Author/s:
Shepherd, DA;Amor, DJ;Moreno-Betancur, M

Title:
Statistical analysis of observational studies in disability research

Date:
2024-11-01

Citation:
Shepherd, D. A., Amor, D. J. & Moreno-Betancur, M. (2024). Statistical analysis of observational studies in disability research. Developmental Medicine and Child Neurology, 66 (11), pp.1408-1418. https://doi.org/10.1111/dmcn.15948.

Persistent Link:
https://hdl.handle.net/11343/351347

Received: 29 December 2023 | Accepted: 4 April 2024

DOI: 10.1111/dmcn.15948

INVITED REVIEW

Statistical analysis of observational studies in disability research

Daisy A. Shepherd[1,2,3] | **David J. Amor**[1,2,3,4] | **Margarita Moreno-Betancur**[1,2]

[1]Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia

[2]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Victoria, Australia

[3]Neurodisability and Rehabilitation, Murdoch Children's Research Institute, Melbourne, Victoria, Australia

[4]Neurodevelopment and Disability, Royal Children's Hospital, Melbourne, Victoria, Australia

Correspondence
Daisy A. Shepherd, Murdoch Children's Research Institute, 50 Flemington Road, Parkville 3052, Victoria, Australia.
Email: daisy.shepherd@mcri.edu.au

Funding information
National Health and Medical Research Council, Grant/Award Number: 2009572

Abstract

Observational studies have a critical role in disability research, providing the opportunity to address a range of research questions. Over the past decades, there have been substantial shifts and developments in statistical methods for observational studies, most notably for causal inference. In this review, we provide an overview of modern design and analysis concepts critical for observational studies, drawing examples from the field of disability research and highlighting the challenges in this field, to inform the readership on important statistical considerations for their studies.

Observational studies have a critical role in health and medical research, providing the opportunity to address a range of research questions. In disability research, such studies can be used, for example, to describe developmental trajectories and prevalence of disabilities, identify individuals at risk of secondary health concerns, or investigate the impact of interventions to improve health outcomes when conducting a trial may not be ethical, timely, or feasible.

Effective use of observational data relies heavily on robust statistical design and analysis methodology, particularly in the disability field, which presents specific challenges. Over the past decades, there have been substantial developments in statistical methods for observational studies, most notably for causal inference. However, the uptake of these methodologies in practice has been slow and often limited by lack of access to biostatistical expertise or training.

In this review, we provide readers with an overview of modern concepts and analytical methods for observational studies, focusing on their use in disability research. We first review the importance of formulating a clear research question and the three key types of research questions (descriptive, predictive, and causal). We then provide an overview of the key statistical considerations relevant to defining and addressing each question type, with focus on important principles to strengthen the quality of observational studies. We conclude with a brief overview of key considerations when reporting and interpreting the results from these studies.

Throughout, we emphasize the importance of planning statistical analyses in advance, regardless of the type of study. A thorough statistical analysis plan (SAP) is vital to ensure a well-designed analysis, with statistical methodological decisions considered and justified. We present the readers with a statistical analysis plan template used in our institutes and hope this tool can provide guidance for others who are planning observational studies.[1]

THE TYPE OF RESEARCH QUESTION

The starting point to any study (trial or observational) is a clearly defined research question that articulates the research aim in terms of a statistical question. This provides a crucial foundation for subsequent design and analytical decisions. Although seemingly trivial, developing a clear research question can be one of the most time-consuming stages of any study.

Abbreviation: DAG, directed acyclic graph; SAP, statistical analysis plan

It can be argued that research questions can be classified into three types according to their purpose: descriptive, predictive, and causal.[2] All three are relevant to the disability field and each entails specific design and analytical concepts as summarized in Table 1. So, how is the type of research question determined? In general, by considering the translational intent of the research.[3] For example, is the study aiming to inform the extent of a health problem (descriptive questions), who is at higher risk (predictive), or how to treat or prevent it (causal)?

What this paper adds

- Descriptive research questions have specific analytical complexities, so careful statistical design before analysis is critical.
- Prediction research aims to produce a model with good predictive ability and requires thorough statistical design prior to analysis.
- Causal research requires careful statistical analysis planning, facilitated by modern causal inference concepts and analytical methods.
- Adopting these approaches will strengthen the quality of observational studies addressing a range of research questions in the disability space.

DESCRIPTIVE RESEARCH QUESTIONS

Descriptive questions are fundamental to disability research and aim to characterize the distribution of a feature or outcome across a population of interest, for example, describing developmental trajectories and disability prevalence over time or describing the characteristics of understudied groups that may not be well understood.[4]

For many years, descriptive research questions garnered the misleading label of being 'simpler' or 'more straightforward' compared to predictive or causal research questions; however, recent work has emphasized the importance of descriptive epidemiology and the statistical rigour required to conduct good descriptive studies.[4–6] Indeed, descriptive research questions present their own analytical complexities and are at risk of selection and measurement bias, so careful design is paramount.

Study design (pre-analysis considerations)

The design framework of Lesko et al.,[4] outlined in Table 2, provides a useful tool to aid the design process for descriptive research questions. The design consists of three steps.

Step 1 is to define the target estimand (i.e. what you are interested in knowing), which is characterized by the target population, outcome of interest, and how the outcome distribution will be summarized (e.g. prevalence, mean). It is important to specify, if relevant, any stratified (or subgroup) analyses of interest. This is particularly relevant in heterogeneous populations, where there is a need to describe the outcomes for different subgroups of the population of interest.

Step 2 is to plan the statistical analysis by specifying the data decisions (e.g. how you will use your study sample) to align as closely as possible with the target estimand, noting any disparities and the potential for bias to be introduced. Analytical decisions should be based on these design considerations, with steps taken to minimize potential biases.

For example, one might plan to use some form of adjustment (i.e. perturbing an estimate in some way) to reduce selection bias (i.e. by rebalancing the covariate distribution in the analytical sample in some way to be more representative of the target population). Causal diagrams can be used to help identify such potential sources of selection bias (see the section on 'Causal research questions' for further discussion). However, for many years the notion of adjustment has not been well understood in the setting of descriptive studies. In fact, understanding the utility of adjustment is complex and we direct readers to the discussions by Kaufman[7] while reinforcing that if an adjustment is to be made in the descriptive setting, the motivation should be clear and justified and the results interpreted accordingly.

Analytic methods

Once the statistical analysis has been designed, step 3 is to estimate the summary measure of interest using an appropriate analytical method. If there are no missing data or other source of selection or measurement bias, this could be as simple as estimating the sample summary statistic.

In Table 2, we present an illustrative example of this process based on a previously published study that describes the mean developmental spoken language trajectories of children (aged from 18 months to 8 years) with cerebral palsy.[8] In the presented example, the use of the framework described enabled clear identification of additional sources of potential selection bias (a result of the convenience sampling design and exclusion of children where no parents could speak Dutch fluently), which was not discussed in the published study. In addition, the framework enhanced transparency of the assumptions underpinning the analytical approach applied (the parametric assumption of the mean trajectory being the same for the whole sample across the age window), allowing researchers to be aware of and thus assess whether this may be a valid assumption (or not).

Within the disability field, one of the most complex challenges in descriptive research questions, broadly speaking, is the potential issue of selection bias. Disabled populations tend to be heterogeneous, with fundamental differences

TABLE 1 General overview of the statistical design and analysis principles relating to each type of research question[a].

	Descriptive	Predictive	Causal
Example(s)	What are the developmental trajectories of spoken language comprehension in children with CP? (See Table 2 for a worked example) What is the prevalence of spoken language comprehension difficulties in younger children with CP compared to older children with CP?	How accurately does parent-reported gross motor function at age 2 years predict significant movement difficulties at age 5 years in children born extremely preterm?	What is the impact of distal rectus femoris transfer in ambulatory children with CP on knee flexion range of motion? (See Table 3 for a worked example)
Study design (pre-analysis considerations and approaches; development of an SAP)	1. Define key aspects of your question (target population, outcome, summary measure). If relevant, specify whether any subgroup analyses are of interest. 2. Plan the analysis to answer the question by specifying data decisions (analytical study sample, outcome measure) to minimize the risk of bias.	1. Specify key aspects of your question (target population, outcome, predictors). 2. Identify whether your data are appropriate (analytical study sample, outcome measure, predictor measures) in addition to whether the sample size is sufficient. 3. Identify an appropriate model building method dependent on considerations in (2).	1. Define the causal effect by specifying the 'target trial'. 2. Identify potential sources of bias by developing causal diagrams (i.e. DAGs). 3. Plan the analysis to minimize identified biases by developing an appropriate emulation strategy to align closely with the 'target trial'.
Data analysis	3. Estimate the summary measure using an appropriate analytical method, for example, sample summary statistic in the setting with no missing data or other source of selection bias and no measurement error.	4. Develop your prediction model. This may involve considering variable selection (if sample size is still a challenge) and functional forms (if specifying a regression or other parametric model). 5. Evaluate the predictive performance of your model using a suitable approach and update the model (if required). 6. Evaluate the performance of the prediction model on a new data set (external validation).	4. Estimate the causal effect using an appropriate analytical method, for example: • Adjust for confounding using multivariable regression, G-methods, doubly robust approaches. • Handle missing data using IPW, multiple imputation, etc. 5. Conduct informal sensitivity analysis or preferably formal quantitative bias analysis to assess the robustness of findings to alternative assumptions.

Abbreviations: CP, cerebral palsy; DAG, directed acyclic graph; IPW, inverse probability weighting; ROM, range of motion; SAP, statistical analysis plan.

[a]We provide these as rough guidelines and do not cover the complete nuances that may be relevant for each individual question. We refer readers to a SAP template that we use at our institutes.[1]

across their ages and developmental stages. Therefore, ensuring that analytical samples are representative of the population we are trying to describe and generalize to (or being aware of how they are not) is critical. Taking time to carefully apply the design framework presented in this review will benefit this process, enabling selection bias to be minimized or clearly acknowledged in the interpretation of studies addressing descriptive research questions.

PREDICTIVE RESEARCH QUESTIONS

Predictive research questions aim to predict the risk or mean of an outcome based on measured factors. Such questions may look at the predictive ability of a single predictor, for example, parent-reported gross motor function at age 2 years to predict movement difficulties at age 5 years in children born extremely preterm.[9] Alternatively, several predictors may be combined into a single model, for example, using multiple movement measures taken in the first 16 weeks of life to predict developmental delay at age 2 years in infants born very preterm or infants of very low birthweight.[10] Accurate prediction models can help identify individuals at higher risk and who may therefore require different clinical care.

Over the last decade, advances in prediction research have focused on analytical methodologies; yet, as with other types of research questions, careful statistical design before analysis is critical. Importantly, the goal of prediction is not unbiased estimation (as is the goal of descriptive and causal questions), but rather to produce a model with good predictive ability (i.e. that can accurately identify individuals at risk).

Study design (pre-analysis considerations)

The design stage for predictive research follows a similar process to descriptive and causal questions, that is, first

TABLE 2 Descriptive research question applying the framework of Lesko et al.[4] to a published study[a].

Target estimand	Data decisions	Potential biases
Target population: Children with CP in the Netherlands aged from 18 months to 8 years. Exclusions: • Children with severe auditory problems. • Children with severe visual problems. • Children with severe cerebral visual impairment.	Analytic sample: Children with CP (aged from 18 months to 8 years) in the Netherlands, recruited between November 2017 and August 2018 using convenience sampling (hospitals, rehabilitation, day care centres). Recruited in three age groups: • Toddlers (18 months to 3 years 11 months). • Preschool children (4 years to 5 years 11 months). • School-age children (6 years to 8 years 11 months). Exclusions: • Children with severe auditory problems (hearing threshold ≥ 31 dB for the best ear). • Children with severe visual problems (<0.3 corrected with spectacles for the best eye). • Children with severe cerebral visual impairment. • Children with parents who did not speak Dutch fluently (required one parent to be able to do this to be included). • Children with baseline measure only (i.e. no follow-up outcome measures).	Selection bias may be an issue due to: 1. convenience sampling design because only children accessing services (hospitals, rehabilitation, day care centres) are recruited; 2. exclusion of children where no parents could speak Dutch fluently; 3. exclusion of participants with no follow-up outcome measures beyond the baseline measure. The above considerations mean that we have a potential risk of selection bias as the analytical sample may not be reflective of the target population, especially if we believe that (1) children not accessing these services differ from those who do, (2) children with non-Dutch speaking parents differ from those who are fluent, and (3) children who did not complete follow-up differed from those who did. In the published study, item (3) was acknowledged, stating that there was a higher proportion of AAC users in dropouts. The analytical sample therefore may underrepresent AAC users relative to the target population. This could potentially lead to selection bias, although this depends on the causes of completing follow-up and how they related to the outcome.[48] The study could have used alternative analytical approaches to minimize the potential selection bias as a result of loss to follow-up (e.g. multiple imputation). Items (1) and (2) were not discussed in the published study.
Outcome measure: SLC assessed with the C-BiLLT (good accessible tool) across 18 months to 8 years.	Outcome measure: SLC assessed with the C-BiLLT (good accessible tool). Raw scores used. The outcome was measured at three time points: (1) baseline (time of recruitment); (2) 12 months after recruitment; and (3) 30 months after recruitment. The age when these measurements occurred varied for each child.	No single participant had outcome measures spanning across the time window (18 months to 8 years) of the developmental trajectories (each individual covers 30 months at most and was measured at varying ages), that is, there were missing data across different age ranges for individuals. Therefore, to estimate the mean developmental trajectory, the authors used an analytical approach (linear mixed models) to model available outcome measures over the age range. This approach smooths over all individuals in the sample by making the parametric assumption that the mean trajectory is the same for the whole sample across the age range of interest (e.g. those measured at a younger age would have the same mean development as those measured at an older age). If this is not a reasonable assumption, then the study is at risk of bias because of misspecification of the parametric model (parametric assumptions bias). Measurement bias may be an issue if the instrument (C-BiLLT) is not an appropriate tool to measure SLC in this population.
Summary measure: Mean vocabulary score trajectory across the age range of 18 months to 8 years.	–	–

Abbreviations: AAC, augmentative and alternative communication; C-BiLLT, Computer-Based instrument for Low motor Language Testing; CP, cerebral palsy; SLC, spoken language comprehension.

[a]The study aimed to describe the developmental trajectories of spoken language comprehension in children with CP across the ages of 18 months to 8 years.[8]

specifying key aspects of the research question before considering how feasible it is to answer it with the observational data at hand (Table 1, steps 1 and 2). Key considerations in predictive research are whether the sample is representative of the target population, whether the outcome has been measured using a valid reliable measure, and whether the predictors are accurately measured and easily reproducible.

The first aspect relates to selection bias and will impact whether the prediction model performs well in the target population (e.g. performance may be poor if the prediction model is developed using a non-representative sample). The second and third aspects relate to measurement bias and how useful the prediction model will be in real-world settings (e.g. utility will be limited if predictors are not measurable in routine practice in a reproducible way).

It is also important to consider whether the sample size is sufficient to develop the prediction model. This may be a challenge in disability research where sample sizes can be restricted. Previous guidance suggested '10 events per variable' as a general rule of thumb, although this has little practical use.[11] More recent advice has suggested considering the complexity of the intended prediction model and its predictive performance.[11] Considering the sample size can help determine whether building a prediction model is feasible, and if so, guide the selection of an appropriate model building approach (e.g. use of machine learning or whether variable selection is required to select a subset of predictors; Table 1, step 3).

Analytic methods

The analytical process tends to follow a consistent workflow (broadly outlined in Table 1, steps 4–6) and broadly involves building a prediction model and assessing its performance before externally validating the prediction model. In the multivariable setting, traditional model building approaches focused heavily on the use of multivariable regression (e.g. logistic regression in the binary outcome setting). This approach has the benefit of an accessible implementation, although it is less flexible because it relies on potentially simplistic assumptions about the form of the relationships between variables, which could have impacts on the model's predictive ability.

Recent methodological advancements have seen a shift towards the use of machine learning methods for prediction.[12] Such methods relax regression assumptions, enabling more flexible modelling when building a prediction model (e.g. the model adapts and learns from the data), which can lead to a better predictive ability. However, machine learning methods also require substantially larger sample sizes,[11] which may not be a feasible option in disability research. Therefore, while the promise of more advanced modelling methods offers some advantages, we encourage readers to consider the most appropriate method for their purpose in light of the available sample size.

After building the model, it is important to assess the predictive performance of the model (Table 1, step 5), which may lead to further model modifications. The metrics used to assess performance depend heavily on the predictive ability of the model, with traditional methods focusing on discrimination (how well the model distinguishes between those experiencing the outcome and those who do not), calibration (how well the predicted outcome distribution matches the observed distribution),[13] and overall measures. However, these metrics do not accurately convey the clinical utility of a prediction model (i.e. is the model worth using at all or whether alternative models, corresponding to different clinical strategies, should be used). Other approaches to performance assessment, such as decision-curve analysis,[14] focus on these key aspects and complement traditional metrics.

Once an appropriate prediction model has been finalized, it is important to externally validate the model on a new data set to understand its wider utility (Table 1, step 6). However, this step is often overlooked in practice. External validation presents an additional challenge in disability research because it relies on the availability of a relevant data set external to the current study, which in the setting of small and local populations is not always available. However, with the growing availability of data resources (e.g. electronic medical records, cohort studies), we hope that the opportunities for external validation will increase in the disability setting.

Designing and conducting a predictive research study can be time-consuming and complex; meanwhile, there are multiple barriers for prediction models to be implemented in practice, particularly if they have not been well validated externally. Therefore, it is critical to consider the need for prediction model development in the first place. If there are existing prediction models (assuming minimal limitations of the model), there may be no utility in building an additional model; rather, efforts may be better invested in validating that model with your own data.[15]

CAUSAL RESEARCH QUESTIONS

Central to disability research are questions of causality, understood here as questions about the impact of exposures, treatments, or interventions on health outcomes.[2] Causal questions are central to informing clinical decision-making and practice. The randomized controlled trial is widely considered to be the criterion standard for answering questions around causation; however, conducting a trial may not always be feasible, timely, or ethical.[2] In addition, the use of randomized controlled trials in the disability field may not be optimal because of the heterogeneous nature of the populations involved.[16]

Using observational data to answer causal questions presents an important alternative. However, causal effect estimation in observational studies presents its own challenges beyond potential heterogeneity because it relies on several (sometimes strong) assumptions,[17] in particular, that confounding bias due to the lack of randomization can be adjusted for. This highlights the need for careful statistical analysis planning, which is facilitated by modern causal inference concepts and analytical methods, as has already been highlighted in the disability field.[18]

TABLE 3 Example of the target trial framework based on a previously published observational study that estimated the impact of distal RFT[a].

Protocol component	Target trial *If we could conduct a randomized controlled trial, what would this look like?*	Emulation strategy[b] *Using our observational data, how will we emulate the target trial to minimize potential biases?*	Remaining biases and comments *Even after the emulation strategy, what biases remain?*
A. **Eligibility criteria**	**Target population:** Ambulatory children (GMFCS levels I–III) aged 6–16 years with CP and spastic bilateral CP who are eligible to receive an SEMLS. Exclusions: • Previous lower-limb surgery. • Dorsal rhizotomy. • Diagnosis of dystonic CP.	**Analytic sample:** Ambulatory children (GMFCS level I–III) aged 6–16 years with CP and spastic bilateral CP who receive routine clinical evaluation at the James R. Gage Center for Gait and Motion Analysis (at Gilette Children's Speciality Healthcare). Authors did not specify the time (e.g. years). Exclusions: • Patients who did not provide written consent for the use of their medical records. • Patients who did not have preoperative and postoperative gait analysis evaluations within a 2.5-year time span.	Selection bias may still be present because of: 1. Selection of participants for the study; not all ambulatory patients with CP and spastic bilateral CP are seen in the gait laboratory or consent to be in the study. 2. Study did not mention the exclusion of those with previous lower-limb surgery or dorsal rhizotomy. 3. Selection of participants into the analytical sample due to loss to follow-up or missing data only included individuals who have both before and after outcome measures. These considerations mean we have a potential risk of selection bias because the analysis may exclude a certain subpopulation and thus may not be representative of the target population (e.g. if those seen in the gait laboratory differed from those not seen, or if those who consented differed from those who did not). In the published study, items (1) and (3) are acknowledged as a repercussion of standard of care health referrals, health insurance coverage, and patient and provider preferences, although not explicitly acknowledged as potential sources of selection bias. Item (2) is not discussed in the published study. In Figure 1, we present example DAGs that consider potential selection bias due to item (3) and outline potential approaches to deal with it. The authors aimed to reduce selection bias due to missing confounder information by treating missing values as a category in the propensity score model. This approach may be appropriate in certain settings.[49] An alternative approach to managing both incomplete confounder and follow-up data could be applied to reduce bias (e.g. multiple imputation).
B. **Treatment strategies**	**Treatment arms in the trial:** Intervention: Individuals receive bilateral RFT (administered in a standardized way) as part of an SEMLS (unspecified type of SEMLS). Comparator: Individuals only undergo an SEMLS (unspecified type of SEMLS) without bilateral RFT.	**Treatment and exposure measures:** Intervention: individuals receive bilateral RFT (administered in a standardized way) as part of an SEMLS (any type of SEMLS). Comparator: individuals only undergo an SEMLS (any type of SEMLS) without bilateral RFT.	The specified treatment strategies contain unavoidable variation because of the variation in the type of SEMLS performed. The type of SEMLS received is specific to each individual; therefore, the treatment strategies will naturally vary. A better approach to accommodate for this variation may be to group similar individuals with similar SEMLS and estimate the causal effects within each group. This would help to separate the impact of the RFT from the impact of the type of SEMLS and the related individual circumstances that represent different subpopulations for which effects may well vary.
C. **Assignment procedures**	**Randomization strategy:** Random assignment to either the treatment or control arm at the time of surgery, unblinded to participants and clinicians.	**Selection of confounders:** Age, sex, maximum knee extension, maximum knee flexion, RFT spasticity (all measured at the preoperative physical examination). **Approach to adjustment:** PD. Propensity score matching (approach used in published study).	Several biases may still be present: 1. Confounding bias: by adjusting for potential confounders, the study aims to minimize confounding bias. However, this assumes the specified confounders are sufficient to adjust for all confounding. There may still be the potential for residual confounding (e.g. if confounders are unmeasured or mismeasured), which needs to be considered when interpreting the results (e.g. differences in concomitant surgery between treatment groups that could introduce bias). This was acknowledged in the published study, with sensitivity analyses conduced to investigate the robustness of findings with respect to unobserved bias. 2. Parametric assumption bias: the use of propensity score matching may have limitations as acknowledged in the study. In addition, this may introduce bias due to the assumptions of the approach (e.g. that the propensity score model is correctly specified). Other approaches to adjustment may be more appropriate in some settings (e.g. other g-methods, doubly robust approaches).

TABLE 3 (Continued)

D. **Follow-up period**	**Start and end times:** Starts at randomization (i.e. from surgery); ends 1 year after surgery.	**Timing of measures:** Starts from surgery; ends at the postoperative examination (a maximum of 2.5 years after surgery).	There may be a risk of measurement bias. In the target trial, the follow-up would be measured at a consistent time for everyone, whereas in the observational study, this varies for individuals and covers a wider range of times. This may introduce a potential for measurement bias because the outcomes will vary depending on when they are measured. This would need to be considered when interpreting the causal effect, which was not acknowledged in the published study.
E. **Outcome measure**	**Outcome:** Knee flexion ROM, measured at a postoperative clinical examination, conducted by a study nurse and physiotherapist. Outcome assessment could be blinded.	**Outcome measure:** Knee flexion ROM, measured at a postoperative clinical examination. Outcome assessment was not blinded.	There may be a risk of measurement bias if the postoperative examination is not blinded or standardized as it would be in a trial. This is not discussed in the paper.
F. **Subgroup analyses**[c]	**Subgroups of the population for which it is important to obtain separate effects (if relevant):** According to severe flexed-knee gait pattern (yes/no). Severe flexed-knee gait pattern was classified as having a knee flexion in the stance phase ≥ 35 degrees and peak knee flexion on swing ≥ 1 SD of the norm, as measured at a preoperative clinical examination performed 1 day before surgery.	**Subgrouping variable(s) measure(s) and approach:** Stratified analyses using a derived categorical variable representing a severe flexed-knee gait pattern (based on the measured knee flexion in the stance phase and peak knee flexion on swing from the preoperative clinical examination). The timing of this examination is not clear in the published study.	There may be a risk of measurement bias due to: 1. The varied timing of the preoperative examination affecting how accurately this is measured. 2. A challenging examination affecting the ability to accurately measure knee flexion in the stance phase, for example. In addition, depending on the analytical approach used for the subgroup analyses, there may be a risk of parametric assumptions bias due to the assumptions underpinning the selected approach.
G. **Causal contrast of interest and causal effect measure**	Difference in mean outcome in the intervention arm versus the control arm in the target population (average causal effect in the difference scale).	–	–

Abbreviations: CP, cerebral palsy; DAG, directed acyclic graph; GMFCS, Gross Motor Function Classification System; RFT, rectus femoris transfer; ROM, range of motion; SEMLS, single-event multilevel surgery.

[a]The study estimated the impact of distal RFT as part of an SEMLS in ambulatory children with CP on knee flexion ROM (limited to one outcome for brevity in this example).[34]

[b]The presented emulation strategy is based on the approach conducted in the published study.

[c]Subgroup analyses were not an objective in the published observational study.[34] Instead, we used the secondary aim from the originally published randomized controlled trial[50] for illustrative purposes.

Study design (pre-analysis considerations)

One of the biggest challenges in causal inference is designing the study optimally, such that potential biases are minimized (confounding, selection, and measurement bias[19]). Over recent years, several tools have been developed to guide the planning and interpretation of causal analyses using observational data.[20–24] The target trial framework is one such example and involves specifying (via several key components) a hypothetical randomized controlled trial (i.e. the 'target trial') that could be used to answer the causal question of interest.[25,26] Given the complexity of designing trials people with disability,[16] specification of the target trial (even if just hypothetical) most probably will be challenging; however, doing so provides clarity and precision to what exactly we are trying to estimate, that is, the causal effect of interest (Table 1, step 1).

The analysis of observational data to answer causal questions can then be conceptualized as an attempt to emulate this target trial to obtain an estimate of the causal effect specified. Beyond the lack of randomization, there are many other potential differences between the target trial and the observational study that need to be considered.[27] The use of causal diagrams (also referred to as directed acyclic graphs [DAGs]) is an effective tool to help identify such potential sources of bias (Table 1, step 2). DAGs are a visual depiction (via nodes and one-way arrows) of the assumed causal structure underlying the observational data, driven by substantive content knowledge and often incorporating unverifiable assumptions. Their use is increasing, including in recent publications in this journal.[28–30] The use of DAGs can be extremely beneficial in making assumptions transparent, particularly when the causal structure may be highly subjective or complex (e.g. the aetiology of cerebral palsy). We refer readers elsewhere for further information on DAGs[20,24,31] and encourage their use in disability research.

By referring to the target trial and using DAGs to identify potential biases, an appropriate emulation strategy (e.g. analytical decisions on how the observational data will be used) can be designed to minimize potential sources of bias (e.g. selection of a confounding adjustment set; see VanderWeele[32] for a more in-depth discussion) and identify any potentially remaining ones (Table 1, step 3). This highlights the study strengths and limitations and thus informs the interpretation of findings. Application of the target trial framework in health research is increasing,[33] although we are not aware of any published studies in the disability field.

To illustrate, we present an example (Table 3) applying the target trial framework to a study from this journal[34] alongside two example DAGs (Figure 1). The example provides a clear demonstration of the benefit of applying the target trial framework and particularly how vital this can be for a heterogeneous patient population of interest. The target trial enables clear definition of the causal estimand of interest, including clarity around the intervention of interest (rectus femoris transfer as part of any type of single-event multilevel surgery) and helps identify potential sources of bias that were not originally acknowledged in the published study (e.g. the potential measurement bias due to the varied follow-up times in the observational study). The example DAGs depict situations where (a) the bias introduced by loss to follow-up could not be minimised and (b) constructing a DAG can help identify an appropriate adjustment set that may be able to minimize this potential bias (by adjusting for health insurance, if measured).

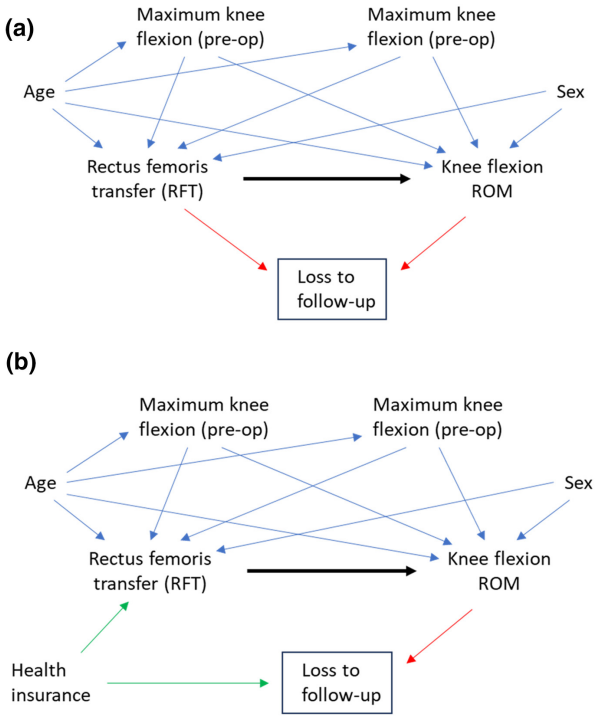

FIGURE 1 Example directed acyclic graphs (DAGs) for a previously published observational study[34] for illustrative purposes (Table 3), with focus on the identification of a potential source of selection bias. The **black** arrow represents the causal effect of interest: the impact of distal rectus femoris transfer (RFT) (exposure) on knee flexion range of motion (ROM) (outcome). The **blue** arrows represent the confounders and their assumed relationships with the exposure, the outcome, and each other (determined by subject-matter knowledge). For example, the arrow from age to the exposure represents the assumption that the child's age influences whether they receive an RFT or not. The arrow from age to the outcome represents the assumption that the age of the child may affect their current ROM. Note that we have assumed that age also influences the knee flexion measurements preoperatively, as indicated by the arrow. The **red** arrows represent the relationship with loss to follow-up. Loss to follow-up is shown in a box to indicate that analysis has been restricted to those with only complete outcome measures. Based on the rules of causal diagrams,[26–28] this opens up a non-causal backdoor pathway between exposure and outcome, thus introducing potential selection bias. Based on the DAG in (a), we assumed that loss to follow-up is influenced by the surgery in addition to the outcome (e.g. individuals with worse outcomes may not attend a postoperative examination). If we believe that this causal structure is the most appropriate, we may not be able to minimize the potential selection bias and would acknowledge this as a limitation of our study. The **green** arrows represent a potential relationship underlying the loss to follow-up. For example, whether a family has health insurance affects whether they receive the surgery; it also impacts whether they have a postoperative examination and thus a measurement of the outcome after surgery. Based on the assumed causal structure in (b), we may be able to minimize selection bias due to loss to follow-up by adjusting for health insurance (if measured).

Analytical methods

After thoroughly defining the target trial and a suitable emulation strategy, an appropriate analytical method is required to estimate the causal effect of interest (Table 1, step 4). In this review, we focus on methods that aim to minimize potential confounding bias through adjustment for the selected confounder set, broadly referred to as confounding adjustment methods. However, we emphasize that additional analytical approaches would be required in a study to manage selection bias, including due to missing data (e.g. multiple imputation) and measurement error (e.g. regression calibration).

For many years, the most common confounding adjustment method for causal effect estimation has been multivariable outcome regression, where the outcome is regressed on the exposure and confounders. However, in the point exposure setting, this method requires the following assumptions for unbiased estimation: that the specified model is correctly specified (an untestable assumption) and that the causal effect is constant across the substrata of the confounders. In many practical settings, particularly those in disability research with complex diseases and heterogeneous populations, these assumptions are not realistic. Furthermore, this approach is not applicable in complex settings, such as problems with time-varying exposures or time-varying confounding.

More flexible methods have been developed that relax the assumptions of regression and are more widely applicable, and thus may be more appropriate in this field. Such approaches include g-methods (e.g. inverse probability weighting, g-computation, g-estimation)[35] in addition to the more recently developed 'doubly robust' methods (e.g. augmented inverse probability weighting,[36] targeted maximum likelihood estimation[37]). The implementation of g-methods is not too far distanced from multivariable regression in the point exposure setting,[38] although these approaches also require a correctly specified model (outcome or propensity score model), which again is an untestable assumption.

The 'doubly robust' methods, garnering their name through a double modelling procedure (using both an outcome and propensity score model), have the added advantage of being more robust to model misspecification by only requiring one of the two models to be consistently estimated. In addition, they have the advantage of being able to incorporate machine learning in fitting the two models, thus further avoiding strong modelling assumptions.

However, a caveat of doubly robust methods incorporating machine learning falls with their reliance on larger sample sizes in addition to ongoing research regarding valid inference with very flexible machine learning approaches.[39] Therefore, g-methods or doubly robust methods coupled with parametric modelling approaches may have greater use in this research field, given the presence of heterogeneous populations, where we expect causal effects to vary more across individuals, and potential restrictions on sample size (e.g. in the setting of low-prevalence disabilities).

We also acknowledge that in most settings, the risk of bias is unavoidable despite best efforts to minimize it. Informal sensitivity analyses (e.g. considering different confounder sets or different approaches to handle missing data) or formal quantitative bias analysis have great utility in this setting, with the latter allowing quantification of the potential magnitude and direction of biases, while providing an estimate of uncertainty arising from systematic errors.[40] Application of such methods in disability research would undoubtedly contribute to more robust clinical recommendations from observational studies.

STATISTICAL REPORTING AND INTERPRETATION

Reporting the statistical aspects of a study can be challenging, particularly when complex design and analytical methods have been applied. We outline several key considerations applicable to all three types of research questions.

First, statistical design decisions should be communicated clearly. Restrictive word limits can present a problem to the level of detail; therefore, presenting additional information in supplementary material may be a viable option. Furthermore, when reporting study results, it is important to interpret results in light of these design considerations and methodological choices, acknowledging potential remaining biases as limitations of the study.

When presenting results, we reinforce the need to move away from dichotomous questions (e.g. 'Is there an effect?') and the related notions of null hypothesis testing and 'statistical significance'. These approaches lead to a dichotomous interpretation of confidence intervals and *p*-values that can be problematic for reasons well outlined elsewhere.[41,42] The interpretation of results should consider instead the magnitude of estimates (alongside a confidence interval and a *p*-value), and consider whether they have any clinical relevance, in addition to the uncertainty in the study due to systematic errors, which the framework described helps to understand. In addition, we caution readers to consider how results from multivariable regression are presented and avoid the common 'Table 2 fallacy'.[43]

Finally, we highlight the availability of reporting frameworks (e.g. Strengthening the reporting of observational studies in epidemiology [STROBE],[44] Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis [TRIPOD],[45] and the upcoming Transparent reporting of observational studies emulating a target trial [TARGET] framework[46]) and encourage their use for transparent reporting of observational studies in the disability research space. We also refer readers to the previous publication by Rigby, which provides a useful angle to some of these concepts from the perspective of a statistical reviewer.[47]

CONCLUDING REMARKS

Given the possibilities of observational studies to address important research questions in the field of disability, it is critical that best practice is applied in their analysis to ensure that studies are of high quality with robust conclusions. This review has provided an overview of modern advancements in concepts and methods, with key focus on the importance of defining the question and carefully considering the design to guide the data analysis and aid in the interpretation of findings. Adopting these modern approaches will undoubtedly be a worthwhile endeavour to strengthen the quality of observational studies addressing a range of research questions in the disability space.

ACKNOWLEDGEMENTS

We thank the Victorian Centre for Biostatistics-Clinical Epidemiology and Biostatistics Unit causal team for their contributions to the development of the short course 'Observational studies: modern concepts and analytic methods' and related material on which this review draws. We particularly thank Dr Marnie Downes and Dr Rushani Wijesuriya for their key contributions to the development of the referenced statistical analysis plan template and the prediction material respectively. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

DS and DA were supported by the Lorenzo and Pamela Galli Medical Research Trust. MMB was supported by a National Health and Medical Research Council Investigator grant (no. 2009572).

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no data sets were generated or analysed during the current study.

ORCID

Daisy A. Shepherd https://orcid.org/0000-0001-8540-0473
David J. Amor https://orcid.org/0000-0001-7191-8511
Margarita Moreno-Betancur https://orcid.org/0000-0002-8818-3125

REFERENCES

1. Downes M, Shepherd DA, Wijesuriya R, Dashti G, Chen T, Carlin J, et al. Statistical analysis plan template for observational studies [Internet]. 2023. Available from: https://doi.org/10.26188/12471380.v5
2. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. Chance. 2019;32(1):42–9.
3. Moreno-Betancur M. The Target Trial: A Powerful Device beyond Well-defined Interventions. Epidemiology. 2021;291–4.
4. Lesko CR, Fox MP, Edwards JK. A framework for descriptive epidemiology. Am J Epidemiol. 2022;9–25.
5. Fox MP, Murray EJ, Lesko CR, Sealy-Jefferson S. On the Need to Revitalize Descriptive Epidemiology. Am J Epidemiol [Internet]. 2022; Available from: https://pubmed.ncbi.nlm.nih.gov/28459981/
6. Platt RW. The importance of descriptive epidemiology. Am J Epidemiol. 2022;1–6.
7. Kaufman JS. Statistics, adjusted statistics, and maladjusted statistics. Am J Law Med. 2017;43(2–3):193–208.
8. Vaillant E, Oostrom KJ, Beckerman H, Vermeulen RJ, Buizer AI, Geytenbeek JJM. Developmental trajectories of spoken language comprehension and functional communication in children with cerebral palsy: A prospective cohort study. Dev Med Child Neurol. 2023;(December 2022):1–11.
9. Costa R, Sarrechia I, Anna- AMA, Seppänen V, Ådén U, Zemlin M, et al. Prediction of movement difficulties at 5 years from parent report at 2 years in children born extremely preterm. 2023;(December 2022):1–11.
10. Caesar RA, Boyd RN, Cioni G, Ware RS, Doherty J, Jackson MP, et al. Early detection of developmental delay in infants born very preterm or with very low birthweight. Dev Med Child Neurol. 2023;65(3):346–57.
11. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ [Internet]. 2020;368(March):1–12. Available from: https://doi.org/10.1136/bmj.m441
12. Finlayson SG, Beam AL, Smeden M van. Machine learning and statistics in clinical research articles - moving past the false dichotomy. JAMA Pediatr. 2023;177(5):448–9.
13. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. JAMA - J Am Med Assoc. 2017;318(14):1377–84.
14. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74.
15. van Royen FS, Moons KGM, Geersing GJ, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. Eur Respir J [Internet]. 2022;60(3). Available from: https://doi.org/10.1183/13993003.00250-2022
16. Rosenbaum P. The randomized controlled trial: An excellent design, but can it address the big questions in neurodisability? Dev Med Child Neurol. 2010;52(2):111.
17. Day SM, Reynolds RJ. Rectus femoris transfer surgery in cerebral palsy: can causal inferences be made from observational data? Dev Med Child Neurol. 2021;63(2):129.
18. Konigorski S. Causal inference in developmental medicine and neurology. Dev Med Child Neurol. 2021;63(5):498.
19. Hernan M. Structure of Bias [Internet]. Available from: https://www.hsph.harvard.edu/miguel-hernan/research/structure-of-bias/
20. Hernán M, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
21. Pearl J. Causality: Models, reasoning and inference. Cambridge University Press, New York; 2000.
22. Rubin D. Estimating Causal Effects of Treatments in Experimental and Observational Studies. J Educ Psychol. 1974;66(5):688–701.
23. Robins JM. A New Approach To Causal Inference in mortality Studies with a sustained exposure - Application To Control of the Healthy Worker Survivor Effect. Math Model. 1986;7:1393–512.
24. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.
25. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016;183(8):758–64.
26. Matthews AA, Danaei. Target trial emulation: applying principles of randomised trials to observational studies. Stat Med. 2020;39(8):1199–236.
27. Lodi S, Phillips A, Lundgren J, Logan R, Sharma S, Cole SR, et al. Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples with Apples. Am J Epidemiol. 2019;188(8):1569–77.
28. Burgess A, Boyd RN, Chatfield MD, Wotherspoon J. Hand function and self-care in children with cerebral palsy. Dev Med Child Neurol. 2023;65(3).
29. Larsen ML, Rackauskaite G, Greisen G, Laursen B, Uldall P, Krebs L, et al. Declining prevalence of cerebral palsy in children born at term in Denmark. 2021;(May):715–22.

30. Sorg A Lisa, Kries R von, Klemme M, Gerstl L, Weinberger R, Beyerlein A, et al. Risk factors for perinatal arterial ischaemic stroke : a large case – control study. 2019;513–20.
31. Gaskell AL, Sleigh JW. An Introduction to Causal Diagrams for Anesthesiology Research. Anesthesiology. 2020;(5):951–67.
32. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol [Internet]. 2019;34(3):211–9. Available from: https://doi.org/10.1007/s10654-019-00494-6
33. Hansford HJ, Cashin AG, Jones MD, Swanson SA, Islam N, Douglas SRG, et al. Reporting of Observational Studies Explicitly Aiming to Emulate Randomized Trials: A Systematic Review. JAMA Netw Open. 2023;6(9):E2336023.
34. Schwartz MH, Ries AJ. Rectus femoris transfer in children with cerebral palsy: comparing a propensity score-matched observational study to a randomized controlled trial. Dev Med Child Neurol. 2021;63(2):196–203.
35. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017;46(2):756–62.
36. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. Polit Anal. 2009;18(1):36–56.
37. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol. 2017;185(1):65–73.
38. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. Am J Epidemiol. 2011;173(7):731–8.
39. Naimi AI, Mishler AE, Kennedy EH. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. Am J Epidemiol. 2021;(Ml).
40. Lash TL, Fox MP, Maclehose RF, Maldonado G, Mccandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43(6):1969–85.
41. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50.
42. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat [Internet]. 2016;70(2):129–33. Available from: https://doi.org/10.1080/00031305.2016.1154108
43. Westreich D, Greenland S. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013;177(4):292–8.
44. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Epidemiology. 2007;18(6):800–4.
45. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement [Internet]. 2015 Jan 6 [cited 2023 Jul 13];162(1):55–63. Available from: https://www.acpjournals.org/doi/10.7326/M14-0697
46. Hansford HJ, Cashin AG, Jones MD, Swanson SA, Islam N, Dahabreh IJ, et al. Development of the TrAnsparent ReportinG of observational studies Emulating a Target trial (TARGET) guideline. BMJ Open. 2023;13(9):1–6.
47. Rigby AS. Statistical recommendations for papers submitted to Developmental Medicine & Child Neurology. Dev Med Child Neurol. 2010;52(3):299–304.
48. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. Am J Epidemiol. 2018;187(12):2705–15.
49. Blake HA, Leyrat C, Mansfield KE, Seaman S, Tomlinson LA, Carpenter J, et al. Propensity scores using missingness pattern information: a practical guide. Stat Med. 2020;39(11):1641–57.
50. Dreher T, Götze M, Wolf SI, Hagmann S, Heitzmann D, Gantz S, et al. Distal rectus femoris transfer as part of multilevel surgery in children with spastic diplegia - A randomized clinical trial. Gait Posture. 2012;36(2):212–8.

How to cite this article: Shepherd DA, Amor DJ, Moreno-Betancur M. Statistical analysis of observational studies in disability research. Dev Med Child Neurol. 2024;00:1–11. https://doi.org/10.1111/dmcn.15948