon called "progression" and "time of progression" refers to the date of the doctor visit that first detected progression.

We excluded patients with other cancers or whose LOT1 indicated a treatment for a cancer other than ovarian, peritoneal, or fallopian tube cancer, patients who progressed during the first 45 days after end of LOT1, and patients that received LOT2 before time of progression.

A.2. Directed Acyclic Graph (DAG)

DAGs are a simple but powerful tool to produce a network of causal variables. DAGs consist of variables and their direct causal effects represented as nodes and edges, respectively. The edges are directed and presented as arrows pointing in one direction. Missing edges symbolize the hypothesis of no direct causal effect. Besides nodes and edges, a DAG always includes at least one node with no "parent", that is, a variable causing the variable of interest, and at least one variable without any children, that is, variables that are caused by the variable of interest. Nodes that have no direct or indirect causal effect on the outcome and no direct or indirect causal effect on the exposure may be omitted from the graph.

DAGs visualize the research question and may help (1) identifying potential biases, especially confounding, (2) identifying variables that need to be controlled for, (3) determining the direction of the bias when it is not appropriately controlled for, and (4) determining which statistical methods are appropriate to perform the analysis.
VanderWeele and colleagues have shown that depending on the direction of the relations, one can estimate the direction of the bias \(^8\). For example, if an uncontrolled confounder-by-indication simultaneously increases the probability of the more effective intervention and the probability of the bad outcome, the estimated hazard ratio of the intervention on the outcome may be overestimated, leading to a bias against treatment effectiveness.

A.2.1. Identification of Potential Biases Using Causal Graphs

To identify potential biases and confounders, we established an expert panel of clinical and epidemiological specialists. Based on clinical and epidemiological substance matter discussions, this panel developed a causal diagram, more specifically, a DAG shown in Figure 1. We used the DAG to identify potential biases and variables that were needed for the analysis of the research question of the effect of LOT2 on overall survival. The panel determined whether those variables should have been included as baseline variables or as repeated measurements.

The DAG (Figure 1) shows that some parameters categorized as potential confounding variables were not or just sporadically available in the database. Where possible, we identified proxy variables. For example, we used CA-125 values as proxy variable to handle missing information on disease severity. However, there might be an additional independent effect of disease severity on survival. Further, the DAG shows that time-independent, time-dependent, and unmeasured confounding is present.

Besides confounding and confounding by indication, we concluded the presence of immortal time bias (patients have to live long enough to be eligible for treatment) as a potential bias.

A.2.2. Estimation of the Direction of Potential Biases using Causal Graphs

Our expert panel judged the direction of potential biases in the outcome estimation. Immortal time bias typically underestimates the treatment HR (i.e., bias in favor of the treatment) \(^14,62,78\). Other biases needed discussions and the expert panel used the DAG to estimate the direction of the bias and followed the techniques described by VanderWeele and colleagues \(^8\).

For the unmeasured factors and those that were just sporadically measured in the dataset, our expert panel estimated that both baseline and time-dependent confounding (abnormal CT scan results, disease severity, symptoms, and stage), likely lead to an overestimation of the HR, (i.e., bias against treatment). The only baseline
confounding caused by the unmeasured factor of education would likely underestimate the HR. The direction of uncontrolled confounding caused by comorbidities could not be clearly estimated. (See Table 1)

As a source for time-dependent confounding, we identified the biomarker CA-125. As the biomarker increases, the probability of dying as well as the probability of starting LOT2 increases. Simultaneously, the biomarker lies on the causal pathway of LOT2 on mortality. Applying the techniques described by VanderWeele and colleagues, we found that not controlling for the confounder CA-125 as well as incorrectly controlling for the biomarker using traditional methods would result in an overestimation of the HR, leading to a bias against LOT2.

Our expert panel believed that the direction of the immortal time bias is in favor of LOT2. Patients living long enough and having a prognosis suggesting a beneficial effect of the treatment are more likely to receive LOT2. In the ever versus never analytic strategy, women with the better prognosis would be allocated to the treatment arm yielding a HR overestimating the true treatment effect.

A.3. Target Trial

The target trial approach proposes to structure any data analysis as if one would design a randomized controlled trial as described by Hernan, Robins, Cain and colleagues. This implies carefully defining each trial component including research question, inclusion criteria (at time of study start), time horizon, compared treatment strategies, and treatment group assignment. Especially, the definition of treatment strategies and their assignment is crucial for a causal interpretation of the results. In a randomized controlled trial, the treatment strategy would most likely not be defined as “take a treatment at any point in time”. More likely, there would be a clear inclusion definition and at the time of inclusion, the study would immediately start with a clearly defined protocol. We included that approach into our analyses.

Another important component of the target trial approach is the assignment procedure. As in a randomized controlled trial, we would like to have comparable groups assigned to each treatment strategy. In observational data, patients cannot randomly receive a treatment, which would be necessary to perform an unbiased intention to treat (ITT) analysis. However, the patients can be randomly assigned to treatment groups, irrespective of the actual treatment received. A per protocol (PP) can then be conducted. Hernan and other suggest replicating the data, including all patients in each treatment arm, and performing a causal PP analysis and artificially censoring the
patients violating the protocol. However, censoring in this case is informative which needs to be reflected in the choice of statistical method. Inverse probability of censoring weighting (IPCW) is adjusting for informative censoring.

A.4. Inverse Probability of Censoring Weighting (IPCW)

We estimated the probability of following the protocol for each individual \(i\) at each time point \(t\) by fitting a weight model. Second, we then calculated stabilized weights for each individual at each time point. Finally, we fitted an outcome model including the weights to estimate the probability of survival.

As we did not have sufficient data to fit a model at each time point, we created seven-day intervals and fitted a pooled logistic regression model. The first interval started at day one as the first day of the trial. We assumed the day of randomization and treatment allocation as day 0. Within each interval, we indicated the time-varying variables. These are (1) CA-125 value (absolute value), (2) CA-125 value (change from nadir over time), (3) time since progression, and (4) artificial censoring or death. Time-varying variables such as the biomarker CA-125 are not routinely measured at each clinical visit. We populated the intervals with the value of the last measurement. We assumed that this reflects the knowledge of the physician. For each patient and interval, we estimated the probability of receiving LOT2 using pooled logistic regression, once with just baseline confounders, and once with both, baseline and time-varying confounders. We then used these estimators to estimate the probability of following the protocol \(P(\text{comply})\) for each individual at each time point. For the different comparative strategy, \(P(\text{comply})\) had to be calculated differently. (See Table 4)

Table 4. Weight function at corresponding strategies and intervals

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Interval</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never LOT2 strategy</td>
<td>All intervals</td>
<td>((1-P(\text{tx}</td>
</tr>
<tr>
<td>Immediate LOT2 strategy</td>
<td>Interval (week) 1-5:</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Interval (week) 6:</td>
<td>(P(\text{tx}</td>
</tr>
<tr>
<td></td>
<td>Interval (week) &gt;6:</td>
<td>1</td>
</tr>
<tr>
<td>Delayed LOT2 strategy</td>
<td>Interval (week) 1-6:</td>
<td>((1-P(\text{tx}</td>
</tr>
<tr>
<td></td>
<td>Interval (week) &gt;6:</td>
<td>1</td>
</tr>
</tbody>
</table>

\(tx\): receiving treatment; \(V\): vector of baseline confounders; \(L_{(k)}\): vector of time-dependent confounding at time \(k\).
For the weights, we estimated the probability of following the protocol ($P(\text{comply})$) up to the given interval for each individual, by multiplying together all the probabilities of following the protocol up to interval $t$ ($\prod_{k=0}^{t} P(t \text{ comply})$), including and excluding time-varying confounders. To calculate the stabilized weights, we divided the weights using the function without time-varying confounders by the weights using the function with time-varying confounders.

For the outcome model, we fitted a pooled logistic regression model predicting dying in the next interval using the stabilized weights for each interval and adjusting only for time-varying confounders. We used robust variance estimators to adjust for dependence between (1) different intervals of patients and (2) the two replicates generated from each patient.

As very high weights (e.g., > 100) may generate unstable results, we assessed the distribution of estimated weights. Different weight models were used to assess robustness of results, in sensitivity analyses. Similar techniques with inverse probability of censoring 2014 weighting have been described and applied for adjusting for treatment switching in randomized clinical trials.

A.5. 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 that is described in the following paragraphs.

A.5.1. Simple "Ever versus Never" Comparison

In analysis 1 in the “ever versus never” comparison, we applied a simple univariable Cox regression for overall survival without adjustment for covariates (“Crude Cox”). We provided Kaplan Meier curves and results from a univariable Cox proportional hazards model. The Cox proportional hazard assumption was tested by inspecting the log(-log(survival)) curves as a function of time (log scale). If the assumption of proportional hazards was violated, an interaction between treatment and time to model a time-dependent treatment effect was included. In analysis 2 (“Adjusted Cox”), we controlled for baseline confounders (i.e., age, nadir, CA-125 at time of progression, time since first-line treatment) by including them as covariates into the Cox model.
A.5.2. “Treated versus Untreated Person Time” Comparison

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 Analysis 2. Additionally, to treatment, CA-125 value was included as time-varying covariate as it changed over time.

A.5.3. “Immediate versus Never” Target Trial

In analysis 5 ("Target trial: PP"), we followed the target trial approach \(3,5,16,27,28\), which estimated the per protocol effect. We applied a Cox model to compare the “immediate versus never” treatment strategy regarding overall survival. 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. This approach is called “artificial censoring”. In other words, the data of each patient were copied. Copy 1 was assigned to “immediate LOT2” and copy 2 was assigned to “never LOT2”. In each treatment arm and for each patient, only the person time and related information was used that was compatible with the assigned treatment at each time point. In the “immediate treatment” arm, patients were censored at six weeks after progression (i.e., grace period) if they had not started LOT2 at that time point. In the “never treatment” arm, patients were censored at the time they started LOT2.

In analysis 6 ("Target trial: causal PP"), we considered the fact that artificial censoring is potentially informative. The causal graph suggests that LOT2 initiation is driven by the biomarker CA-125 which also determines the outcome (i.e., overall survival). Hence, we applied a marginal structural Cox model adjusting for informative censoring by inverse probability of censoring weighting (IPCW) \(4,11,28-31\). First, we estimated the probability of following the protocol for each individual \(i\) at each time point \(t\) by fitting a weight model. Second, we used the weight model to calculate stabilized weights for each individual at each time point. Finally, we fitted a pooled logistic regression outcome model including the stabilized weights to estimate the probability of dying in the next interval, approximating a marginal structural Cox model. We calculated the 95%CI using bootstrapping to adjust for dependence between 1) different intervals of patients and 2) the two replicates generated from each subject. As very large weights (e.g., > 100) may generate unstable results, we assessed the distribution of weights. In sensitivity analyses, we assessed the robustness of results of the outcome model using different weight models \(33-35\) (Table 3).
B. Results

B.1. Descriptive Analysis

We performed descriptive analyses for the patients receiving LOT2 at any time and for patients never receiving LOT2 (ever versus never). The majority (74%) received platinum and taxane-based first-line chemotherapy. Another 12% of patients received single-agent platinum, 0.3% had a combination of platinum and a non-taxane based treatment, and 13% had another first-line chemotherapy regimen. The probability of first-, second-, and third-year survival was 49%, 20%, and 8%, respectively.

The results are shown in Table 5.
Table 5. Descriptive statistics

<table>
<thead>
<tr>
<th>Age</th>
<th>Ever Tx (N=1,255)</th>
<th>Never Tx (N=327)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>n (%)</td>
</tr>
<tr>
<td>&lt;30 years N (%)</td>
<td>2 (0.16)</td>
<td>1 (0.31)</td>
</tr>
<tr>
<td>30-55 years N (%)</td>
<td>192 (15.3)</td>
<td>50 (15.3)</td>
</tr>
<tr>
<td>56-65 years N (%)</td>
<td>338 (26.9)</td>
<td>62 (19.0)</td>
</tr>
<tr>
<td>&gt;65 years N (%)</td>
<td>723 (57.6)</td>
<td>214 (65.4)</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>67 (11)</td>
<td>68 (12)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>68 (15)</td>
<td>71 (16)</td>
</tr>
<tr>
<td>1st-line treatment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single agent platinum</td>
<td>144 (11.5)</td>
<td>51 (15.6)</td>
</tr>
<tr>
<td>Combination platinum (no taxane)</td>
<td>3 (0.24)</td>
<td>2 (0.61)</td>
</tr>
<tr>
<td>Platinum and taxane based</td>
<td>954 (76.0)</td>
<td>216 (66.1)</td>
</tr>
<tr>
<td>Other</td>
<td>154 (12.3)</td>
<td>58 (17.7)</td>
</tr>
<tr>
<td>Baseline CA-125 value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>334 (983)</td>
<td>249 (506)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>96 (171)</td>
<td>87 (129)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>263 (21.0)</td>
<td>61 (18.7)</td>
</tr>
<tr>
<td>Comorbidity Charlson Score (≥3)</td>
<td>144 (11.5)</td>
<td>32 (9.8)</td>
</tr>
<tr>
<td>Ascites</td>
<td>276 (22.0)</td>
<td>42 (12.8)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>249 (19.8)</td>
<td>72 (22.0)</td>
</tr>
<tr>
<td>Progression free survival after LOT1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;6 months N (%)</td>
<td>559 (44.5)</td>
<td>138 (42.2)</td>
</tr>
<tr>
<td>6-11 months N (%)</td>
<td>425 (33.9)</td>
<td>93 (28.4)</td>
</tr>
<tr>
<td>12-24 months N (%)</td>
<td>200 (15.9)</td>
<td>62 (19.0)</td>
</tr>
<tr>
<td>&gt;24 months N (%)</td>
<td>71 (5.7)</td>
<td>34 (10.4)</td>
</tr>
<tr>
<td>Mean days (SD)</td>
<td>270(251)</td>
<td>330(349)</td>
</tr>
<tr>
<td>Median days (IQR)</td>
<td>200(217)</td>
<td>211(297)</td>
</tr>
<tr>
<td>Treatment free survival</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean days (SD)</td>
<td>350 (287)</td>
<td>608 (569)</td>
</tr>
<tr>
<td>Median days (IQR)</td>
<td>270 (264)</td>
<td>406 (573)</td>
</tr>
<tr>
<td>Survival - 5years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 year N (%)</td>
<td>706 (56.2)</td>
<td>77 (23.5)</td>
</tr>
<tr>
<td>2 year N (%)</td>
<td>290 (23.1)</td>
<td>33 (10.1)</td>
</tr>
<tr>
<td>3 year N (%)</td>
<td>109 (8.7)</td>
<td>18 (5.5)</td>
</tr>
<tr>
<td>4 year N (%)</td>
<td>56 (4.5)</td>
<td>9 (2.8)</td>
</tr>
<tr>
<td>5 year N (%)</td>
<td>21 (1.7)</td>
<td>4 (1.2)</td>
</tr>
<tr>
<td>Median days (IQR)</td>
<td>419 (490)</td>
<td>116 (312)</td>
</tr>
</tbody>
</table>

Tx: treatment; SD: standard deviation; IQR: interquartile range; CA-125: cancer-antigen 125; LOT1: first-line therapy;
B.2. Violated constant hazard assumption

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. After accounting for the observed violation of the proportional hazards’ assumption, the HR of the treatment at time zero was estimated to be 0.27 (95%CI: 0.22-0.34) with a treatment-time interaction HR greater than one indicating an increasing treatment HR over time. When adjusting for baseline confounding (analysis 2), those HRs, assuming and not assuming constant HRs, were 0.51 (95%CI: 0.44-0.59) and 0.25 (95%CI: 0.21-0.31), respectively.

B.3. Weight Distribution

The distribution of the weights was assessed. The stabilized weights ranged between 0.43 and 1.45 with a mean and median of 1 and a standard deviation of 0.1. We felt comfortable using these weights in the outcome model expecting to generate stable results. Those HRs, assuming and not assuming constant HRs, were

B.4. Mapping the Target Trial Population

To make the study population of our cohort comparable with the reference RCT study population, we standardized results to the several features. For example, inclusion/exclusion criteria were mapped as much as possible to the same criteria as in the Rustin trial. However, some differences persist. As typical for using secondary databases, we used data that provided the diagnoses as ICD-9 codes and dates, doctor visit dates, medications, and dates of treatment administration. Hence, treatments were not explicitly related to the diagnoses. We were not able to ensure that all patients received a surgery prior to first-line chemotherapy, but we assumed that the recommendations were followed. Moreover, to ensure that the chemotherapy under assessment has been administered for ovarian cancer, we excluded other cancers and patients, where the drug combination indicated the use for other cancer treatments than ovarian cancer treatment. Also, we excluded patients receiving the treatment prior to the documented progression. These exclusions were necessary for linking the treatment under assessment (LOT2) to the diagnosis of interest (progressive ovarian cancer). Such data restrictions were used to improve data quality and comparability to the reference RCT. Using standardization enabled us to match for age between the populations, however, as survival of older patients is shorter, our mean follow-up remained slightly shorter than in the Rustin trial.
C. **Interpretation of Bias Magnitude and Direction**

In the case of LOT2 in progressive ovarian cancer, we estimated that confounding (time-independent, time-dependent, and unmeasured) was mainly biasing the results against LOT2 while immortal time bias was biasing the results in favor of LOT2. Further, we found that in this case, immortal time bias played a major role. Controlling for immortal time bias by including time-varying covariates in the Cox model and by applying the target trial concept with g-methods yielded almost the same results. However, the target trial concept should be the preferred strategy as it is following the counterfactual approach and is also controlling for time-dependent confounding.

D. **Target Trial Approach in Context**

When analyzing observational data, it is essential that the study design reflects the counterfactual causal approach, which includes the fact that the time of treatment strategy allocation reflects the actual decision point as well as the start of the study. Aligning the decision point, treatment allocation, determination of eligibility, and the start of the study is assured by the target trial and prevents immortal time bias. The start of the actual (observed) treatment may be triggered by the course of disease, for example by progression, occurrence of symptoms, comorbidities, or any other event. In some cases, this treatment start may be triggered by the availability of a scarce treatment. In transplantation medicine, for example, this is the time point when the transplant is available. However, if the treatment triggering event, in this example the availability of the transplant, is not collected for all patients receiving or not receiving the transplant in observational data, a target trial approach may not be feasible. In the case of scarce treatments, the application of time-varying covariates may be a better choice than the target trial if no residual time-dependent confounding is present. However, one would have to acknowledge and discuss the disadvantages of not following the counterfactual approach.

In the case of second-line chemotherapy in women with progressed ovarian cancer, the research question is ideal for applying the target trial approach as the true question is a dynamic treatment question. Physicians have diverse opinions about when second-line chemotherapy should be provided (when the biomarker indicates progression, when progression is visible in the CT scan, when symptoms occur, or never). However, unfortunately, we could not compare all those strategies as no data on CT scans or symptoms were available. In analyses 7 and 8, we therefore summarized and approximated the treatment start at suggestive CT scan or symptoms to a delayed treatment strategy. We recognize that CA-125 indicates a progression before it shows in a CT scan or symptoms and delayed
treatment starts would be based on new disease information. This information would most likely come from CT scans or symptoms.

E. SAS Code

Analysis 1
Title1 'Ever vs. Never';

Title2 'Crude Cox; Without interaction of time and LOT2';

proc phreg data = studyPop;
model survival_studyend*death_indicator_new(0) = EverVsNever_Tx/ ties = efron RL;
run;

Title 'Crude Cox; With interaction of time and LOT2';

proc phreg data = studyPop;
model survival_studyend*death_indicator_new(0) = EverVsNever_Tx  int_EverVsNever_yrs/ ties = efron RL;
int_EverVsNever_yrs = survival_studyend * EverVsNever_Tx;
label int_EverVsNever_yrs='survival(years) x 2nd LOT';
run;

Analysis 2
Title1 'Adjusted Cox';

Title2 'Without interaction of time and LOT2';

proc phreg data = studyPop;
class nadir_cat(ref='1') stage_cat(ref='5');

model survival_studyend*death_indicator_new(0) = EverVsNever_Tx std_CA_value_at_prog
time_since_lot1_month stage_cat age_progression_10yrs nadir_cat
/ ties = efron RL;
run;

Title 'With interaction of time and LOT2';

proc phreg data =studyPop;
class nadir_cat(ref='1') stage_cat(ref='5');

model survival_studyend*death_indicator_new(0) = EverVsNever_Tx int_EverVsNever_yrs std_CA_value_at_prog
time_since_lot1_month stage_cat age_progression_10yrs nadir_cat / ties = efron RL;
int_EverVsNever_yrs =survival_studyend *EverVsNever_Tx;
label int_EverVsNever_yrs='survival(years) x 2nd LOT';
run;

Analysis 3
Title1 'Time-var. Cox';
Title2 'Crude time-var. Cox';

proc phreg data =studyPop;

model survival_studyend*death_indicator_new(0) = time_dep_Tx /rl ties=efron;

if (timeToLOT2_new =. or survival_studyend < timeToLOT2_new) then time_dep_Tx=0;
else time_dep_Tx=1;

run;

Analysis 4
Title2 'Adjusted time-var. Cox';

proc phreg data=studyPop;

class nadir_cat (ref='1') stage_cat(ref='5');

model survival_studyend*death_indicator_new(0) = time_dep_Tx std_CA_value_at_prog time_since_lot1_month stage_cat age_progression_10yrs nadir_cat STD_CA125_Value / ties = efron RL;

if (timeToLOT2_new = . or survival_studyend < timeToLOT2_new) then time_dep_Tx=0;

else time_dep_Tx=1;

STD_CA125_Value=0;

array CA125Value_std (*) STDCA125value1 - STDCA125value102;

array CA125time (*) getCA125time1 - getCA125time102;

do i=1 to 101;

if CA125time[1] = . then STD_CA125_Value=0;else

if survival_studyend < CA125time[1] then STD_CA125_Value=0;

else if survival_studyend >= CA125time[i] and CA125time[i] ne . then STD_CA125_Value=CA125value_std[i];

else do i=1 to dim(CA125value_STD);
if CA125time[i] ne . and CA125time[i] <= survival_studyend < CA125time[i+1] then STD_CA125_Value = CA125value_std[i];
end;
end;

label STD_CA125_Value='CA125 (SD in U/mL)';
run;

Analysis 5
Title1 'Immediate vs. Never';
Title2 'Target Trial Approach';
Title3 'Target trial PP';

proc surveylogistic data=causal_naive_duplicates_3;
effect timespline = spline(int_start_day / naturalcubic details knotmethod=percentiles(5));
model deathind2_new (event='1') = EarlyTX_cloned timespline age_progression_10yrs STD_CA_Value_at_Prog
dummy_nadir_0 dummy_nadir_2 dummy_nadir_3 time_since_lot1_month dummy_stage_1 dummy_stage_2
dummy_stage_3 dummy_stage_4/PARMLABEL;;
cluster patient_id;
where Censorint = 0;
run;
Analysis 6

Title 'Weight Calculation';

**proc logistic** data=NeverTreatStrat;

effect timespline = spline(int_start_day_Inmonth / naturalcubic details knotmethod=percentiles(5));

model lot2int (event='1') = timespline age_progression_10yrs STD_CA_Value_at_Prog dummy_nadir_0
dummy_nadir_2 dummy_nadir_3
dummy_stage_1 dummy_stage_2 dummy_stage_3 dummy_stage_4 time_since_lot1_month /PARMLABEL;;

output out=base_splinetime predicted=predbase_splinetime;

**proc logistic** data=NeverTreatStrat;

effect timespline = spline(int_start_day_Inmonth / naturalcubic details knotmethod=percentiles(5));

model lot2int (event='1') = timespline CA125Int age_progression_10yrs STD_CA_Value_at_Prog dummy_nadir_0
dummy_nadir_2 dummy_nadir_3
dummy_stage_1 dummy_stage_2 dummy_stage_3 dummy_stage_4 time_since_lot1_month /PARMLABEL;;

output out=full_splinetime predicted=predfull_splinetime;

run;

data wt_splinetime_Never_fin;

set wt_splinetime_Never;

by patient_id;

retain swt_splinetime_never;

if first.patient_id then do; swt_splinetime_never = 1; end;
swt_splinetime_never = swt_splinetime_never * (1 - predfull_splinetime) / (1 - predbase_splinetime);

swt_splinetime_never_corr = swt_splinetime_never * 0.5;

run;

proc sql;
create table weight_early as
select
s.*,
case when s.int_start_day=42 then
(int_1.predfull_splinetime
*int_2.predfull_splinetime
*int_3.predfull_splinetime
*int_4.predfull_splinetime
*int_5.predfull_splinetime
*int_6.predfull_splinetime
*int_7.predfull_splinetime)/
(int_1.predbase_splinetime
*int_2.predbase_splinetime
*int_3.predbase_splinetime
*int_4.predbase_splinetime
*int_5.predbase_splinetime
*int_6.predbase_splinetime
*int_7.predbase_splinetime)/
*int_5.predbase_splinetime

*int_6.predbase_splinetime

*int_7.predbase_splinetime)

else 1

end as swt_splinetime_early,

calculated swt_splinetime_early/2 as swt_splinetime_early_corr

from wt_splinetime_early s

left join wt_splinetime_early int_1 on s.patient_id = int_1.patient_id and int_1.int_start_day = 0

left join wt_splinetime_early int_2 on s.patient_id = int_2.patient_id and int_2.int_start_day = 7

left join wt_splinetime_early int_3 on s.patient_id = int_3.patient_id and int_3.int_start_day = 14

left join wt_splinetime_early int_4 on s.patient_id = int_4.patient_id and int_4.int_start_day = 21

left join wt_splinetime_early int_5 on s.patient_id = int_5.patient_id and int_5.int_start_day = 28

left join wt_splinetime_early int_6 on s.patient_id = int_6.patient_id and int_6.int_start_day = 35

left join wt_splinetime_early int_7 on s.patient_id = int_7.patient_id and int_7.int_start_day = 42

order by s.patient_id, s.int_start_day;

quit;

data OutcomeModel_earlyNever;

set OutcomeModel_earlyNever;

swt_splinetime_Overall_corr= swt_splinetime_early_corr;
if swt_splinetime_Overall_corr=. then swt_splinetime_Overall_corr=swt_splinetime_never_corr;

run;

Title1 'Immediate vs. Never';

Title2 'Target Trial Approach';

Title3 'Target trial causal PP (IPCW');

**proc surveylogistic** data=Outcomemodel_earlyNever;

effect timespline = spline(int_start_day / naturalcubic details knotmethod=percentiles(5));

model deathint (event='1') = EarlyTX_cloned timespline age_progression_10yrs STD_CA_Value_at_Prog
dummy_nadir_0 dummy_nadir_2 dummy_nadir_3 time_since_lot1_month dummy_stage_1 dummy_stage_2
dummy_stage_3 dummy_stage_4 /PARMLABEL;;

weight swt_splinetime_Overall_corr;

where Censorint = 0;

cluster patient_id;

run;
Potentially included patients (n=3,688)

Excluded:
- Other cancer n=235
- Refractory patients n=1391
- LOT2 prior to progression n=435
- Death at day of progression n=45

Included patients (n=1,582)

<table>
<thead>
<tr>
<th>Replicate assigned to immediate LOT2 (n=1,582)</th>
<th>Replicate assigned to no LOT2 (n=1,582)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths</td>
<td>Deaths</td>
</tr>
<tr>
<td>Censored (protocol violation)</td>
<td>Censored (protocol violation)</td>
</tr>
<tr>
<td>Lost to follow-up</td>
<td>Lost to follow-up</td>
</tr>
<tr>
<td>n=277</td>
<td>n=660</td>
</tr>
<tr>
<td>n=1229</td>
<td>n=863</td>
</tr>
<tr>
<td>n=76</td>
<td>n=59</td>
</tr>
</tbody>
</table>
### Analytic Strategies

**Ever vs. Never**

1. **“Crude Cox”**
   - Without interaction of time and LOT2
   - With interaction of time and LOT2

2. **“Adjusted .Cox”**
   - Without interaction of time and LOT2
   - With interaction of time and LOT2

**Treated vs. Untreated Person Time**

3. "Crude time-var. Cox"
4. "Adjusted time-var. Cox"

**Immediate vs. Never**

5. "Target trial PP"
6. "Target trial causal PP" (IPCW)

**Immediate vs. Delayed**

7. "Partially emulated trial" (only strategies)
8. "Fully emulated trial" (strategies, population)
<table>
<thead>
<tr>
<th>Analytic Strategy</th>
<th>Intervention</th>
<th>Comparator</th>
<th>Controlling for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>never receiving LOT2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crude</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 “Crude Cox”</td>
<td>LOT2 anytime during follow-up</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 “Adjusted Cox”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 “Crude time-var Cox”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 “Adjusted time-var Cox”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 “Target trial PP”</td>
<td>LOT2 immediately after progression</td>
<td>never receiving LOT2</td>
<td></td>
</tr>
<tr>
<td>6 “Target trial causal PP” (IPCW)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 “Partially emulated trial” (only strategies)</td>
<td>LOT2 immediately after progression</td>
<td>delayed LOT2 &gt; 6 wks after progression</td>
<td></td>
</tr>
<tr>
<td>8 “Fully emulated trial” (strategies, population)</td>
<td>LOT2 immediately after progression</td>
<td>delayed LOT2 &gt; 6 wks after progression</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bias</th>
<th>Confounding</th>
<th>Time-step</th>
<th>Unmeas</th>
<th>Arm Time</th>
<th>Sibling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Causal
Highlights

- To assess potential biases in real-world evidence (RWE), this paper compares the HR of the reference trial to the estimated hazard ratios (HRs) following different analytic approaches.
- Biases resulting from the different analytic approaches varied in size and direction, ranging from 75% underestimating the HR to 36% overestimating the HR.
- The full causal analysis (including the target trial emulation using a marginal-structural-Cox-model) yielded the smallest bias, overestimating the HR by 10%.
- In RWE, a thorough causal-inference-based design and analysis is important.

What is new

- We systematically assessed type and impact of potential biases in real-world observational database analysis by applying a stepwise analytic approach ranging from simple crude to full causal analyses with trial emulation and comparing it to a reference randomized controlled trial.
- 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 adjustment method.
- This conceptual paper using a real-world case example offers a good overview of potential biases that may occur in real-world data analysis and provides a summary for causal inference tools and potential adjustment methods including causal graphs, target trial emulation and g-methods.
- The findings of this paper provide confidence in causal analytic methods that are currently discussed – if applied correctly and completely – and underline the need of carefully designing observational studies based on real world evidence.
- It is important to increase the knowledge about causal analytic frameworks that go beyond simple regression analyses in clinical research as well as in guideline development and health technology assessment, to ultimately make sure patients receive treatments with causally substantiated benefits that outweigh the harms.
Declaration of Interest

**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)

**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 Contributions**
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.

**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.
Author Contributions

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.