DRAFT, August 2025

A benchmark causal inference (ABCi) data set: simulated data on time-varying treatments, confounders and outcomes based on patients with type 2 diabetes

Ruth Keogh[1], Nan van Geloven[2], Daniala Weir[3].

[1] Medical Statistics Department, London School of Hygiene & Tropical Medicine, London, UK.
[2] Leiden University Medical Center, Leiden, The Netherlands.
[3] Utrecht University, Utrecht, The Netherlands.

1. Introduction

It is increasingly recognised that observational or 'real world' data provide opportunities to address questions about causal effects of medical interventions. Sources of such data include electronic health records, patient registries, administrative health databases and insurance claims databases. While randomized controlled trials (RCTs) remain the gold standard for evaluation of effects of treatments, they are not always feasible or ethical, and not able to address all questions of interest. Observational data enable, under certain assumptions, investigation of effects of interventions in large patient populations that are observed in real clinical practice, and that are more diverse than would be included in an RCT. Such data also allow for investigations into whether effects of interventions differ according to patient characteristics (treatment effect heterogeneity), which are not usually well-powered in RCTs, and investigations of outcomes observed over long-term follow-up.

There is an extensive and rapidly growing literature on causal inference methods that enable us to address causal questions, such as about the effects of treatments, using observational data, under certain assumptions. The increasing availability and complexity of these methods raises challenges for methodological researchers as well as for those who may be end users of such methods and those wishing to learn about them. Firstly, it is important to make comparisons between the performance of different methods and to assess their suitability for addressing different causal questions, but there is a lack of detailed and neutral comparisons. Secondly, there is a lack of openly accessible data that can be used to enable researchers to learn, and teach others, how to implement the methods and assess their suitability for the types of data they may face in practice. In this paper we introduce a simulated data resource that is designed to mimic complex longitudinal observational data. The data resource is intended to enable comparisons of methods for addressing a range of different types of causal question, as well as helping researchers to learn and teach how to implement methods.

Friedrich and Friede (2023) discussed the roles of benchmarking data sets and simulation studies in methods comparison investigations. A benchmarking data set typically refers to an openly accessible real data set to which different methods can be applied. Simulation studies, on the other hand, generate data according to a known data generating mechanism, ideally designed to test different features and assumptions of methods under comparison. An advantage of simulation studies is that the 'ground truth', i.e. the true value of the target of estimation, is known or can be generated. For assessment of causal inference methods knowing this ground truth is crucial to enable assessments of bias, which is in contrast to what is needed for comparisons of methods aimed purely at prediction, since comparisons of predictive performance using different methods can be assessed based on observed data alone. However, simulated data sets tend to be simpler than real world data sets and are sometimes designed such that certain analysis methods are favoured over others (Pawel et al. 2023, Nießl et al. 2023). Benchmarking data sets provide a more data-driven approach to comparisons, and have the advantage of reflecting the complexities of real world data, but the disadvantage that the ground truth cannot be known. Recommendations made by Friedrich and Friede (2023) to improve methods comparison studies and make them more ‘neutral’ included enhancing simulation studies with real data examples and conducting simulation studies based on real data. They also recommended establishing infrastructure, databases and gold standards within methods development communities, including by extending databases and data repositories.

The simulated data resource presented in this paper is referred to as the ABCi data set (A Benchmark Causal inference data set). It aims to mimic some of the complexities of real observational data, but by being entirely simulated it retains the advantage that the data-generating process is known, meaning that it is possible to generate the ground truth – in particular, the true causal effect of an intervention. The simulated data is based on a cohort of patients with type-2 diabetes (T2D) who have initiated treatment on Metformin, and may subsequently initiate second-line treatments. The data has the following features: (1) longitudinal data on use of four different second-line treatments, (2) time-fixed covariates and several longitudinal covariates of different types, (3) time-to-event outcomes for two competing events, and censoring. While the data generating mechanism for the ABCi data is known, we aimed to design it such that it does not naturally favour particular analysis methods applied with their standard settings, therefore enabling more neutral comparison studies.

The paper is organised as follows. In Section 2 we describe the motivating setting of patients with TD2, and outline several example causal research questions that could be addressed concerning choices about second line treatments. These range from simpler questions about point treatments and average treatment effects on overall mortality, to more complex questions about time-varying treatment strategies, conditional average effects, and predictions under interventions for outcomes involving competing risks. In Section 3 we outline how the data are generated and present descriptive statistics, with full details being provided in Supplementary materials. In Section 4 we illustrate the use of the data to compare estimators for example causal estimands. As part of this we explain how to generate corresponding ‘ground truth’ data sets in which counterfactual outcomes are provided for all individuals under the treatment strategies of interest. The paper concludes with a discussion in Section 5.

The ABCi data, the code used to generate it, and code to perform the illustrative analyses, are available at https://github.com/ruthkeogh/ABCi_data.

2. Motivating case-study: Second-line treatments in Type 2 Diabetes

The ABCi data is motivated by studies of the effects of different second-line treatments for people with T2D. When an individual is diagnosed with T2D, metformin monotherapy (an antihyperglycemic medication) is usually the first choice of treatment. Clinical guidelines recommend adding a second antihyperglycemic agent if HbA1c (a measure of glycemic control) is not well controlled using single therapy (NICE 2025). Medication classes that can be used as an add on treatment to metformin include Sulfonylureas (SU), Thiazolidinediones (TZDs), Dipeptidyl peptidase 4- inhibitors (DPP-4 i's), Glucagon Like Peptide-1 Receptor Agonists (GLP-1 RA) and Sodium Glucose Cotransporter-2 Inhibitors (SGLT2-i). In practice not all patients initiate metformin as their first line treatment, but we assume this in the data and focus on other treatments as potential add-ons. The choice of which add-on treatment to prescribe is dependent on patient characteristics such as: atherosclerotic cardiovascular disease, chronic kidney disease, heart failure, need for weight loss and risk of hypoglycemia, among others (NICE 2025). People with T2D are at increased risk of cardiovascular disease and it is of interest to investigate the impacts of different treatment choices on this outcomes, with different antihyperglycemic medications known to have different cardiovascular safety profiles. The simulated data includes two outcomes: major adverse cardiovascular event (MACE - defined as a combination of stroke, myocardial infarction, hospitalization for heart failure or cardiac death) and death from other causes.

There is a large literature on the impacts of second line treatments for T2D and many of these studies have used large scale observational data. The ABCi data set is based on a cohort of individuals with T2D who were initially treated with metformin, some of whom subsequently add second line treatments. Although we did not aim to mimic any specific real data set, the data generation was informed by published studies (see Supplementary Section S1). In the simulated data individuals can initiate one of four add-on treatments. Although we had specific treatments in mind, we label these as treatments A, B, C, D in the data to avoid any suggestion that any results derived from the data should be interpreted as real results relating to particular treatments. The start of follow-up for each individual is the date of initiating metformin, and each individual is followed for up to 5 years. The data set includes the following information on each individual: (1) time-fixed covariates recorded at the start of follow-up; (2) time-dependent covariates of different types, recorded every 30 days; (3) time to the initiation on one of four add-on treatments A, B, C, or D; (4) time of occurrence of MACE and death from other causes recorded in days relative to time zero; (5) time of censoring/loss to follow-up. Individuals are followed up until the earliest of MACE or death from another cause, loss to follow-up, or the end of the follow-up period (5 years after the start of follow-up). The time-fixed covariates are: sex, smoking status. The time-dependent covariates are: age, time since diabetes diagnosis, BMI, HbA1c, history of a cardiovascular event; kidney disease; dyslipidemia; hypertension, pancreatitis. **Figure 1** illustrates the general structure of the data. **Table 1** provides a data dictionary.

Table 2 provides examples of causal questions about the impacts of treatments on outcomes that could be addressed using the ABCi data. The questions cover treatment strategies of different complexity: 1. Point treatment strategies, 2. Sustained longitudinal treatment strategies including sustained non-treatment, 3. Treatment strategies that allow a grace period for treatment initiation, and 4. Dynamic treatment strategies. Outcomes of interest are incidence of MACE and death from other causes up to 5 years. These could be considered as competing events or as a composite event. For each treatment strategy and outcome type, interest may lie in different types of causal estimand, including average treatment effects (ATEs), conditional average treatment effects (CATEs), or individualised predictions of risk under different treatment strategies (prediction under interventions) conditional on baseline characteristics. Questions about the first set of treatment strategies (1. Point treatment strategies) can be addressed using methods that only require control for confounding due to baseline covariates. Tackling questions about treatment strategies 2, 3, and 4 requires control for time-dependent confounding.

3. Generating the ABCi data

In this section we outline how the ABCi data were generated, and provide summary statistics. The full data generating mechanism is outlined in **Supplementary Section S1** and **Supplementary Table 1**. To generate the data required assumptions about the following: covariate distributions at baseline; trajectories of time-dependent covariates and treatment use over time; how covariates affect treatment use, and how treatment use in turn affects later values of time-dependent covariates; how covariates and treatments affect the outcomes. After an individual starts an add-on therapy they always continue and there are also some individuals who never start an add-on therapy. In the real-world clinical setting patients can switch between second line treatments or add third or subsequent treatments, but we did not incorporate this into our simulated cohort. The outcomes of MACE and death from other cause were generated based on hazard models. Although not all events included in MACE mean an individual has died, follow-up ends at MACE for individuals who have MACE. The data generating mechanism includes non-linear associations, including interaction terms, which are incorporated to challenge standard implementations of analysis methods which tend to assume linear relationships.

The ABCi data includes 10,000 simulated individuals. Their characteristics at the start of follow-up are summarised in **Table 3**. **Table 4** shows the number (%) of individuals who have initiated each treatment at years 1-5. The percentage of individuals initiating treatments A,B,C,D over the 5 years of follow-up are respectively 19.3%, 14.8%, 11.0%, 4.2%. **Figure 2** shows the cumulative incidences for MACE and other deaths. There are 2699 MACE events and 466 other deaths over the 5 years, 1315 people are lost to follow-up (i.e. censored before time 5), with the remaining 5894 individuals administratively censored at 5 years of follow-up. The estimated overall cumulative incidences are 28.9% for MACE and 4.9% for other deaths.

4. Illustration of use of the ABCi data

We illustrate the use of the ABCi data to address the questions set out in **Table 1**. In the main text we focus on the point treatment strategies 1.A, 1.B, 1.C, 1.D and their impacts on the outcomes of MACE, with death from other causes as a competing event. Below we outline the causal questions, estimands, data set-up steps and analysis methods. We also outline how the ground truth for the estimates of interest can be generated.

Supplementary Section S2 provides further details, where we also provide illustrations of estimating the effects of the other treatment strategies listed in **Table 1**.

4.1 Causal question and estimands

For treatment strategies 1.A, 1.B, 1.C, 1.D, the population of interest is individuals with T2D using metformin, who have not yet started a second line treatment, whose HbA1c is above 7.5, and whose doctor had decided to initiate a second line treatment. The focus is on estimating population average treatment effects. More specifically, the causal estimands that we consider are the marginal cumulative incidences for MACE up to 4 years if all individuals starting a 2nd line treatment had started treatment A, and had they all started B, C or D. We focus on cumulative incidences in the real world in which competing events are not eliminated, the differences in such cumulative incidences have been described as ‘total effects’ (Young et al. 2020). The focus is on cumulative incidences up to 4 years as the maximum follow-up in the ABCi data is 5 years, and for the question the time zero for the analysis is after the start of follow-up for most individuals, meaning that very few have 5 years of follow-up from time zero.

4.2 Data set-up and estimation methods

Prior to the analysis step, the data need to be set up, with the individuals who will contribute to the analysis identified. The data set-up stage is a challenging aspect of a study, and as a learning and teaching resource the ABCi data provides opportunities for consideration of this aspect. For this illustrative question, the analysis data set is created by identifying individuals meeting the eligibility criteria. Different individuals in the cohort meet the eligibility criteria at different times. We identify individuals who initiate treatment A, B, C or D and retain those individuals whose HbA1c was above 7.5 at the most recent measurement time prior to initiation – this is referred to as the analysis cohort. The remaining individuals do not contribute to the analysis cohort, and this is individuals who never start a 2nd line treatment, and individuals who start treatment when their HbA1c is less than 7.5. The ‘time zero’ for the analysis is the time of starting a 2nd line treatment. The analysis will make use of treatment status (A, B, C, D), the time-fixed covariates, the values of the time-dependent covariates as measured most recently prior to initiation of the treatment, and the time of the outcome or censoring, where these times are measured relative to time zero for the analysis cohort, which is not the same as the start of follow-up in the overall cohort.

A challenge for estimation is confounding of the association between treatment use and the outcome. As treatment status does not change over time in the analysis cohort, the confounding is by individual features up to the time zero for the analysis, i.e. the covariate history prior to treatment initiation. We consider four methods for estimation of the marginal cumulative incidences of interest corresponding to the four treatment strategies: (i) a naive analysis without adjustment for confounders, using an unweighted Aalen-Johansen estimator applied separately in each treatment group; (ii) inverse probability of treatment weighted (IPTW) Aalen-Johansen estimator performed separately in each treatment group; (iii) using cause-specific Cox models including treatment and potential confounders, followed by obtaining conditional cumulative incidences and standardising (a version of the g-formula); (iv) a doubly robust targeted maximum likelihood estimation procedure for competing events outcomes described by Diaz et al. (2024). All methods rely on the assumptions of positivity, consistency, conditional exchangeability, and conditionally independent censoring. While these assumptions are known to be met in the ABCi data, the methods also require specification of certain models and incorrect specifications could lead to bias. The IPTW approach (ii) requires correct specification of the model for treatment. The outcome modelling approach (iii) requires correct specification of the cause-specific hazard models for MACE and death from other causes. Method (iv) is a doubly robust approach, which gives consistent estimates if either the treatment model and the censoring models are correctly specified, or the cause-specific hazard models are correctly specified. In our implementation of the methods, which use standard settings, none of the models used are correctly specified, for example because we did not include non-linear terms.

Further details on the estimation methods and their assumptions are given in **Supplementary section S2.1**. All analyses were performed using R. Methods (i)-(iii) were performed using userwritten code and method (iv) was implemented using the lmtp package in R (Williams et al. 2025). For the naïve approach we used standard estimates of 95% confidence intervals, for the IPTW and g-formula approaches we used percentile-bootstrap 95% confidence intervals, and for the doubly robust approach we used the 95% confidence intervals based on the efficient influence function provided by the lmtp package.

4.3 Calculating the 'ground truth'

As noted earlier, a key advantage of simulated data is that the true value of the estimands of interest can be calculated. However, in a complex longitudinal setting the true value of the estimand of interest does not typically correspond to single parameter that is one of the inputs to the data generating procedure, and a simulated-based approach is needed. This is the case for the ABCi data. The steps for estimating the true values of the cumulative incidence for MACE under treatment strategy 1.A (that everyone initiating a 2nd line treatment initiates treatment A) are as follows:

1. Generate the baseline characteristics as for the observed ABCi data.
2. Generate time-dependent covariates, outcomes, and treatment initiation times as for the observed ABCi data, up to the point of initiation of any treatment A,B, C or D. The hazards for treatment initiation are as in the observed data.
3. At the time of initiation of any treatment (A,B,C or D) set the treatment initiated to be treatment A, which is counter-to-fact for an individual simulated to initiate treatment B, C or D.
4. For all times after treatment initiation the time-dependent covariates and outcomes are generated as though all individuals initiating treatment had initiated treatment A.
5. The analysis data set for obtaining the true values of the estimands is then the subset of individuals who initiated treatment (which was fixed to be treatment A for everyone) and whose HbA1c was above 7.5 at that time. Individuals that never initiate treatment or that initiate treatment with HbA1c below 7.5 are excluded. The time zero for the analysis is the time of treatment initiation, and the cumulative incidence for MACE is estimated using the Aalen-Johansen estimator.

These steps are then repeated for treatments B, C and D, to obtain true values of the cumulative incidences for MACE and death from other causes if all treatment initiators had taken a specified treatment (strategies 1.A, 1.B, 1.C, 1.D). As the truth is obtained via a simulation procedure it is subject to Monte Carlo error. It is important, therefore, to use a large enough sample size for the start cohort such that there is little variability in the 'true' values obtained across cohorts generated with different random seeds.

4.4 Results

Figure 3 shows the estimated cumulative incidences of MACE under treatment strategies 1.A, 1.B, 1.C, and 1.D using the four analysis methods, and compared with the true values, obtained using the above procedure. In this example, the estimated cumulative incidences from the naïve analysis tend to be further from the true values than those from the other methods, which control for confounding. As we would expect, the IPTW analysis gives somewhat wider confidence intervals compared to the G-formula and doubly robust results. Estimates from the doubly robust analyses tend to be closer to the true values at most times. The computation time required for the doubly robust analyses was substantially greater than that for the other analyses.

5. Discussion

This paper has described the creation of the ABCi data set, which contains longitudinal data on treatments, covariates and time-to-event outcomes based on a case-study in type 2 diabetes. The ABCi data set is intended as a resource for use as a benchmark data set for causal inference methods and for learning and teaching of such methods. We have suggested example causal questions that could be of interest to address using the data, targeting different types of treatment strategy and different causal estimands. We illustrated the use of the ABCi data to address some of these questions using different analysis methods, focusing on population average treatment effects. The simulated data were generated to be realistic in terms of the associations. However, it is important to emphasize that the data are not intended to be used to draw conclusions about the effects antihyperglycemic medications in real-world clinical practice.

When new methods are developed they are often assessed and compared with alternatives using a simulation study (Morris et al. 2019, Heinze et al. 2023). However, the simulated data used in such studies tend to be simpler than real world data sets and are often designed, intentionally or otherwise, such that the new analysis method is favoured over others (Pawel et al. 2023, Nießl et al. 2023). Often, more complex real data are used for an illustration of methods alongside a simulation, but the real data are often not accessible to the readers. The ABCi data has a role as an example data set for method illustrations, following traditional simulation studies that are tailored to assess different aspects of method performance, for example by generating data such that certain models or assumptions hold.

Access to data is also important for learning of methods, including to assess their suitability for future use in practice. Data are also needed in teaching, such as courses focused on causal methods, and for workshops. It is therefore useful for individuals and groups to have easy access to data with some of the features they will face in practice. However, gaining access to real data is at the very least time consuming, can be expensive, and can be impossible due to data governance, particularly for students and workshop participants. Many educators or workshop organisers develop their own simulated resources. This is time-consuming and the ABCi data provides a ready-made resource. A version of the ABCi data was originally designed for a workshop held at the Lorentz Center (Leiden, Netherlands) to enable several groups to explore different analysis methods for prediction under interventions.

Strengths of the ABCi data resource include that the ground truth can be generated (as illustrated in our code for the four example questions), that it reflects the complex structure of longitudinal data, and that the data generating models have complexities such as non-linear terms. Different analysis methods require different models to be specified. These include models for treatment initiation, and models for the outcome which may be marginal or conditional on covariates at a particular time or conditional on time-updated covariates. The data have been generated such that none of the models typically required in different causal inference methods can be easily correctly specified. For example, a multinomial logistic model for treatment initiation including only main effects of covariates with untransformed versions of continuous variables would not be correctly specified, though the correct form of the model can be known from the data generating mechanism. The hazard models for the outcomes (MACE and death from other causes) depend on time-updated continuous variables in a non-linear way. The forms of marginal outcome models or outcome models conditional on covariates at a particular time point (e.g. time 0), would be difficult to specify correctly (or to derive what the correct specification would be). These features reflect reality, in which the correct forms of models needed for analysis are not known. It does mean, however, that the data are not suitable for testing performance of methods under correct specification of certain models. This arguably makes the data suited to more 'neutral' comparisons of methods. Oher simulation techniques have been developed for use in assessment of causal inference methods under correct model specification, for example which enable correct specification of marginal models (Keogh et al. 2021, Seaman and Keogh 2024, Evans and Didelez 2023)

There exist other accessible data resources that have been designed for, or can be used for, learning or assessment of causal inference methods. The R package 'causaldata' R package (Huntington-Klein and Barrett 2021) gathers together data sets used in several causal inference textbooks. In a tutorial on causal inference methods focused on continuous outcomes and point exposures Goetghebeur et al. (2020) provided a 'simulation learner' based on a study of the effects of breastfeeding interventions on child development. In 2022, the American Causal Inference Conference (ACIC) hosted a data challenge (https://acic2022.mathematica.org/, Thal and Finucane 2023) based on a freely available simulated data set mimicking a real Medicare data set, containing time-fixed exposure and covariates and subsequent outcome. This data resource has been used to study new methods (e.g. Kokandakar et al. 2023). Another data challenge was set in 2023, which featured longitudinal data on treatments, covariates and continuous outcome (https://github.com/zalandoresearch/ACIC23-competition). With the exception of the last data set, most available data sets suitable for assessment of causal inference methods do not feature longitudinal variables, and there is lack of example featuring time-to-event outcome. With the exception of the simulation learner of Goetghebeur et al. (2020), these examples also do not provide a way of generating the ground truth.

An alternative to fully simulated data is to use so-called plasmode simulation, which involves generating data by resampling from a real data set (Schrek et al 2024). Outcomes are then simulated according to specified mechanisms, which could be estimated from the data that is resampled. An advantage of plasmode simulations is that they result in realistic dependence structures as found in real data, which are more difficult to represent in a parametric simulation. A disadvantage is that only part of the data generating mechanism is known, which may be insufficient for some purposes. They are also not well suited to generating complex longitudinal dependences. Examples of plasmode simulation in the context of assessing causal inference methods include those of Franklin et al (2014) and Souli et al (2023). Souli et al. (2023) generated longitudinal data including time-varying confounding. However, only the baseline covariates and baseline exposure were resampled from the original data, with time-varying covariates and outcome being generated based on models fitted to the data. A disadvantage of plasmode simulation is that there are typically restrictions on the sharing of resampled data from, and the data was not shared in either of the above-mentioned examples.

Related to plasmode data is synthetic data based on a real data set. The Clinical Practice Research Datalink (CPRD) collects anonymised patient data from GP practices in the UK. The anonymised data are available to researchers through a formal approvals process with submission of a detailed protocol, and there is a cost. The CPRD has also developed synthetic data sets (https://www.cprd.com/synthetic-data, Wang et al. 2021) based on the real data while preserving patient privacy, which can be requested by researchers and which are designed for training purposes or testing of algorithms. The synthetic data sets are cross-sectional. Being based on real data, the underlying data generating mechanisms are not specified for these data, meaning that the true values of parameters of interest in estimation procedures are not known.

In summary, there is a lack of openly accessible data sets for assessment, learning and teaching of causal inference methods. Especially lacking is data that reflect complexities encountered in real data, including longitudinal data on treatments, covariates and outcomes, non-linearities, and competing events. This addresses some of the recommendations of Franklin and Friede (2023) regarding establishment of benchmark data sets within methods communities, use of benchmark data sets alongside simulations, and use of neutral comparison studies. We provide a single 'benchmark' data set and the code for generation. The data generating mechanism could also be used to generate repeated data sets for use in a simulation study. There are several commonly-encountered features of observational data that are not incorporated in the ABCi data resource, but which can present additional challenges for estimating causal effects and which methods are increasingly developed to address. These include missing data in time-fixed or time-dependent covariates, informative observation of time-dependent covariates (which are updated every 30 days in the ABCi data), dependent censoring, and unmeasured confounding. The data generating mechanism could be modified to accommodate these complexities. We encourage other researchers to make modifications to the data generating mechanism, and to make new versions available for others to use to expand the usefulness of the resource.

Acknowledgements

A version of the ABCi data was originally designed for a workshop held at the Lorentz Center (Leiden, Netherlands) to enable several groups to explore different analysis methods for prediction under interventions.

References

Athey S, Imbens GW, Metzger J, Munro J. Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations. 2020. https://arxiv.org/pdf/1909.02210.pdf

Díaz I, Hoffman KL, Hejazi NS. Causal survival analysis under competing risks using longitudinal modified treatment policies. Lifetime Data Anal. 2024 Jan;30(1):213–36.

Evans RJ, Didelez V. Parameterizing and Simulating from Causal Models [Internet]. arXiv; 2023 [cited 2025 Aug 15]. Available from: http://arxiv.org/abs/2109.03694

Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014 Apr;72:219–26.

Friedrich S, Friede T. On the role of benchmarking data sets and simulations in method comparison studies. Biometrical Journal 2023; 2200212. https://doi.org/10.1002/bimj.202200212

Goetghebeur E, le Cessie S, De Stavola B, Moodie EEM, Waernbaum I. Formulating causal questions and principled statistical answers. Statistics in Medicine 2020; 39: 4922-4948

Huntington-Klein N, Barrett M. causaldata: Example Data Sets for Causal Inference Textbooks [Internet]. 2021 [cited 2025 Aug 4]. p. 0.1.4. Available from: https://CRAN.R-project.org/package=causaldata

Keogh R, Seaman S, Gran J, Vansteelandt S. Simulating longitudinal data from marginal structural models using the additive hazard model. Biometrical journal Biometrische Zeitschrift. 2021 May 13;63(7):1526–41.

Kokandakar AH, Kang H, Deshpande SK. Bayesian Causal Forests & the 2022 ACIC Data Challenge: Scalability and Sensitivity [Internet]. arXiv; 2023 [cited 2025 Aug 14]. Available from: http://arxiv.org/abs/2211.02020

NICE 2025. Diabetes - type 2. Scenario: Management – adults. https://cks.nice.org.uk/topics/diabetes-type-2/management/management-adults/. Accessed 15/8/2025

Nießl C, Hoffmann S, Ullmann T, Boulesteix A-L. Explaining the optimistic performance evaluation of newly proposed methods: A cross-design validation experiment. Biometrical Journal 2023; 2200238. DOI: 10.1002/bimj.202200238

Pawel S, Kook L , Reeve K. Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. Biometrical Journal 2023; 2200091. DOI: 10.1002/bimj.202200091

Schreck N, Slynko A, Saadati M, Benner A. Statistical plasmode simulations–Potentials, challenges and recommendations. Statistics in Medicine. 2024;43(9):1804–25.

Strobl C, Leisch F. Against the “one method fits all data sets” philosophy for comparison studies in methodological research. Biometrical Journal. 2024;66(1):2200104.

Thal DRC, Finucane MM. Causal Methods Madness: Lessons Learned from the 2022 ACIC Competition to Estimate Health Policy Impacts. Observational Studies. 2023;9(3):3–27.

Williams N, Díaz I, Campbell-Brown B. lmtp: Non-Parametric Causal Effects of Feasible Interventions Based on Modified Treatment Policies. 2025. Available from: https://cran.r-project.org/web/packages/lmtp/index.html

Seaman SR, Keogh RH. Simulating Data From Marginal Structural Models for a Survival Time Outcome. Biometrical Journal. 2024;66(8):e70010.

Souli Y, Trudel X, Diop A, Brisson C, Talbot D. Longitudinal plasmode algorithms to evaluate statistical methods in realistic scenarios: an illustration applied to occupational epidemiology. BMC Medical Research Methodology. 2023 Oct 18;23(1):242.

Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2021;37(2):819–51.

Table 1: Data dictionary summarising variables included in the ABCi data.

Variable type	**Variable name**	**Description**
Identifier	id	Simulated person ID
Time and outcome variables	Tstart	Times at which time-dependent variables are updated. In days from 0, 30, 60,…, 1800 [time increments of 30 days].
	Tstop	End of time interval, which is tstart plus 30 days if the person remains event-free and uncensored at the end of the 30-day period. If the event or censoring occurs within the interval then tstop is the tme (in days) of the event or censoring.
	Mace_time	Time in days to MACE. [Same in every row.] NA if MACE does not occur.
	Death_time	Time in days to death. [Same in every row.] NA if death does not occur.
	Cens_time	Time in days to censoring. [Same in every row.] NA if MACE or death occurs.
	Mace_status	Time-updated MACE status. 0 in rows where individual has not yet had MACE at tstop. 1 in final row for individual where Mace_time equal to tstop.
	Death_status	Time-updated death status. 0 in rows where individual has not yet died at tstop. 1 in final row for individuals where death_time equal to tstop.
	Cens_status	Time-updated censoring status. 0 in rows where individual has not yet been censored at tstop. 1 in final row for individuals where cens_time equal to tstop.
	Event_time	Earliest of time to MACE, death or censoring. [Same in every row.]
	Event_type	Status at event_time. 0: censored 1: MACE 2: death 3: administratively censored [Same in every row.]
Time fixed covariates	sex	0: male 1: female
	Age	Age at baseline in years
	smok	Smoking status at baseline. 1: never smoker 2: former smoker 3: current smoker
Time-dependent covariates	Age	Age in years

(updated every 30 days)		
	Diabdur	Time since diabetes diagnosis, in years.
	Bmi	Body mass index
	Hb	HbA1c (%)
	Hyp	Indicator of hypertension (0/1)
	Dys	Indicator of dyslipidemia (0/1)
	Cvd	Indicator of history of cardiovascular disease (0/1)
	Kidney	Indicator of kidney disease (0/1)
	panc	Indicator of pancreatitis (0/1)
Treatment variables	Treat_time_A	Time of starting treatment A. NA if individual never starts treatment A. [Same in every row.]
	Treat_time_B	As above, for treatment B
	Treat_time_C	As above, for treatment C
	Treat_time_D	As above, for treatment D
	Treat_status_A	Time-updated indicator of using treatment A. 0 in rows where individual has not started treatment A before tstart. 1 in rows where tstart is >= treat_time_A.
	Treat_status_B	As above, for treatment B
	Treat_status_C	As above, for treatment C
	Treat_status_D	As above, for treatment D

Table 1. Example causal questions to be addressed using the ABCi data. Specification of treatment strategies, the population of interest for applying those strategies, and the outcomes of interest.

Treatment strategies	**Population of interest**	**Outcomes**
1. Comparative effectiveness point treatment strategies		
1.A. Start 2nd line treatment A immediately and continue to the time horizon of interest. 1.B, 1.C, 1.D. As in 1.A, but using treatment B, C, D.	• People with type 2 diabetes on Metformin. • Not yet started a 2nd line treatment. • HbA1c >7.5. • Doctor has decided to prescribe a 2nd line treatment.	• Incidence of MACE up to t years ($t \leq 5$) • Incidence of death from other causes up to t years ($t \leq 5$) • Incidence of the composite of MACE or death from other causes up to t years (t≤5)
2. Sustained treatment strategies		
2.A. Start 2nd line treatment A immediately and continue to the time horizon of interest. 2.B, 2.C, 2.D. As in 2.A, but using treatment B, C, D. S0. Do not start any 2nd line treatment up to the time horizon of interest.	• People with type 2 diabetes on Metformin. • Not yet started a 2nd line treatment. • HbA1c >7.5.	
3. Treatment strategies with a grace period		
3.A. Start 2nd line treatment A, allowing a grace period of 90 days, and continue to the time horizon of interest. 3.B, 3.C, 3.D. As in 3.A, but using treatment B, C, D.	• People with type 2 diabetes on Metformin. • Not yet started a 2nd line treatment. • HbA1c >7.5.	
4. Dynamic treatment strategies		
4.A. Start 2nd line treatment A when HbA1c goes above 7.5, and continue to the time horizon of interest.	• People with type 2 diabetes on Metformin. • Not yet started a 2nd line treatment.	

Table 3: Summary of the baseline characteristics of the ABCi data

Variable	Summary
Sex, n(%)	
Male	5538 (55.4)
Female	4462 (44.6)
Age (years), mean (SD)	61.87 (11.09)
Smoking status, n(%)	
Never smoker	3884 (38.8)
Former smoker	4005 (40.1)
Current smoker	2111 (21.1)
Diabetes duration (years), mean (SD)	0.45 (0.43)
BMI, mean (SD)	29.98 (6.20)
HbA1c, mean(SD)	8.18 (1.48)
Hypertension, n(%)	6221 (62.2)
Dyslipidemia, n(%)	4167 (41.7)
Cardiovascular disease, n(%)	1756 (17.6)
Kidney disease, n(%)	596 (6.0)
Pancreatitis, n(%)	330 (3.3)

Table 4: Number (%) of individuals in the ABCi data who have started treatment A, B, C, D (or any of these) after 1,2,3,4,5 years, among those who remain under observation at that time.

Treatment	1 year	2 years	3 years	4 years	5 years
Any	2275 (26.8%)	3016 (39.8%)	3292 (48.1%)	3339 (53.8%)	3291 (59.2%)
A	874 (10.3%)	1155 (15.2%)	1234 (18.0%)	1240 (16.2%)	1206 (21.7%)
B	672 (7.9%)	886 (11.7%)	995 (14.5%)	1008 (16.2%)	990 (17.8%)
C	544 (6.4%)	739 (9.8%)	804 (11.7%)	823 (13.3%)	827 (14.9%)
D	185 (2.2%)	236 (3.1%)	259 (3.8%)	268 (4.3%)	268 (4.8%)

Figure 1: Illustration of the overall structure of the SimCI data.

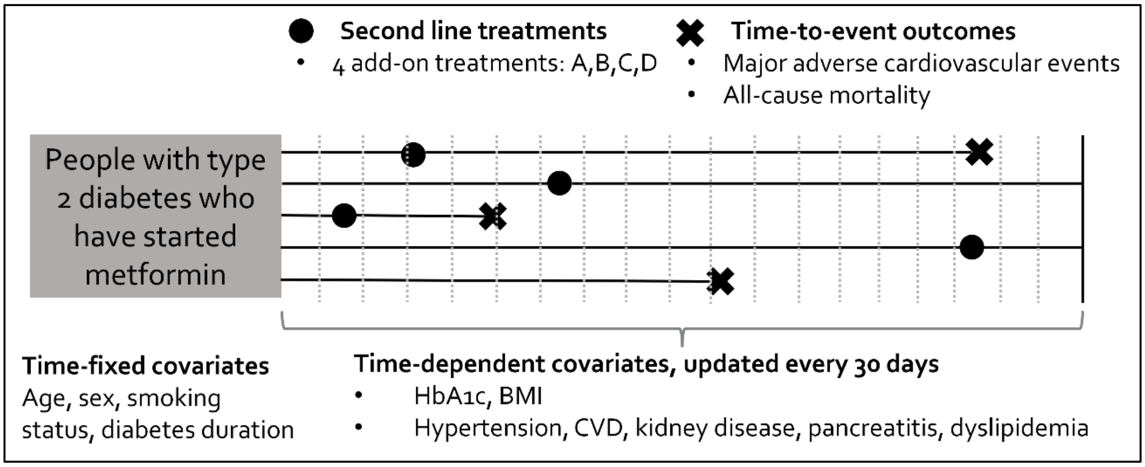

Second line treatments
4 add-on treatments: A,B,C,D
Time-to-event outcomes
Major adverse cardiovascular events
All-cause mortality
People with type 2 diabetes who have started metformin
Time-fixed covariates
Age, sex, smoking status, diabetes duration
Time-dependent covariates, updated every 30 days
HbA1c, BMI
Hypertension, CVD, kidney disease, pancreatitis, dyslipidemia

Figure 2: Overall cumulative incidences of MACE and other deaths in the ABCi data.

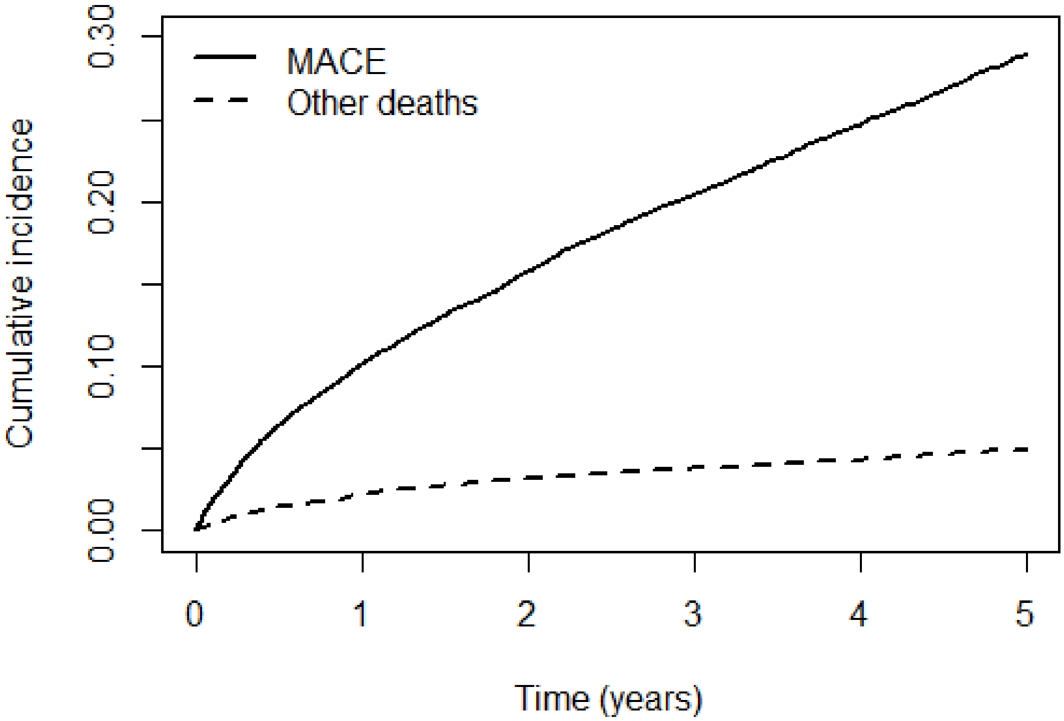
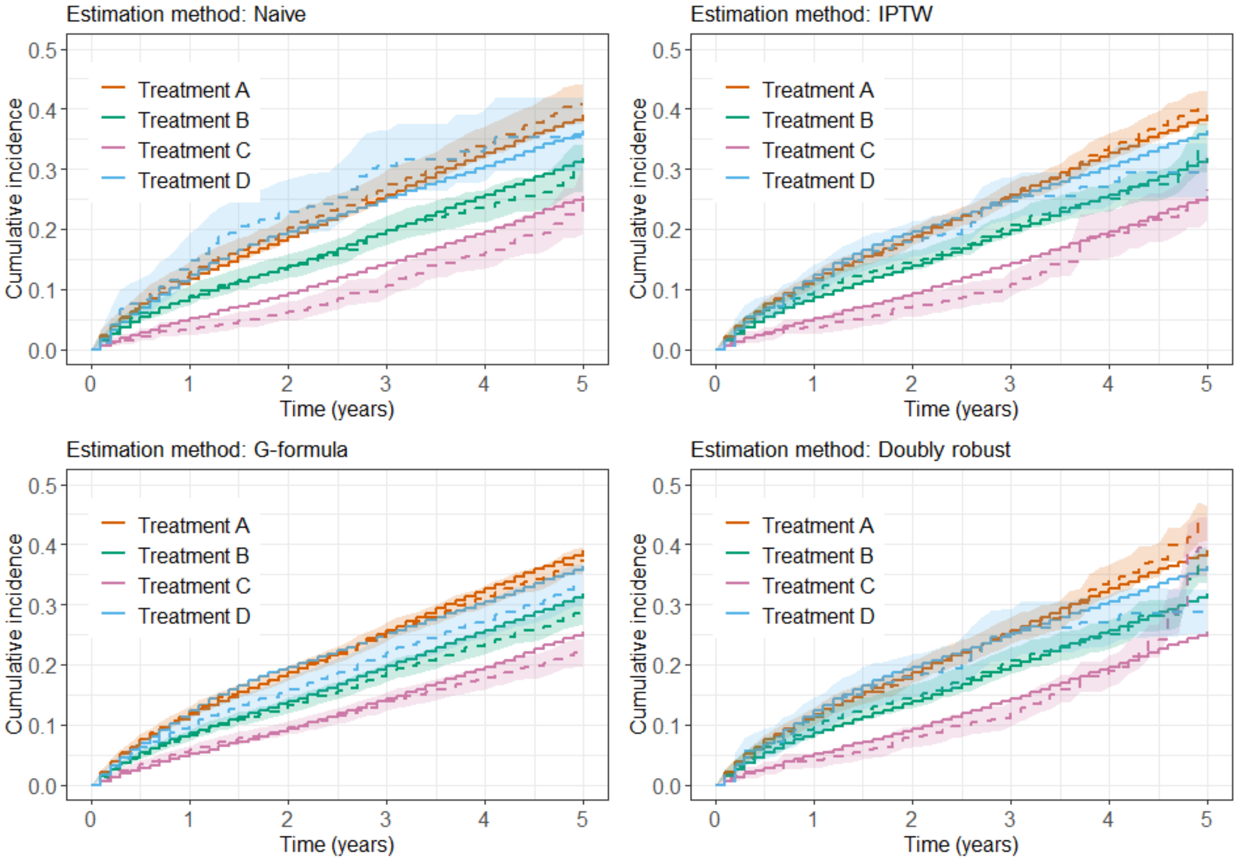


Figure 3: Estimated cumulative incidences of MACE under treatment strategies 1.A, 1.B, 1.C, and 1.D. Solid lines show the true cumulative incidences, and dashed lines show the estimates obtained using the (i) naïve analysis, (ii) IPTW analysis, (iii) g-formula analysis, (iv) doubly robust analysis. The shaded areas show the 95% confidence intervals.


Estimation method: Naive
Estimation method: IPTW
Estimation method: G-formula
Estimation method: Doubly robust
Treatment A
Treatment B
Treatment C
Treatment D
Cumulative incidence
Time (years)