

BMJ Open Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review

Stan R W Wijn ¹, Maroeska M Rovers ², Gerjon Hannink¹

To cite: Wijn SRW, Rovers MM, Hannink G. Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review. *BMJ Open* 2022;**12**:e058977. doi:10.1136/bmjopen-2021-058977

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-058977>).

Received 10 November 2021
Accepted 11 February 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Radboud University Medical Center, Radboud Institute for Health Sciences, Department of Operating Rooms, Radboudumc, Nijmegen, The Netherlands

²Radboud University Medical Center, Radboud Institute for Health Sciences, Department of Operating Rooms and Health Evidence, Radboudumc, Nijmegen, The Netherlands

Correspondence to

Stan R W Wijn;
stan.wijn@radboudumc.nl

ABSTRACT

Objectives To adjust for confounding in observational data, researchers use propensity score matching (PSM), but more advanced methods might be required when dealing with longitudinal data and time-varying treatments as PSM might not include possible changes that occurred over time. This study aims to explore which confounding adjustment methods have been used in longitudinal observational data to estimate a treatment effect and identify potential inappropriate use of PSM.

Design Mapping review.

Data sources We searched PubMed, from inception up to January 2021, for studies in which a treatment was evaluated using longitudinal observational data.

Eligibility criteria Methodological, non-medical and cost-effectiveness papers were excluded, as were non-English studies and studies that did not study a treatment effect.

Data extraction and synthesis Studies were categorised based on time of treatment: at baseline (interventions performed at start of follow-up) or time-varying (interventions received asynchronously during follow-up) and sorted based on publication year, time of treatment and confounding adjustment method. Cumulative time series plots were used to investigate the use of different methods over time. No risk-of-bias assessment was performed as it was not applicable.

Results In total, 764 studies were included that met the eligibility criteria. PSM (165/201, 82%) and inverse probability weighting (IPW; 154/502, 31%) were most common for studies with a treatment at baseline (n=201) and time-varying treatment (n=502), respectively. Of the 502 studies with a time-varying treatment, 123 (25%) used PSM with baseline covariates, which might be inappropriate. In the past 5 years, the proportion of studies with a time-varying treatment that used PSM over IPW increased.

Conclusions PSM is the most frequently used method to correct for confounding in longitudinal observational data. In studies with a time-varying treatment, PSM was potentially inappropriately used in 25% of studies. Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding are available, but were only used in 45% of the studies with a time-varying treatment.

Strengths and limitations of this study

- We systematically mapped the literature from inception up to January 2021 for the most commonly used methods to correct for confounding in longitudinal observational data.
- This study was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.
- No risk-of-bias assessment was performed because the scope of this mapping review targets the statistical methods that have been used in the included studies, so a risk of bias assessment was not applicable.
- For some studies we were not able to identify if patients were treated at baseline or during follow-up (fortunately, this issue was only apparent in 8% of the included studies).

INTRODUCTION

The increasing availability of real-world data derived from electronic health records, registries, wearables and surveys can be a valuable source of data to evaluate the effectiveness of a treatment.¹ Deriving inference directly from real-world data can be challenging as it is prone to confounding. To adjust for confounding, researchers use methods such as propensity score matching (PSM) to create two comparable groups in which both the treated- and untreated patients have similar observable characteristics (like age, pain scores, weight, etc.) similar to a randomised trial.²

Although these methods can be sufficient when a patient is treated at the start of a study (baseline), more advanced methods might be required when dealing with longitudinal data and time-varying or repeated treatments. Adjustment at baseline in the presence of longitudinal data and time-varying treatment might not include possible changes that occurred over time. These can include

Box 1 Empirical example using data from the Osteoarthritis Initiative

To investigate the influence of the different confounding adjustment methods on the outcome, two previously published empirical examples with a time-varying treatment were selected: (1) the effect of meniscectomy (surgical removal of the meniscus) and (2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.^{20 21} Data from the Osteoarthritis Initiative (OAI) was used for both examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access (<https://data-archivenimh.nih.gov/oai/>). A large set of variables was extracted from the OAI, measured at baseline and annual follow-up visits. These include general patient characteristics, clinical variables, quality of life measurements, functional scores and time-varying treatments.

In total, we compared nine commonly used adjustment for both empirical examples: four methods that corrected using baseline covariates, four time-dependent methods and no matching. We found in the first example (meniscectomy) that adjustment using baseline covariates resulted in larger estimates of the treatment effect compared with time-dependent methods, while results were consistent in the second example (intra-articular corticosteroid injection; figure 1). These results show that the selected adjustment method can influence the detected treatment effect when dealing with potential time-varying confounding. See online supplemental file S2 for more details.

changes in treatment regimens or disease progression, but can also comprise weight changes, pain scores or changes in behaviour (eg, stopped smoking). These changes can alter the balance between treated and untreated patients and can result in different estimates of the treatment effect (see box 1).^{3 4}

Methods like time-dependent PSM and the g-methods (inverse probability weighting (IPW), parametric

g-formula or g-estimation) can incorporate time-varying covariates and time-varying treatments and can take feedback between the treatment and outcome over time into account.^{2 5–8} It is however unclear if these methods are regularly used in practice when dealing with longitudinal observational data with a time-varying treatment. Therefore, this mapping review aimed to identify and describe which methods have been used to adjust for confounding bias in longitudinal observational data and identify potential inappropriate use of baseline adjustment methods (like PSM).

METHODS

A mapping literature review was performed to determine which confounding adjustment methods were used in longitudinal observational data to estimate a treatment effect. Mapping reviews are designed to map out and categorise existing literature and explore trends and identify gaps by study design and other key features.⁹ This study was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.¹⁰

Patient and public involvement

Patients and/or public were not involved.

Search strategy

We searched in PubMed from inception up to January 2021 for papers in which a treatment was evaluated using longitudinal observational data. Search terms used were time varying, longitudinal observational data, and commonly used adjustment methods and terms (eg, matching, g-methods). The search strategy can be found in online supplemental file S1. Methodological, non-medical and cost-effectiveness papers were excluded as well as non-English studies or studies that did not study a treatment effect. Studies that used no adjustment method or used the adjustment method solely as sensitivity analysis were also excluded.

All papers were screened based on title and abstract and papers that met the inclusion criteria were screened full-text. The title, author(s), journal, research theme, publication date, confounding adjustment method and time of treatment (at baseline or time-varying) were extracted from all papers that met the inclusion criteria. A treatment at baseline was defined as an intervention performed at the start of follow-up for all included patients (eg, all treated patients received surgery at the start of follow-up). Time-varying treatment was defined as a treatment that was received asynchronously during follow-up (eg, patients received surgery at different moments during follow-up) or when dealing with a repeated treatment of which the timing was not identical for all treated patients (eg, personalised medication intake over time). If the time of treatment was not defined, studies were categorised as unclear.

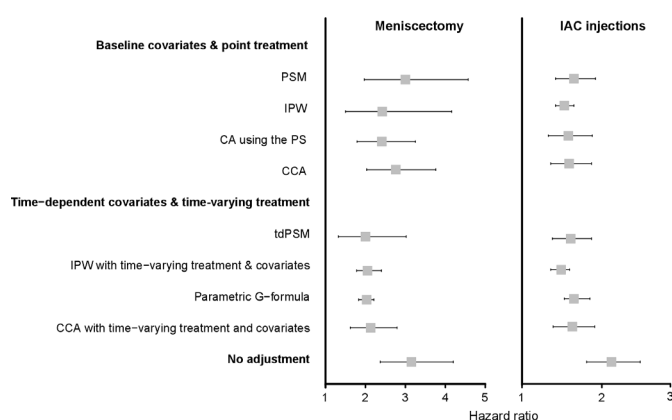


Figure 1 Forest plot displaying the results of the two empirical examples (left: meniscectomy, right: intra-articular corticosteroid (IAC)). Four methods were compared using baseline covariates, four methods using time-dependent covariates and time-varying treatment and one without correction. CCA, conventional covariate adjustment; IPW, inverse probability weighting; PSM, propensity score matching; tdPSM, time-dependent propensity score matching.

Study selection and data extraction were performed by one reviewer (SRWW). Any issues during study selection, data extraction or analysis were discussed and resolved by all authors. No risk of bias assessment was performed because the scope of this paper targets the statistical methods that have been used in these papers, and therefore a risk of bias assessment was not applicable.

Analysis

Study selection was performed in Rayyan.¹¹ Study characteristics (author, publication year, journal), time of treatment (at baseline, time-varying or unclear) and confounding adjustment method were extracted and analysed in R (V.4.1.0, The R Foundation for Statistical Computing, Vienna, Austria). Studies were sorted based

on publication year, time of treatment and confounding adjustment method and described using descriptive statistics. If a study used multiple adjustment methods or a combination of methods, we included all methods, that is, more methods than papers could be identified. Cumulative time series plots were used to investigate the use of different methods over time for treatments at baseline and time-varying treatments using the Plotly package.¹²

RESULTS

Our search identified 2140 articles of which eventually 764 met the eligibility criteria after title and abstract review, and subsequent full-text review (see also figure 2).

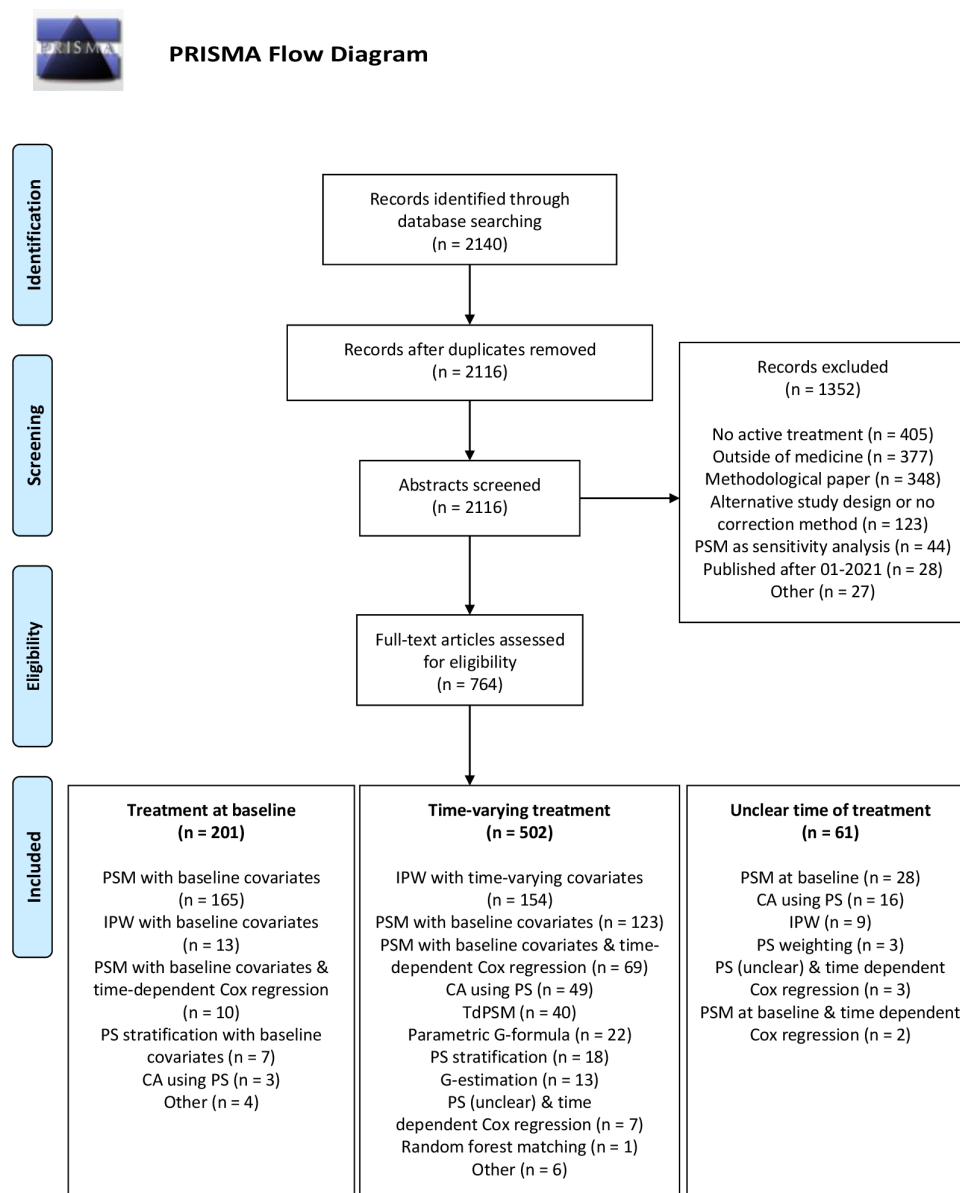


Figure 2 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram of the flow of papers in the mapping review. In total, 764 studies were included and categorised according to the time of treatment. CA, covariate adjustment; IPW, inverse probability weighting; PS, propensity score; PSM, propensity score matching; TdPSM, time-dependent propensity score matching.

The main reasons for exclusion were the lack of intervention/treatment ($n=405$), a scope outside of medicine ($n=376$), a methodological paper ($n=348$) or the study did not use longitudinal observational data or did not correct for confounding ($n=123$). Of all included papers, 201 (26%) had a treatment at baseline, 502 (66%) had a time-varying treatment and 61 (8%) papers had no clearly defined time of treatment. Of the papers with a treatment at baseline, the majority used PSM with baseline covariates ($n=165$, 82%) as a method to correct for confounding. Studies that had a time-varying treatment most often used IPW (154 papers, 30%), PSM with baseline covariates was used in 123 papers (25%), PSM with baseline covariates combined with time-dependent Cox regression in 69 papers (14%), covariate adjustment using the propensity score in 49 papers (10%), time-dependent PSM in 40 papers (8%), parametric G-formula in 22 papers (4%), propensity score stratification in 18 papers (2%) and G-estimation in 13 papers (3%). Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding (IPW, parametric g-formula or g-estimation) were used in 45% of the papers with a time-varying treatment. In the last 5 years, the proportion of studies with a time-varying treatment that used PSM with baseline covariates over IPW increased (199 vs 158 in 2020, for PSM with baseline covariates and IPW, respectively) (figure 3). For papers of which the time of treatment was unclear, PSM at baseline was most frequently used in 28 papers (46%). We added an overview of the most commonly used methods found in our search and when they should be used (figure 4).

DISCUSSION

Although advanced methods are available to correct for confounding in longitudinal observational data, we showed that these methods are not always used in studies that have a time-varying treatment. Instead, 25% of the studies that had a time-varying treatment used PSM with baseline covariates to correct for confounding which can potentially result in a biased treatment effect.⁴

Our findings confirm the results by Clare *et al* whom provided a summary of new methods that have been used in literature to deal with time-varying confounding. They concluded that IPW was the most commonly used, more robust methods (like g-estimation) were underused.¹³ Our results are also in agreement with the findings by Austin and Stuart whom reported a rapidly increasing use of IPW in the literature in the last decade.¹⁴ Nonetheless, we detected a similarly rapid growth in the use of PSM in studies with a time-varying treatment, which can potentially result in biased results as PSM does not correct for time-varying confounding. Although time-dependent methods like tdPSM, parametric g-formula and IPW are extensively described in the literature,^{5 8 15} adjusting at baseline in observational data is still common in literature and was used in 25% of the papers with a time-varying treatment we included in our mapping review.¹⁶ The

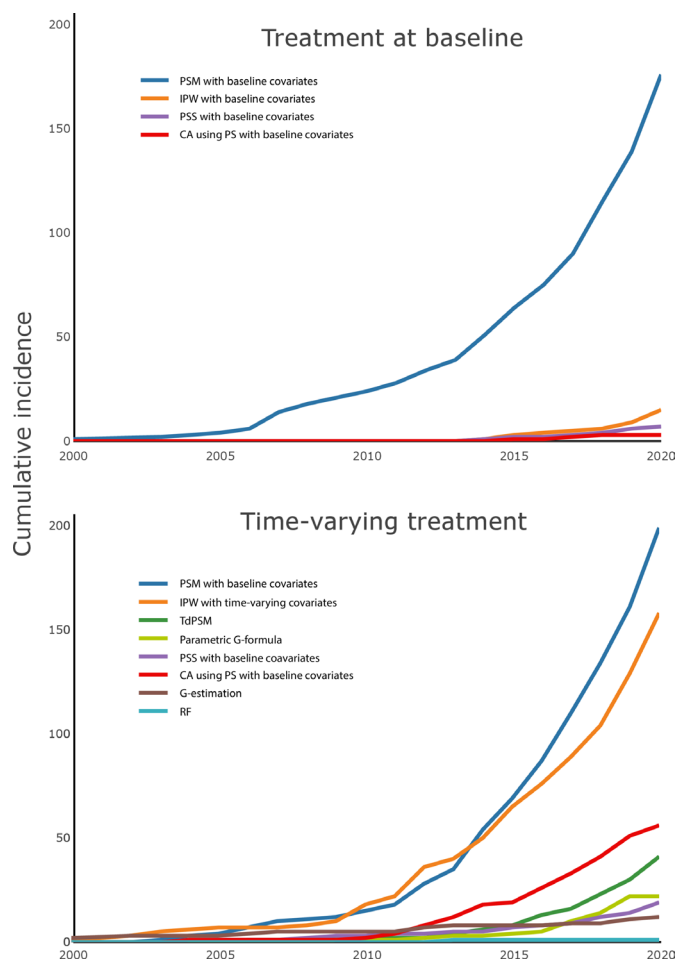


Figure 3 Cumulative incidence of the different confounding adjustment methods that are used in practice. Some studies used multiple methods. CA, covariate adjustment; IPW, inverse probability weighting; PS, propensity score; PSS, propensity score stratification; PSM, propensity score matching; RF, random forest matching; TdPSM, time-dependent propensity score matching.

proportion of studies with a time-varying treatment that used PSM over IPW even increased in the last 5 years.

Some potential limitations should also be discussed. First, the main limitation of a mapping review is the broad descriptive level at which studies are analysed and described. However, it does provide a general overview of the published literature and suggests that methods to deal with confounding in studies with a time-varying treatment are underused. Furthermore, no risk of bias assessment of the included studies was performed and study selection and data extraction were performed by one reviewer. Using a second reviewer throughout the entire study screening process could increase the number of relevant studies identified for use in a systematic review.¹⁷ However, as we targeted the overall trends in data analysis of studies with longitudinal observational data, this would likely not affect our conclusions much. Second, although it is common to search multiple databases in a systematic review, our mapping review was limited to PubMed. We found over 2000 papers in PubMed which was ample



Common methods to correct for confounding

Multiple methods are used to correct for confounding. Here we list the most common types and when they should be used.

- 
Covariate adjustment using propensity score
 - The outcome variable is regressed on an indicator variable denoting treatment status and the estimated propensity score. (included in the analysis of study)
 - Not recommended for eliminating baseline differences as it does not allow balancing of covariates across treated and control groups
- 
Propensity score stratification
 - Stratifying patients into mutually exclusive subsets based on their estimated propensity score. (separates design from analysis of study)
 - Patients within strata have similar (baseline) values of the propensity score.
 - Not recommended for eliminating baseline differences.
- 
Propensity score matching
 - Creating matched sets of treated and untreated patients who share a similar value of the propensity score. (separates design from analysis of study)
 - PSM is recommended over stratification or covariate adjustment as it eliminates greater proportion of systemic differences in baseline characteristics between treated and untreated. Not recommended for time-varying treatment or time-varying confounding.
- 
Inverse probability weighting
 - Generates a pseudo-population in which exposures are independent of confounders, enabling estimation of marginal structural model parameters. (separates design from analysis of study)
 - Suitable for baseline imbalances and time-varying confounding. Not recommended when propensities are small (close to 0) as weights can be unstable.
- 
Parametric G-formula
 - Models the joint density of the observed data to generate potential outcomes under different hypothetical treatment strategies (included in the analysis of the study).
 - Suitable for longitudinal data with a time-varying treatments and can adjust for time-varying confounders that are affected by prior exposures.
- 
G-estimation
 - Exploits the conditional independence between the exposure and potential outcomes to estimate structural nested model parameters (included in the analysis of the study).
 - Suitable to estimate the joint effect of a sequence of treatments, when dealing with continuous exposures or when standard assumptions fail.

Sources:

• Naimi AJ, Cole SR, Kennedy EH. An introduction to g-methods. *Int J Epidemiol*. 2017 Apr 1;46(2):756-762.
 • Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011 May;46(3):399-424.
 • Hernan MA, Robins JM (2020). *Causal Inference: What If* (2020). Boca Raton: Chapman & Hall/CRC.

Figure 4 Common methods to correct for confounding and when they should be used.

for the aim of this study and for a mapping review. It is unlikely that additional searches could alter our conclusions. Third, for some studies we were not able to identify if patients were treated at baseline or during follow-up. Fortunately, this only occurred in 8% of the papers we included.

Implications

From previously published studies we can conclude that time-dependent methods can be important to avoid biased estimates of the treatment effect when adjusting for confounding in longitudinal observational data with potential time-varying confounding.^{4 18} Therefore, we suggest using one of the g-methods (IPW, parametric g-formula, g-estimation) with time-varying covariates and time-varying treatment if the data is available.¹⁸ Yet, these methods are not the panacea for unconfounded analyses in longitudinal observational data. They still rely on relevant confounder selection (based on prior knowledge, possibly supported by a directed acyclic graph), require careful examination of weights and adequate covariate balance.¹⁴ Although there are clear benefits and limitations to each g-method, it is often unclear what the most appropriate method is to correct for confounding.¹⁵ From the g-methods, IPW has three main advantages over the other methods: (1) it is a commonly used method,

(2) it is relatively simple to understand and explain, and (3) it is easy to perform in standard statistical software (like R or STATA). Parametric g-formula is ideal for joint interventions or dynamic interventions but requires more computational power and additional programming.¹⁸ G-estimation is particularly useful for studying the interaction between treatment and time-varying confounders (treatment-confounder feedback), but it can be challenging to implement g-estimation in longitudinal data. G-estimation can also be complex as there are not many practical guidelines or statistical packages that support this method for longitudinal data with a time-varying treatment. The developers of *gestools* R-package (General Purpose G estimation in R) are currently drafting a comprehensive introduction including an explanation of the structural nested mean model types, the g-estimation algorithm, instructions to set up the users' dataset, and a tutorial to perform g-estimation.¹⁹

When dealing with real-world data, g-methods are recommended to evaluate the effectiveness of a treatment to preclude confounding. However, a proper assessment of the required confounding adjustment methods prior to data analysis is appropriate. As we have seen in [box 1](#), different confounding adjustment methods can potentially influence the conclusions of a study. It depends on many (unknown) case-specific aspects and thus it can be challenging to predict how different methods can affect the conclusion of a study. A direct comparison of different methods to correct for confounding is not recommended as this could stimulate selective reporting of (positive) study results. Every analysis of longitudinal observational data should start by selecting the method best suited for the data at hand. [Figure 4](#) provides an overview of the most commonly used methods and can assist researchers to select the most appropriate method available.

CONCLUSION

PSM using baseline covariates is the most used method to correct for confounding in longitudinal observational data, even in the presence of a time-varying treatment. Of the 502 identified studies with a time-varying treatment, 123 (25%) used PSM with baseline covariates, which might be inappropriate. Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding (IPW, parametric g-formula or g-estimation) are available, but were only used in 45% of the papers with a time-varying treatment and this can potentially result in biased estimates of the treatment effect.

Twitter Stan R W Wijn @Stan_Wijn and Maroeska M Rovers @MaroeskaRovers

Contributors SRWW: conceptualisation, methodology, software, validation, formal analysis, writing - original draft, visualisation. MMR: conceptualisation, writing - review and editing, supervision, project administration, funding acquisition. GH: conceptualisation, methodology, validation, writing - review and editing, supervision, project administration, funding acquisition, guarantor.

Funding This work was supported by the Junior Research project (2018) grant provided by the Radboud Institute for Health Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The search strategy is available in the supplemental file and all data extraction documents are available on request to the corresponding author.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Stan R W Wijn <http://orcid.org/0000-0003-3782-6677>

Maroeska M Rovers <http://orcid.org/0000-0002-3095-170X>

REFERENCES

- Blonde L, Khunti K, Harris SB, et al. Interpretation and impact of real-world clinical data for the practicing clinician. *Adv Ther* 2018;35:1763–74.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46:399–424.
- Pazzagli L, Linder M, Zhang M, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: an overview. *Pharmacoepidemiol Drug Saf* 2018;27:148–60.
- Zhang Z, Li X, Wu X, et al. Propensity score analysis for time-dependent exposure. *Ann Transl Med* 2020;8:246.
- Lu B. Propensity score matching with time-dependent covariates. *Biometrics* 2005;61:721–8.
- Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res* 2017;26:1654–70.
- Morgan SL, Winship C. *Counterfactuals and causal inference. Counterfactuals causal inference methods Princ. soc. Res.* Cambridge: Cambridge University Press, 2014: 1–499p.
- Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
- Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info Libr J* 2009;26:91–108.
- Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018;169:467–73.
- Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan-a web and mobile APP for systematic reviews. *Syst Rev* 2016;5:210.
- Sievert C. *Interactive Web-Based Data Visualization with R, plotly, and shiny [Internet]*. Chapman and Hall/CRC, 2020.
- Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding—a systematic review of the literature. *Int J Epidemiol* 2019;48:254–65.
- Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015;34:3661–79.
- Naimi AI, Cole SR, Kennedy EH. An introduction to G methods. *Int J Epidemiol* 2017;46:756–62.
- Kupzyk KA, Beal SJ. Advanced issues in propensity scores. *J Early Adolesc* 2017;37:59–84.
- Stoll CRT, Izadi S, Fowler S, et al. The value of a second reviewer for study selection in systematic reviews. *Res Synth Methods* 2019;10:539–45.
- Mansournia MA, Etminan M, Danaei G, et al. Handling time varying confounding in observational research. *BMJ* 2017;359:j4587.
- Dukes O, Vansteelandt S. A note on G-Estimation of causal risk ratios. *Am J Epidemiol* 2018;187:1079–84.
- Wijn SRW, Rovers MM, van Tienen TG, et al. Intra-Articular corticosteroid injections increase the risk of requiring knee arthroplasty. *Bone Joint J* 2020;102-B:586–92.
- Rongen JJ, Rovers MM, van Tienen TG, et al. Increased risk for knee replacement surgery after arthroscopic surgery for degenerative meniscal tears: a multi-center longitudinal observational study using data from the osteoarthritis initiative. *Osteoarthritis Cartilage* 2017;25:23–9.

Supplement S1: Search strategy

Initial search	(time var*[tiab] OR time dependent*[tiab] AND iptw[tiab] OR inverse probability[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab])	28373
Included propensity mesh to reduce the number of papers	("Propensity Score"[Mesh] OR propensity score*[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR iptw[tiab] OR inverse probability[tiab])	1566
Added risk-set matching	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR iptw[tiab] OR inverse probability[tiab])	1570
Added g-methods to the search and studies published before 2021	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) OR ("g-methods"[tiab] OR "g-formula"[tiab] OR "g-estimation"[tiab] OR "parametric g-formula"[tiab]) OR (iptw[tiab] OR inverse probability[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR longitudinal*[tiab]) Filters: from 1992 - 2020	2081
Added "Marginal structural Cox model" to the search	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) OR ("g-methods"[tiab] OR "g-formula"[tiab] OR "g-estimation"[tiab] OR "parametric g-formula"[tiab]) OR (iptw[tiab] OR inverse probability[tiab] OR Marginal structural Cox model[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR longitudinal*[tiab]) Filters: from 1992 - 2020	2087
Total combined		2140

Supplement S2: Empirical example details from Box 1

Empirical examples

Data from the Osteoarthritis Initiative (OAI) was used for two empirical examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access at <https://data-archive.nimh.nih.gov/oai/>. We extracted a large set of variables from the OAI that were measured at baseline and annual follow-up visits (12 to 108 months), these include general patients characteristics (age, gender, history of knee symptoms, physical activity, weight, care access), clinical variables (knee symptoms, radiographic signs of OA, hand OA), quality of life measurements (12-Item Short Form Survey (SF-12)), functional scores (Knee injury and Osteoarthritis Outcome Score (KOOS), Western Ontario and McMasters Osteoarthritis index (WOMAC)) and time-varying treatments (meniscectomy, knee replacement surgery, corticosteroid injections). Missing values were imputed through single imputation using predictive mean matching for continuous variables and logistic regression for categorical variables.

To investigate the impact of the different confounding adjustment methods on the outcome, two empirical examples with a time-varying treatment were selected that we previously published using data from the OAI: 1) the effect of meniscectomy (surgical removal of the meniscus) on the risk to receive knee replacement surgery and 2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.[19,20]

Statistical methods

In total, we compared nine methods that were the most commonly used adjustment methods found in the mapping review for both empirical examples: four methods that matched using baseline covariates, four time-dependent methods, and no matching. Confounding factors included in all eight correction methods were: patient characteristics (age, gender, BMI, physical activity, health care access, treatment centre, education, family history with OA, occupation), clinical variables (knee

26 medication use, hand OA at baseline, knee symptoms at baseline, radiographic confirmed OA),
27 quality of life scores (SF-12 subscales), and functional scores (KOOS and WOMAC). After adjustment,
28 Cox proportional hazard models were applied to estimate the treatment effect and confidence
29 intervals.

30 The baseline methods consisted of PSM, IPW with a point treatment (yes/no), covariate adjustment
31 using the propensity score, and conventional covariate adjustment (CCA) using baseline covariates
32 and a point treatment. For PSM, the propensity score was calculated for every patient (the
33 probability of a patient being assigned to the treatment given a set of observed covariates) and
34 subsequently treated and control patients were matched using a 1:1 matching ratio without
35 replacement, a caliper of 0.20 and a nearest neighbour matching algorithm, as nearest neighbour is
36 commonly used and results in less biased estimates compared to the other matching algorithms.[21]
37 Covariate balance was assessed by calculating the standardized mean difference (SMD) and by
38 plotting the balance between patients and controls. Balance smaller or equal to 0.10 SMD were
39 assumed to have appropriate balance.[2] IPW was performed to build a marginal structural model
40 able to balance the covariates at baseline (marginal structural model with point treatment; patients
41 were either labelled as treated or untreated). For IPW we used unbalanced weights and the weights
42 were visually inspected. Similar to PSM, a 0.10 SMD was assumed to have an appropriate balance.
43 Confidence intervals were estimated using 1000 bootstraps. Covariate adjustment using the
44 propensity score was performed by calculating the propensity score using logistic regression and
45 subsequently the propensity score was added to the Cox regression. Conventional covariate
46 adjustment was performed by including the same set of covariates in the Cox regression without any
47 prior adjustment.

48 The time-dependent methods consisted of time-dependent propensity score matching (tdPSM), IPW
49 with time-varying treatment, parametric g-formula, and CCA with time-varying treatment and
50 covariates.[5,15,17] Time-dependent propensity score matching was performed by sequentially

51 matching treated patients with all available controls at time of treatment using a 1:1 nearest
52 neighbour matching algorithm without replacement using a caliper of 0.2. After matching a patient
53 to a control, both were removed from the dataset to avoid further matches. Similar to the baseline
54 methods, IPW was used to create a marginal structural model but with time-varying treatment and
55 time-varying covariates. Likewise, we used unbalanced weights and the weights were visually
56 inspected and balance was assessed. Confidence intervals were estimated using 1000 bootstraps.

57 Robins' g-formula (also known as parametric g-formula or parametric g-computation) is an
58 alternative method to recover effects of time-varying treatment under untestable assumptions, given
59 that sufficient covariates are measured to control for confounding by unmeasured risk factors.[22]
60 The causal effect is measured by comparing the treatment effect between an always exposed- and a
61 never exposed scenario. Conventional covariate adjustment with time-varying covariates and
62 treatment was performed by including these variables in the Cox regression.

63 Finally, we performed one crude analysis by only including the time-varying treatment in the Cox
64 regression. All analyses and simulations were performed using R (version 4.0.2, The R Foundation for
65 Statistical Computing, Vienna, Austria) using packages 'mice', 'MatchIt', 'WeightIt', 'gfoRmula',
66 'plotly', 'coxphw', 'boot', and 'survival'. [12,22–29]
67

68 Results

69 In total, nine methods were compared for both empirical examples: four methods that adjust using
70 baseline covariates (PSM, IPW using point treatment, CA using the propensity score, CCA), four time-
71 dependent methods (tdPSM, IPW using time-varying treatment, parametric g-formula, CCA) and one
72 without adjustments. (see figure in Box 1)

73 In the meniscectomy example, patients who underwent meniscectomy had an HR of 3.0 (95% CI:
74 1.97– 4.57), 2.42 (95% CI: 1.50 – 4.16), 2.41 (95% CI: 1.79 – 3.25), and 2.76 (95% CI: 2.03 – 3.76) to
75 receive knee replacement surgery for PSM, IPW, CA using the propensity score, and CCA using the
76 baseline covariates, respectively. The time-dependent strategies resulted in lower hazard ratios: HR
77 of 2.00 (95% CI: 1.32 – 3.02), 2.05 (95% CI: 1.78 – 2.40), 2.03 (95% CI: 1.83 – 2.21) and 2.13 (95% CI:
78 1.62 – 2.79) for tdPSM, IPW, parametric G-formula and CCA, respectively. Without any adjustment,
79 an HR of 3.15 (95% CI: 2.37 – 4.20) was found.

80 The results from intra-articular corticosteroid injection examples were more consistent between the
81 baseline and time-dependent methods. Patients that receive intra-articular corticosteroid injections
82 had a higher risk to receive knee replacement surgery with an HR of 1.64 (95% CI: 1.42 – 1.92), 1.53
83 (95% CI: 1.42 – 1.65), 1.58 (95% CI: 1.33 – 1.88), and 1.59 (95% CI: 1.36 – 1.87) for the baseline
84 methods (PSM, IPW, CA using the propensity score, and CCA, respectively) and an HR of 1.61 (95% CI:
85 1.38 – 1.87), 1.49 (95% CI: 1.36 – 1.57), 1.65 (95% CI: 1.53 – 1.85) and 1.63 (95% CI: 1.39 – 1.91) for
86 the time-dependent methods (tdPSM, IPW, parametric g-formula, CCA, respectively). No adjustment
87 resulted in an HR of 2.12 (95% CI: 1.81 – 2.48).

88