Confounding-adjustment methods for the difference in medians

Daisy A. Shepherd*1,2, Benjamin R. Baer3 & Margarita Moreno-Betancur1,2

1Clinical Epidemiology & Biostatistics Unit, Department of Paediatrics, The University of Melbourne, VIC, Australia
2Clinical Epidemiology & Biostatistics Unit, The Murdoch Children’s Research Institute, The Royal Children’s Hospital, VIC, Australia
3Department of Biostatistics and Computational Biology, The University of Rochester, New York, USA

Correspondence: *Daisy A. Shepherd, The Murdoch Children’s Research Institute, Royal Children’s Hospital, VIC, Australia. Email: daisy.shepherd@mcri.edu.au

Keywords: causal inference, skewed outcomes, potential outcomes, difference in medians, confounding, quantile regression, inverse probability weighted, propensity scores, g-computation

Abbreviations: LSAC, The Longitudinal Study of Australian Children; SDQ, Strengths and Difficulties Questionnaire; QR, quantile regression; WQR; weighted quantile regression; IPW, inverse probability weighted; PS, propensity score
ABSTRACT

With continuous outcomes, the average causal effect is typically defined using a contrast of expected potential outcomes. However, in the presence of skewed outcome data, the expectation may no longer be meaningful and the definition of the causal effect should be considered more closely. When faced with this challenge in practice, the typical approach is to either “ignore or transform” – ignore the skewness in the data entirely or transform the outcome to obtain a more symmetric distribution for which the expectation is interpretable as the central value. In many practical settings, neither approach is entirely satisfactory. An appealing alternative is to define the causal effect using a contrast of median potential outcomes. Despite being a widely acknowledged concept, there is limited discussion or availability of confounding-adjustment methods to estimate this parameter. Within this study, we described and compared confounding-adjustment methods to estimate the causal difference in medians, addressing this gap. The methods considered were multivariable quantile regression, an inverse probability weighted (IPW) estimator, weighted quantile regression and two possible, little-known implementations of g-computation for this problem. The methods were evaluated within a simulation study under varying degrees of skewness in the outcome and applied to an empirical study using data from the Longitudinal Study of Australian Children. Results indicated that the IPW estimator, weighted quantile regression and g-computation implementations minimised bias across all simulation settings, if the corresponding (propensity score or outcome) model was correctly specified, with g-computation additionally minimising the variance in estimates. Multivariable quantile regression offered little improvement compared to an unadjusted analysis. Application to the empirical study illustrated the implementation of these approaches in real world practice. The methods presented within this paper provide appealing alternatives to the common “ignore or transform” approach, enhancing our capability to obtain meaningful causal effect estimates with skewed outcome data.
Causal inference is a central goal of health research, and is widely understood as the task of assessing the impact of intervening on a given exposure on an outcome of interest.\cite{1} In a perfect randomized controlled trial (i.e., with no loss to follow-up, etc.) causal effects can in principle be directly estimated by comparing the estimated mean outcome in those randomised to be exposed with the estimated mean outcome in those randomised to be unexposed. However, estimation of causal effects in observational studies is not as straightforward and requires more sophisticated methods, in particular to adjust for potential confounding due to the lack of randomisation.

With continuous outcomes, causal effects are typically defined as a contrast of the expected potential outcome under exposure versus the expected potential outcome under no exposure, which would be the estimand in the hypothetical target trial that we seek to emulate with observational data.\cite{2,3} However, epidemiological studies may suffer from skewed outcome data, for which the expectation may no longer be interpretable as the central value of the distribution. Examples of skewed outcomes are abundant in health research, particularly in areas where scale scores are used to measure aspects such as self-reported quality of life (via the PedsQL\cite{4}), self-esteem (via the RSES\cite{5}), or childhood behaviour (via the SDQ\cite{6}). Other examples include time-to-event outcomes in the absence of censoring (e.g., survival time), or duration of events (e.g., breastfeeding duration).

When faced with this challenge in practice, the typical approach is to either “ignore or transform” – ignore the skewness in the data entirely and continue to define the estimand as a contrast of expected potential outcomes, or transform the outcome to obtain a more symmetric distribution for which the expectation is interpretable as the central value. Both approaches have their advantages, although neither is entirely satisfactory in many practical settings. Opting to ignore the skewness in the outcome distribution may be appropriate when the expected potential outcomes are of interest, and allows established confounding-adjustment methods to be applied for estimation (e.g., linear regression and standard implementations of g-methods and doubly robust methods\cite{1}). However, if the central value of the outcome distribution is of direct interest, then this approach is not suitable given the expectation is not interpretable as the central value. Transformation of the outcome to be more symmetrically distributed could be an alternative solution. However, this relies on a suitable transformation existing (which may not be feasible for highly skewed distributions), and makes interpretation of the causal effects more complex than interpretation in the original scale (e.g., log-years instead of years).

In such a situation, defining the causal effect using a contrast of median potential outcomes could be an appealing alternative. Although the contrast of expected potential outcomes is the most commonly used estimand, the idea of defining the causal effect using the median is not a novel
one; the causal effect has been generally defined as a contrast of any functional of the distributions of counterfactual outcomes under different exposure values.\cite{7} Despite being a widely acknowledged concept, there is scarce availability and awareness of confounding-adjustment methods for estimation of the causal difference in medians in practice.

A handful of previous studies have acknowledged the need for such methods, and presented derivations of approaches to estimate causal effects as contrasts of distribution quantiles more generally. A study by Zhang et al. (2012) derived a number of methods - a quantile regression estimator, an inverse-probability weighted estimator, and a stratified estimator using propensity scores. Based on results from a simulation study and application to an empirical dataset, the authors suggested that the quantile regression approach was advantageous in terms of reduced bias and improved efficiency.\cite{8} A more recent study in the field of environmental science defined a novel “overlap weighting” estimator which uses a class of balancing weights from functions of the propensity score model to weight each group to a selected target population.\cite{9} The performance of three estimation methods (propensity score regression, inverse probability weighting and overlap weighting) was explored under homogenous and heterogeneous causal effects and across a number of simulation settings. The study recommended the inverse probability weighting approach if the effect is believed to be heterogeneous, with the other two methods preferred otherwise.\cite{9}

Despite these advances, application of these confounding-adjustment methods in health and medical research remains scarce. The current discussion and evaluation of such methods is relatively limited, which has potentially led to a lack of awareness in their existence. Furthermore, to the best of our knowledge, the use of g-computation in the context of the causal difference in medians has not been widely discussed, let alone been studied in relation to other approaches. In addition, there has been little investigation of how these methods perform and compare in realistic settings across various scenarios, specifically in terms of the degree of skewness in the outcome which has only been explored minimally and not for all the methods considered here.\cite{9}

In this paper we aimed to describe and compare confounding-adjustment methods to estimate the causal difference in medians, with the aim to increase understanding of their utility and encourage application in practice where appropriate. The methods considered were multivariable quantile regression, an IPW estimator, weighted quantile regression and two possible, little-known implementations of g-computation. We begin this paper by defining the causal effect of interest and identifiability assumptions alongside an illustrative example from the Longitudinal Study of Australian Children (LSAC). Then we outline the confounding-adjustment methods considered. Thirdly, we outline the methodology and report findings from a simulation study motivated by the LSAC example, assessing the performance of each method in a realistic setting and considering
varying degrees of skewness in the distribution of the outcome variable. The paper concludes with a demonstration of the methods applied to the LSAC example, before summarising the key findings, strengths, limitations and practical recommendations.

2 Defining the causal effect using medians

Let us consider an observational study with continuous skewed outcome variable $Y$, a binary 0/1 exposure variable $A$, and a vector of $K$ confounder variables $C$. We assume that the confounder vector includes only binary or continuous variables, noting that categorical confounders can be represented as a set of binary indicators (e.g., a categorical variable with $Q$ levels would have $Q - 1$ indicators). For simplicity of discussion, we have restricted the exposure to be binary and assume that no variable is subject to missingness.

The example used throughout this paper involves data from 4882 children from a longitudinal cohort study, the LSAC kindergarten cohort. Children aged 4-5 years were recruited in 2004 (wave 1 of LSAC; approved by the Australian Institute of Family Studies Ethics Committee), with follow-ups every two years in subsequent waves. The example investigation examined the impact of maternal mental health on a child’s behaviour in early childhood. The exposure variable ($A$) was a binary indicator variable of probable serious maternal mental illness at wave 1 ($A = 1$ if present, $A = 0$ otherwise), with the outcome variable ($Y$) being the child’s behavioural difficulties as measured by the Strengths and Difficulties Questionnaire (SDQ) completed at wave 3. For SDQ scores, higher values indicate increased behavioural difficulties, with the distribution of the SDQ scores in the LSAC cohort being positively skewed (see supplementary material Figure S1). Potential confounders ($C$) were measured at wave 1 and included demographic information about the child and mother (see Table 1 for a full description of all variables used in this example).

We define $Y^{A=a}$ to be the potential outcome when the exposure is set to level $a$, where $a \in \{0, 1\}$. In the LSAC example, $Y^{a}$ would represent the SDQ score for a child when their mother is set to have a probable serious mental illness ($a = 1$) versus not ($a = 0$). Here we define the shorthand notation $m_a$ to denote the median counterfactual outcome under exposure level $A = a$, such that $m_a = m[Y^{a}]$ and $m_a \in \mathbb{R}$. The median counterfactual outcome under exposure level $A = a$ is defined as the solution to

$$\int_{-\infty}^{m_a} f_{Y^{a}}(y) = 0.5,$$

where $f_{Y^{a}}(y)$ is the density function. In other words, the median $m_a$ is the 50th centile of the cumulative distribution function (CDF) of $Y^{A=a}$, denoted as $F_{Y^{a}}(y) = P(Y^{a} \leq y)$, where $m_a = F_{Y^{a}}^{-1}(0.5)$ under the assumption that the CDF is continuous and strictly increasing for $a \in \{0, 1\}$.\footnote{5}
Table 1: Overview of variables from the Longitudinal Study of Australian Children (LSAC) example [11, 12] examining the impact of maternal mental health on a child’s behavioural difficulties in early childhood. The exposure and confounders are recorded at wave 1 (2004), with the outcome variable recorded at wave 3 (2008).

| Role        | Variable                                           | Values and additional details                                                                 |
|-------------|----------------------------------------------------|-----------------------------------------------------------------------------------------------|
| Outcome     | Behavioural difficulties score                     | Range 0-40; Strengths & Difficulties Questionnaire [6]                                         |
| Exposure    | Probable serious maternal mental illness           | Yes/No; Yes defined as a K10 score < 4 [13, 14]                                               |
| Confounders | Sex of child                                       | Male/Female                                                                                    |
|             | Whether the child has siblings                     | Yes/No                                                                                         |
|             | Child’s physical functioning score                 | Range 0-100; Pediatric Quality of Life Inventory [4]                                           |
|             | Behavioural difficulties score (baseline)          | Range 0-40; Strengths & Difficulties Questionnaire [6]                                         |
|             | Maternal age                                       | Recorded in years                                                                               |
|             | Maternal smoking status                            | Yes/No                                                                                         |
|             | Maternal risky alcohol consumption                 | Yes/No; Yes defined as > 2 standard alcoholic drinks per day                                    |
|             | Maternal completion of high school                 | Yes/No                                                                                         |
|             | Family financial hardship score                    | Range 0-6                                                                                      |
|             | Consistent parenting score                         | Range 1-5                                                                                      |

Therefore, the causal effect (CE) of interest, denoted by $\delta$, is defined as the difference between the median potential outcomes under the two exposure levels:

$$\delta = F_{Y_{a=1}}^{-1}(0.5) - F_{Y_{a=0}}^{-1}(0.5) = m[Y^{a=1}] - m[Y^{a=0}] = m_1 - m_0. \quad (2)$$

For the LSAC example, $\delta$ represents the difference in median SDQ scores if all children were exposed to maternal mental health problems compared to if none of them were exposed. Here it is useful to note that the median of the difference in individual potential outcomes is not equivalent to the difference in median potential outcomes, i.e., $m[Y^{a=1} - Y^{a=0}] \neq m[Y^{a=1}] - m[Y^{a=0}]$. This also holds for other functionals (e.g., risk ratio) as has been noted by Dawid (2000) [15], reiterating the rationale for defining causal effects as contrasts of functionals of the population counterfactual distribution across exposure groups. That is, defining them as effects in a target trial as opposed to defining them as functions of individual causal effects, which are generally not identifiable, even in a trial [16, 2, 3].

The causal difference in medians is identifiable from observational data under the following assumptions as has been shown elsewhere [17, 9]. Firstly we require the consistency assumption, which states that the exposure $A$ corresponds to a well-defined intervention that in turn corresponds to the versions of the exposure in the data [1]. Under these conditions, the potential outcome $Y^a$ is well-defined and would be equal to $Y$ if an individual received exposure level $A = a$ (assumption 1). Secondly, we require the conditional exchangeability assumption given the selected set of confounders, which states that the potential outcome $Y^a$ is independent of the received exposure $A$. 

6
given \( C \), i.e., \( Y^a \perp\!
olimits_{A|C} \) (assumption 2). Thirdly, we require the positivity assumption, which states that every individual in the population has a positive probability of being exposed or unexposed, that is \( P(A = a|C = c) > 0 \), for all \( c \) with positive probability of occurring (assumption 3). Whether these assumptions hold in practice is a matter of debate, however, for the remainder of this paper we assume these conditions do hold.

## 3 Confounding-adjustment methods

Under the aforementioned assumptions, the causal difference in medians \( \delta \) can in principle be estimated from observational data using methods that adjust for the confounder variables. Within this section we will introduce the confounding-adjustment methods investigated in our study.

### 3.1 Multivariable quantile regression

In the presence of confounding, multivariable regression is a common approach to estimate the average causal effect (i.e., defined as the difference in expected potential outcomes in the case of continuous outcomes) under the assumption of a constant causal effect across confounder strata. This approach adjusts for confounding through conditioning on the confounders. To estimate the causal difference in medians \( \delta \), a natural adaptation of this approach is to use quantile regression (QR); a common method for modelling the quantiles of the distribution of a random variable conditional on a set of covariates.\(^{[18]}\) Within this approach, a QR model is fitted for the outcome variable \( Y \) conditional on both the exposure \( A \) and \( K \) confounder variables \( C \). For example, under a main-effects QR model, the \( \tau^{th} \) quantile of \( Y \) is defined as

\[
Q_{\tau}(Y) = \beta_0(\tau) + \beta_1(\tau)A + \beta_3^T(\tau)C,
\]

where \( \beta_3(\tau) \) is a vector of length \( K \). By setting \( \tau = 0.5 \), the coefficient of the exposure variable \( \beta_1(0.5) \) encodes the difference in the conditional median outcome between exposure groups for every substrata, i.e., \( m[Y|A = 1, C = c] - m[Y|A = 0, C = c] \). Under assumptions (1-3), the assumption of a constant causal effect across confounder strata and that the regression model is correctly specified, the estimated exposure coefficient \( \hat{\beta}_1(0.5) \) has been shown to be a consistent estimator for the causal effect \( \delta \).\(^{[9]}\)

Quantile regression is a widely-applied method in practice and has implementation readily available in statistical software (e.g., the \texttt{quantreg} package in R\(^{[19]}\); the \texttt{qreg} function in STATA) making it an accessible approach. It is important to note, however, that the assumption of a constant causal effect across confounder strata may be too simplistic. Therefore less-restrictive confounding-adjustment methods may be required in practice, in particular g-methods.
3.2 IPW estimator

One family of g-methods are those using the framework of inverse probability weighting (IPW).\cite{20, 21} The motivation behind the IPW framework is to create a pseudo-population in which the distribution of covariates $C$ are balanced between exposure groups, such that the association between $A$ and $Y$ in the pseudo-population provides an unbiased estimate for the causal effect of $A$ on $Y$. To create the pseudo-population, observations are re-weighted in a way that is inversely proportional to the probability of their observed exposure group conditional on the confounding variables. These probabilities are estimated from a model for the propensity score (PS) defined as $\pi(c) = P(A = 1|C = c)$ (e.g., a logistic regression model for $A$ conditional on $C$).

A study by Zhang et al. (2012) derived an IPW estimator for the causal effect defined as the difference under exposure and no exposure in a given quantile of the counterfactual outcome distribution, in particular the median which is our focus here.\cite{8} The method can be used to estimate $\delta$ as follows. Given assumptions (1-3) hold, then for a given exposure level $a$ the cumulative distribution function is equal to\cite{8}

$$F_{Y_a}(y) = E\left[ \frac{I(A = a)I(Y \leq y)}{P(A = a|C)} \right]. \quad (4)$$

Under each exposure level $a \in \{0, 1\}$ and given estimates of the denominator probabilities from the fitted PS model, the expectation can be estimated using the sample average\cite{20, 22} and can be regarded as the weighted empirical distribution of $Y$.\cite{8} The IPW estimator of the median outcome value $m_a$ is therefore defined as the solution to the equation

$$\sum_{i=1}^{n} \hat{W}_{a,i} I(Y_i \leq y) = 0.5, \quad (5)$$

where $\hat{W}_{a,i}$ denotes the estimated weight for observation $i = 1, \ldots, n$ under exposure level $A_i = a$. The weights are defined as $I(A_i = a)/[nP(A_i = a|\mathbf{c}_i)]$ and calculated using estimates of the propensity score $\hat{\pi}(\mathbf{c})$. It is important to note that Equation 5 may not have a unique solution, and therefore $m_a$ can be estimated by the value which minimises the difference between the two sides.\cite{8} Following the suggestion of Zhang et al. (2012), normalised weights $\hat{W}_{a,i}^*$ are preferred to improve finite-sample performance\cite{8}, and are calculated by dividing each weight by the sum of all weights in the associated exposure group, such that

$$\hat{W}_{a,i}^* = \frac{I(A_i = a)}{P(A_i = a|\mathbf{c}_i)} \left/ \sum_{i=1}^{n} \frac{I(A_i = a)}{P(A_i = a|\mathbf{c}_i)} \right., \quad (6)$$

with $\hat{W}_{a,i}^*$ replacing $\hat{W}_{a,i}$ in Equation 5 above. To obtain the estimates of the median values, $\hat{m}_1$ and $\hat{m}_0$, Equation 5 is solved for each exposure level $a$. Under assumptions (1-3) and assuming that the
propensity score model is correctly specified, the difference between these two values consistently estimates the causal difference in medians $\delta$.

### 3.3 Weighted quantile regression

An alternative implementation of the IPW approach uses the IP weights to fit an appropriate weighted regression model, that is, weight the score equations of the regression as opposed to the observed outcomes. This can be applied using a univariable QR model such that the $\tau^{th}$ quantile is obtained as

$$Q_\tau(Y) = \beta_0^\ast(\tau) + \beta_1^\ast(\tau)A,$$

and the coefficient $\beta_1^\ast(\tau)$ is estimated by solving a weighted sum of check functions.[9] The weighting is implemented using the same estimated IP weights as outlined for the previous IPW estimator. By setting $\tau = 0.5$, the coefficient of the exposure variable encodes the difference in medians between each exposure level in the pseudo-population. Under assumptions (1-3) and assuming the propensity score model is correctly specified, the estimate $\hat{\beta}_1^\ast(0.5)$ has been shown to be a consistent estimator for the causal difference in medians $\delta$.[9]

Unlike the previous IPW estimator, which needs to be hand-coded at present, implementation of weighted QR is readily available, with many of the aforementioned software packages enabling specification of weights within the QR implementation (e.g., the `weights` argument within the `rq` function in R[19, 23]).

### 3.4 G-computation

G-computation is another popular method to adjust for confounding when estimating average causal effects.[24] The method arises from the g-formula which states that under assumptions (1-3), the marginal density of $Y^a$ using observed data can be identified as

$$f_{Y^a}(y) = \sum_c f_{Y|A,C}(y|a,c)f_C(c),$$

which is equivalent to $\mathbb{E}[f_{Y|A,C}(y|a,c)]$ where the outer expectation is over $C$. This result can be used to identify any functional of the marginal density of $Y^a$, such as the expected or median potential outcome. Intuitively, the right-hand side is standardising the conditional density under an exposure value to the distribution of the confounders in the whole sample, thus addressing the imbalance in outcome predictors between exposure groups due to non-randomisation and rendering the contrast under the two exposure values comparable.
When the expected potential outcome is of direct interest, the above expression implies that this parameter is identified as the expectation of the conditional expectations across the sample, and thus it is straightforward to estimate by plugging in estimates of the conditional expectations in the g-formula. In practice, the conditional expectations are estimated parametrically by fitting a regression model for the outcome variable $Y$ conditional on both the exposure $A$ and confounder variables $C$. The fitted model is used to obtain individual predictions of the mean outcome $Y$ under exposure level $A = a$ for each record given the confounder values, with the mean of these providing an estimate of the expected potential outcome under that exposure level.

When interest is in the median potential outcome $m_a = m[Y^a]$, it is noted that the g-formula (Equation 8) implies that Equation 1 can be rewritten as

$$\int_{-\infty}^{m_a} \mathbb{E}[f_{Y|A,C}(y|a, c)] dy = 0.5$$

(9)

Here we note that the term inside the integral is the expectation of a density, and thus the median potential outcome is not identified by something as simple as, say, the median of the conditional medians or conditional expectations across the sample. Instead, estimation of the median potential outcome using g-computation requires estimation of the conditional density within the integral. Next we describe two possible implementations of g-computation to estimate $m_a$ based on the above expression: a Monte Carlo integration-based approach and an approximate approach, denoted as $g$-comp (MC) and $g$-comp (approx), respectively.

3.4.1 G-comp (MC)

The first implementation, g-comp (MC), uses Monte Carlo simulation to perform draws from the conditional density of $Y^a$, $f_{Y^a}(y)$, which can then be used to estimate the median of $Y^a$. Specifically, we posit a model for the conditional density of $Y$ given $A = a$ and $C$ and then perform a fixed, large number of random draws from it for each record which by Equation 8 corresponds to draws from $f_{Y^a}(y)$, or equivalently, the inner expectation in Equation 9. The median potential outcome under exposure value $a$ is estimated as the median of these draws.

Implementing this approach to estimate $m_a$ requires an assumption about the distribution of $Y$ conditional on $A$ and $C$, such that a suitable model can be specified (referred to as the outcome model). Given the skewed nature of the outcome in this setting, one possible approach, which we consider in our simulations and case study, is to assume that $Y \sim A, C$ follows an approximate log normal distribution, with the mean of the underlying normal distribution dependent on $A$ and $C$. The MC g-computation method would therefore be implemented as follows:

1. Fit a linear model for log($Y$) conditional on $A$ and $C$ to the observed data.
2. Using the fitted linear model, obtain predictions of the mean outcome (on the log scale) for every observation (where \( i = 1, \ldots, n \)) twice:

(a) Set each observation to be exposed (i.e., set \( A_i = 1 \)) to obtain predictions \( \hat{\mu}_1^i = \hat{E}[\log(Y_i)|A_i = 1, C_i = C] \), \( i = 1, \ldots, n \).

(b) Set each observation to be unexposed (i.e., set \( A_i = 0 \)) to obtain predictions \( \hat{\mu}_0^i = \hat{E}[\log(Y_i)|A_i = 0, C_i = C] \), \( i = 1, \ldots, n \).

3. For \( a = 1, 0 \), repeatedly perform \( R \) draws (say \( R = 1000 \)) for every observation from a log normal distribution parametrised with the mean (on the log scale) equal to \( \hat{\mu}_a^i \) and standard deviation \( \hat{\sigma} \) equal to the estimate obtained via the residual deviance of the model fitted in step 1.

4. For \( a = 0, 1 \), the median of the combined \( R \) samples drawn for each of the \( n \) observations are used to obtain an estimate of the median potential outcome \( \hat{m}_a \) under \( A = a \).

3.4.2 G-comp (approx)

In practice, the above simulation approach may be computationally intensive, particularly for a data set with a large number of observations. Therefore, an alternative approach, g-comp (approx), is to approximate \( \hat{E}[f_{Y|A,C}(y|a,c)] \) by obtaining estimates of it across a grid of candidate \( y \) values, denoted as \( y^* \), and then solve Equation 9 numerically to obtain the estimated median potential outcome. This approach again requires a model for the density, which as with the previous implementation could be based on an assumption of log-normality. Thus a linear regression model for \( \log(Y) \) conditional on \( A \) and \( C \) would again be fitted to the observed data and used to obtain predictions \( \hat{\mu}_1^i, \hat{\mu}_0^i \) and an estimate of \( \hat{\sigma} \) as before (steps 1 and 2 outlined above). For \( a = 0, 1 \) and for each candidate \( y^* \), we then estimate \( \hat{f}_{Y|A,C}(y^*|a,c_i) \) for each record \( i \) by assuming a log normal density with the mean (on the log scale) equal to \( \hat{\mu}_a^i \) and standard deviation \( \hat{\sigma} \). Averaging the conditional densities over the sample yields the estimated expectation for candidate value \( y^* \):

\[
\hat{E}[f_{Y|A,C}(y^*|a,c)] = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{Y|A,C}(y^*|a,c_i).
\]  

By repeating this process for every candidate value \( y^* \), the expectation within the integral is estimated across the range of \( Y \). Equation 11 is then solved by approximating the integral by adding up these values cumulatively, and finding the minimum value of \( y^* \) for which this sum is equal to 0.5. This yields the estimate \( \hat{m}_a \), with the process repeated under \( a = 1 \) and \( a = 0 \).

For both g-computation implementations, the difference between the estimates \( \hat{m}_1 \) and \( \hat{m}_0 \) consistently estimate the causal difference in medians \( \delta \) under assumptions (1-3) and assuming that
the outcome model (i.e., the model for the conditional density of \( Y \) given \( A \) and \( C \)) is correctly specified.

4 Simulation Study

A simulation study was conducted motivated by the LSAC example, to investigate the performance of each confounding-adjustment method in a realistic setting and under varying degrees of skewness in the outcome variable.

4.1 Design of the simulation study

We explored four different skewness scenarios (i.e., different skewed distributions for \( Y \)), with the simulated datasets generated as detailed below. A full outline of each variable and their generating distribution is provided in the supplementary material (Table S1).

We generated 1000 samples consisting of 1000 observations for each of the four skewness scenarios. For each scenario, dataset and record, five confounder variables \( C_k \) for \( k = 1, \ldots, 5 \) (three binary and two continuous) were generated sequentially in that order based on variables in the LSAC data set. Observations for the binary confounders \( C_1, C_3 \) and \( C_4 \) (based on sex, maternal education and financial hardship, respectively) were generated from a binomial distribution with the success probability defined by a logistic regression model using all previous confounders as predictors. For the continuous confounders \( C_2 \) and \( C_5 \) (maternal age and log-transformed baseline SDQ, respectively) a similar approach was used; values were generated from normal distributions with the means defined by a linear regression model including all previous confounders as predictors.

Secondly, a binary exposure \( A \) was generated for each record from a binomial distribution with success probability defined by a logistic regression model including all confounders as main effects. A skewed continuous outcome \( Y \) was then generated for each record. Values for \( \log(Y) \) were generated from a normal distribution with the mean defined by a linear regression model including the exposure and all confounders as predictors. The linear regression model was specified to include all main effects and two exposure-confounder interaction terms based on interactions observed in the LSAC example. Different skewed distributions in the outcome variable were established by setting the standard deviation in the generating normal distribution for \( \log(Y) \) to \( \sigma = 0.75, 1, 1.25, 1.5 \) for each of the four increasing skewness scenarios, respectively. Values for \( \log(Y) \) were then exponentiated to obtain the outcome value \( Y \), with the corresponding distribution of \( Y \) being positively skewed (see Figure S1 in the supplementary material for a visualisation of these distributions). Here we note that \( \sigma \) characterises other properties of the outcome distribution beyond the measure of skewness.
Therefore, findings across skewness scenarios are not expected to behave in a monotonic pattern, for example, increasing bias with increasing skewness.

Aside from varying skewness, a further consideration of this study was the strength of confounding bias in the simulated datasets and whether this impacted the methods’ performance. Therefore the data generation process was conducted under two different settings - weak confounding bias (approximately 10% relative bias in the unadjusted estimate relative to the true value) and strong confounding bias (approximately 20% relative bias). Modifications to coefficients in the outcome-generating model were used to achieve this.

The true causal difference in medians $\delta$ under this data generation set-up was computed by empirical methods\cite{9}, with further details about implementation and the resulting true values provided in the supplementary material (Section \text{S2.3} and Table \text{S2}). Coefficients in the outcome-generating model were modified to ensure the true value was large enough to be estimable with the given sample size with adequate power (approximately 80%) in an unadjusted analysis.

The confounding-adjustment methods outlined in Section 3 (alongside an unadjusted approach contrasting the medians across exposure groups in the observed data) were applied to each simulated dataset to estimate the causal difference in medians $\delta$. For the multivariable quantile regression approach, the outcome model included $A$ and $C$ as predictors with main effect terms only, as is the default in most software. For both the IPW estimator and the weighted quantile regression approaches, the propensity score model regressed $A$ on the confounders including main effect terms only. For the g-computation approaches, the outcome model was specified as $\log(Y)$ conditional on $A$ and $C$, including main effect terms and two exposure-confounder interaction terms. Here we note that both the propensity score model and the outcome model were correctly specified (i.e., consistent with the data generation approach). For g-comp (MC), we performed $R = 1000$ draws per observation. For g-comp (approx), the candidate $y^*$ values ranged over the values $[0.01, 8]$ in increments of 0.01.

Based on previous recommendations\cite{8} and for ease of implementation and comparability, standard errors and confidence intervals (CIs) were estimated using bootstrap procedures (using the percentile method for CIs) for all methods. For each confounding strength and skewness scenario, metrics assessing the performance of each method (e.g., bias, relative bias, coverage, empirical and model-based standard errors for $\hat{\delta}$) were calculated using the formulae in Morris et al. (2019)\cite{25} and averaged over the 1000 samples, with key focus on bias in the point estimate. Monte Carlo standard errors were estimated for each metric. All analysis was conducted in R 4.0.2\cite{23}, using the quantreg\cite{19} and boot\cite{26,27} packages within self-developed code (available at https://github.com/daisyshep/CI-medians.git), with the IPW estimator implemented using
4.2 Results from the simulation study

Within the simulation study, estimates for $\delta$ differed depending on the confounding-adjustment method used for analysis, with the variation in estimates tending to increase with increasing skewness for both confounding strengths (Figure 1). Estimates obtained using quantile regression (QR) were biased across all skewness scenarios (relative bias range: 6.6% to 8.1% weak confounding, 9.0% to 14.4% strong confounding; Table 2); an expected result given the method’s strict assumption of a constant causal effect across stratum which did not hold in the data generating mechanism. In contrast, methods which relaxed this assumption (the IPW estimator, weighted quantile regression and both implementations of g-computation) performed well across both confounding strength settings, with minimal bias in estimates for $\delta$ (relative bias $< 5\%$ in the majority of skewness scenarios).

The IPW estimator and weighted quantile regression consistently had similar performance to one another, yielding similar results across all skewness scenarios and both confounding settings. In addition, both g-comp (MC) and g-comp (approx) had similar performance to one another across all simulation scenarios (i.e., relative bias differing by $<0.1\%$ between the methods), with implementation of the latter being substantially quicker computationally. Here we also note that both implementations of g-computation consistently yielded lower empirical standard errors comparatively to all other confounding-adjustment methods across all skewness and confounding settings, therefore minimising both the bias and variance simultaneously.

There was little bias in estimating the model standard error (SE) with the bootstrap for all methods, although in a number of settings the IPW estimator and weighted QR overestimated the model SE by a larger margin comparatively to other approaches. The coverage probability was close to nominal level for all confounding-adjustment methods (range: 93.60% to 97.70%) across all skewness scenarios and both confounding settings. Monte Carlo SEs of performance estimates were all acceptable (supplementary material Table S3), and therefore we feel comfortable drawing conclusions about the methods based on the 1000 samples used per skewness scenario.

5 Application to LSAC

Methods were applied to the LSAC data set using complete cases only ($n = 3245$). For the multivariable quantile regression approach, the outcome model included $A$ and $C$ (consisting of 9 confounders; Table 1) as predictors with main effect terms only. The propensity score model (used for the IPW estimator and weighted quantile regression) regressed $A$ on $C$ including main effect terms only. For both g-comp (MC) and g-comp (approx), the outcome model was specified as $\log(Y)$
Table 2: Performance of confounding-adjustment methods across confounding and skewness scenarios in the simulation study with maximum Monte Carlo standard errors (SE) provided in the table footnote (see supplementary material Table S3 for all Monte Carlo SEs).

| Confounding Skewness scenario | Method            | Bias   | Relative bias (%) | Empirical SE | Model SE | Error SE (%) | Coverage (%) |
|-------------------------------|-------------------|--------|-------------------|--------------|----------|--------------|--------------|
| Weak 1                        | Unadjusted        | 0.090  | 10.07             | 0.369        | 0.381    | 3.21         | 94.60        |
|                               | QR                | 0.059  | 6.63              | 0.384        | 0.375    | -2.24        | 94.80        |
|                               | IPW estimator     | -0.004 | -0.48             | 0.442        | 0.454    | 2.78         | 94.90        |
|                               | Weighted QR      | -0.023 | -2.53             | 0.440        | 0.453    | 2.98         | 95.20        |
|                               | G-comp (MC)       | -0.026 | -2.91             | 0.308        | 0.298    | -3.18        | 93.70        |
|                               | G-comp (approx)   | -0.026 | -2.90             | 0.308        | 0.298    | -3.16        | 93.60        |
| Weak 2                        | Unadjusted        | 0.123  | 10.08             | 0.526        | 0.528    | 0.32         | 94.90        |
|                               | QR                | 0.099  | 8.13              | 0.540        | 0.517    | -4.35        | 94.20        |
|                               | IPW estimator     | 0.050  | 4.09              | 0.625        | 0.640    | 2.43         | 93.70        |
|                               | Weighted QR      | 0.023  | 1.91              | 0.621        | 0.636    | 2.54         | 94.20        |
|                               | G-comp (MC)       | 0.006  | 0.48              | 0.436        | 0.420    | -3.60        | 94.00        |
|                               | G-comp (approx)   | 0.006  | 0.45              | 0.436        | 0.420    | -3.62        | 93.90        |
| Weak 3                        | Unadjusted        | 0.160  | 10.00             | 0.684        | 0.727    | 6.27         | 95.60        |
|                               | QR                | 0.114  | 7.14              | 0.693        | 0.695    | 0.31         | 94.70        |
|                               | IPW estimator     | 0.019  | 1.17              | 0.806        | 0.870    | 7.94         | 95.00        |
|                               | Weighted QR      | -0.020 | -1.23             | 0.803        | 0.864    | 7.62         | 95.20        |
|                               | G-comp (MC)       | -0.016 | -1.00             | 0.550        | 0.563    | 2.47         | 95.20        |
|                               | G-comp (approx)   | -0.015 | -0.97             | 0.550        | 0.563    | 2.40         | 95.10        |
| Weak 4                        | Unadjusted        | 0.193  | 10.09             | 0.833        | 0.894    | 7.24         | 96.80        |
|                               | QR                | 0.133  | 6.95              | 0.864        | 0.842    | -2.62        | 95.50        |
|                               | IPW estimator     | 0.123  | 6.45              | 1.032        | 1.098    | 6.40         | 95.30        |
|                               | Weighted QR      | 0.081  | 4.22              | 1.017        | 1.091    | 7.24         | 95.50        |
|                               | G-comp (MC)       | 0.062  | 3.25              | 0.695        | 0.712    | 2.40         | 95.00        |
|                               | G-comp (approx)   | 0.063  | 3.28              | 0.694        | 0.710    | 2.20         | 94.90        |
| Strong 1                      | Unadjusted        | 0.171  | 10.17             | 0.391        | 0.398    | 1.66         | 93.20        |
|                               | QR                | 0.098  | 11.58             | 0.398        | 0.390    | -2.10        | 94.30        |
|                               | IPW estimator     | 0.003  | 0.31              | 0.441        | 0.466    | 5.75         | 95.40        |
|                               | Weighted QR      | -0.017 | -1.95             | 0.440        | 0.464    | 3.33         | 95.30        |
|                               | G-comp (MC)       | -0.023 | -2.65             | 0.308        | 0.306    | -0.93        | 94.10        |
|                               | G-comp (approx)   | -0.023 | -2.67             | 0.308        | 0.306    | -0.86        | 94.10        |
| Strong 2                      | Unadjusted        | 0.240  | 20.05             | 0.544        | 0.568    | 4.40         | 94.00        |
|                               | QR                | 0.117  | 9.81              | 0.556        | 0.556    | -0.03        | 95.00        |
|                               | IPW estimator     | 0.030  | 2.55              | 0.621        | 0.673    | 8.34         | 96.00        |
|                               | Weighted QR      | 0.004  | 0.33              | 0.617        | 0.670    | 8.61         | 95.40        |
|                               | G-comp (MC)       | -0.025 | -2.08             | 0.446        | 0.441    | -1.22        | 94.30        |
|                               | G-comp (approx)   | -0.025 | -2.10             | 0.446        | 0.440    | -1.22        | 94.00        |
| Strong 3                      | Unadjusted        | 0.307  | 20.11             | 0.729        | 0.754    | 0.72         | 93.50        |
|                               | QR                | 0.219  | 14.39             | 0.726        | 0.709    | -2.32        | 93.70        |
|                               | IPW estimator     | 0.117  | 7.68              | 0.847        | 0.897    | 5.89         | 95.40        |
|                               | Weighted QR      | 0.082  | 5.35              | 0.843        | 0.891    | 5.77         | 95.50        |
|                               | G-comp (MC)       | 0.076  | 4.97              | 0.570        | 0.579    | 1.62         | 95.10        |
|                               | G-comp (approx)   | 0.075  | 4.93              | 0.569        | 0.579    | 1.75         | 95.20        |
| Strong 4                      | Unadjusted        | 0.420  | 20.02             | 1.004        | 1.052    | 4.76         | 95.10        |
|                               | QR                | 0.189  | 8.98              | 1.002        | 0.993    | -0.88        | 94.50        |
|                               | IPW estimator     | 0.038  | 1.81              | 1.189        | 1.252    | 5.37         | 95.50        |
|                               | Weighted QR      | -0.009 | -0.43             | 1.180        | 1.241    | 5.19         | 95.50        |
|                               | G-comp (MC)       | -0.028 | -1.34             | 0.759        | 0.806    | 6.11         | 95.40        |
|                               | G-comp (approx)   | -0.028 | -1.32             | 0.760        | 0.791    | 4.12         | 95.70        |

Maximum Monte Carlo SE (performance measure): 0.038 (bias), 0.018 (relative bias), 0.010 (empirical SE), 0.029 (model SE), 6.956% (relative error in model SE), 0.796% (coverage).
conditional on \( A \) and \( C \), including two-way interaction terms between \( A \) and three confounders (child’s sex, maternal completion of high school, consistent parenting) deemed plausible based on substantive knowledge. For g-comp (MC), we performed \( R = 1000 \) draws per observation. For g-comp (approx), the vector of candidate \( y^* \) values ranged over the values \([0.01, 18]\) in increments of 0.01.

Applying the confounding-adjustment methods to the LSAC example yielded a lower estimate of the causal difference in medians for all methods comparatively to the unadjusted approach (Figure 2). All methods estimated the median SDQ score for children exposed to maternal mental health problems to be higher than the median SDQ score for children if they were not exposed, suggesting moderately increased behavioural problems in early childhood for children of mothers with mental illness. Estimated effects were more consistent across the methods than observed within the simulation study, although the quantile regression approach yielded lower estimates of \( \delta \) than the other methods. Notably both the IPW estimator and weighted quantile regression produced the same point estimates and bootstrap confidence intervals as one another when applied to the LSAC example. Additionally, both implementations of the g-computation approach produced similar estimates to one another, and estimated a slightly smaller causal effect than the IPW estimator and weighted quantile regression.
Figure 2: Estimated causal difference in medians $\hat{\delta}$ for the Longitudinal Study Of Australian Children (LSAC) example\textsuperscript{11, 12} obtained under each confounding-adjustment method, where $\hat{\delta} = \hat{m}_1 - \hat{m}_0$ is the estimated difference in median SDQ scores if all children were exposed to maternal mental health problems compared to if none of them were exposed. Point estimates and their corresponding 95% confidence intervals are presented alongside the figure.

6 Discussion

In the presence of skewed outcome data, the common approach to “ignore or transform” may not be optimal, and defining the causal effect using a contrast of median potential outcomes may be more appropriate. Despite being a widely acknowledged concept, there is scarce availability and awareness of confounding-adjustment methods to estimate the causal difference in medians in practice. A handful of previous studies have proposed approaches to estimate causal effects as contrasts of distribution quantiles more generally\textsuperscript{9, 8}, but investigation of these methods and their application in health and medical studies remain scarce.

In this paper we aimed to address this gap, by describing and evaluating methods identified from previous literature (multivariable quantile regression\textsuperscript{9}, an IPW estimator\textsuperscript{8} and weighted quantile regression\textsuperscript{9}) alongside two implementations of g-computation that, to the best of our knowledge, have not been widely described or studied alongside other methods. The confounding-adjustment methods were selected and described with a key focus on their accessibility and ease of implementation in a practical setting, with code made available, to encourage their use in practice where applicable.

Results from the simulation study indicated varied performance of the confounding-adjustment methods when estimating the causal difference in medians. As anticipated, the multivariable quan-
tile regression was too simplistic for the realistic setting reflected in our simulated datasets, where the causal effect is not constant across strata and therefore resulted in biased estimates for the causal difference in medians. The IPW estimator, weighted quantile regression and both implementations of g-computation yielded estimates with minimal bias, with g-computation additionally minimising the variance in estimates; an expected observation as IPW estimates tend to be more variable than those obtained via g-computation. We also note that the g-comp (approx) implementation was computationally more efficient than g-comp (MC) provided a suitable $y^*$ range is used to ensure an accurate estimate can be obtained.

These findings need to be interpreted in light of the fact that under our data generation approach, both the propensity score model (used for the IPW estimator and weighted quantile regression) and outcome model (used for g-computation) were correctly specified. Correct specification of these models is a critical assumption when applying the singly robust methods as noted by Zhang et al. (2012) in the case of IPW. However, in the context of estimating the causal difference in medians, the outcome model for g-computation relates to specifying a model for the whole density. In practice, correctly specifying this model may be harder to achieve than a correctly specified propensity score model, and thus could be considered a stronger assumption than for the weighted methods (IPW estimator and weighted quantile regression).

A strength of this work was the design of our simulation study which was motivated by the LSAC example, allowing us to investigate the performance of these methods in a realistic scenario. Further we considered a range of settings investigating varying skewness distributions in the outcome variable, alongside two different strengths of confounding bias present in the datasets, resulting in a more complex and realistic study than those explored in previous papers. Additionally, our inclusion of the g-computation approach (under two implementations) in an accessible and clear manner, has brought light to the approach; something we have not previously seen in practice and in the context of the causal difference in medians. The g-comp (approx) method in particular provided a well-performing implementation with a lower computational burden than g-comp (MC), which may be more feasible to apply in practice.

A potential limitation of our study was the restriction to singly robust methods only, which rely on correct specification of the respective model. However, previous studies have presented a handful of promising doubly robust methods. Zhang et al. (2012) derived an adaptation of the augmented inverse probability weighted estimator and a further adaptation incorporating stratification based on the propensity score, with results suggesting both estimators are preferable over the singly robust alternative if precision is of concern. A later study by Xu et al. (2018) compared the performance of the doubly robust estimators of Zhang et al. and a proposed targeted maxi-
mum likelihood estimator (TMLE) for quantiles by Díaz (2017) with a Bayesian non-parametric (BNP) approach for causal inference on quantiles. Results indicated the BNP approach performed consistently well, whilst the frequentist approaches suffered with under-coverage and large bias under some simulation scenarios, and the TMLE approach had high computation times and bias under some simulation settings. Xie et al. proposed an additional multiply robust estimator, using multiple outcome and propensity score models, and found the estimator to be less biased under model misspecification compared to the doubly robust estimators of Zhang et al. Results from these studies indicated the proposed doubly robust methods performed well, with the natural advantage being the methods’ robustness to model misspecification. However, the doubly robust methods have not yet been evaluated in a range of complex and realistic scenarios as is done here for singly robust methods, with their implementation not as accessible or readily available in software and therefore their application remains scarce in practice. Both of these factors would be useful to pursue in future work.

Within this study, we restricted our simulation design to explore a binary exposure only. Therefore, potential future directions could explore the performance of methods in the presence of a continuous, or multi-level exposure variables. In addition, the outcome distributions explored within this study were all heavily right-skewed, so results may represent an extreme representation of the methods’ performance under heavily skewed outcome data. Furthermore, we also reiterate the need for the causal estimand of interest to be driven by the research question at hand. Even in the presence of skewed outcome data, it may be appropriate or preferred to define the average causal effect as a contrast of expected potential outcomes. Therefore we caution the reader to not be purely data-driven, but to consider the usefulness of the target estimand when drawing conclusions about their study.

In conclusion, when estimating the causal difference in medians the IPW estimator, weighted quantile regression or g-computation present promising approaches, provided a richly specified model is used such that correct specification of the propensity score model or outcome model is likely. Implementation of the IPW estimator and weighted quantile regression methods is readily available and therefore accessible to implement (e.g., \{rq\} function in \texttt{R}, open-source code of the IPW estimator). Implementation of the g-computation approach (approximate or MC) is not as readily available, but we have provided source code which can guide practitioners in the implementation of this method. We also support the use of the approximate implementation of g-computation for computational efficiency provided that a suitable range of candidate \(y^*\) values is used. Overall, these methods provide appealing alternatives to the common “ignore or transform” approach or the stringent constant-effect assumption of multivariable quantile regression, enhancing
our capability to obtain meaningful causal effect estimates with skewed outcome data.

**Author contributions**

Development of the project and study design was conducted by DS and MMB. DS conducted the simulation study and data analysis, with regular input from MMB. The manuscript was developed by DS, with critical input provided by MMB. BB provided input into the g-computation methodology, and feedback on the manuscript.

**Financial disclosure**

DS was supported by the Lorenzo and Pamela Galli Medical Research Trust. MMB was supported by an Australian Research Council Discovery Early Career Researcher Award (project number DE190101326) funded by the Australian Government and a NHMRC Investigator Grant (ID 2009572).

**Data availability statement**

Data used for the illustrative example within this paper are available from the Department of Social Services (DSS) with access provided by the National Centre for Longitudinal Data (NCLD). Restrictions apply to the availability of these data, which were used under license for this study. Data are available at [https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/F2YRL5](https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/F2YRL5) with the permission of the DSS.

**Conflict of interest**

The authors declare no potential conflict of interests.

**References**

[1] M. Hernán and J. Robins, *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[2] M. Hernán and J. Robins, “Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available,” *American Journal of Epidemiology*, vol. 183, no. 8, pp. 758–764, 2016.

[3] M. Moreno-Betancur, “The target trial: A powerful device beyond well-defined interventions,” *Epidemiology*, vol. 32, no. 2, pp. 291–294, 2021.
[4] J. Varni, M. Seid, and C. Rose, “The pedsql: measurement model for the pediatric quality of life inventory,” *Medical Care*, vol. 37, pp. 126–139, 1999.

[5] M. Rosenberg, *Society and the Adolescent Self-Image*. Middletown, CT: Wesleyan University Press, 1989.

[6] R. Goodman, “Strengths and difficulties questionnaire (sdq) [database record],” 1997.

[7] M. Hernán, “A definition of causal effect for epidemiological research,” *Journal of Epidemiology & Community Health*, vol. 58, no. 4, pp. 265–271, 2004.

[8] Z. Zhang, Z. Chen, J. Troendle, and J. Zhang, “Causal inference on quantiles with an obstetric application,” *Biometrics*, vol. 68, pp. 697–706, 2012.

[9] S. Sun, E. Moodie, and J. Nešleuhová, “Causal inference for quantile treatment effects,” *Environmetrics*, p. e2668, 2021.

[10] A. Tsiatis, M. Davidian, S. Holloway, and E. Laber, *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Boca Raton: Chapman & Hall/CRC, 2020.

[11] A. Sanson, J. Nicholson, J. Ungerer, S. Zubrick, K. Wilson, J. Ainley, D. Berthelsen, D. Broom, L. Harrison, B. Rodgers, M. Sawyer, S. Silburn, L. Straxdins, G. Vimpani, and M. Wake, *Introducing the Longitudinal Study of Australian Children*. Australia: Australian Institute of Family Studies - Commonwealth of Australia, 2002.

[12] D. Christensen, M. Fahey, R. Giallo, and K. Hancock, “Longitudinal trajectories of mental health in australian children aged 4-5 to 14-15 years,” *PLOS ONE*, vol. 12, no. 11, pp. 1–20, 2017.

[13] R. Kessler, G. Andrews, L. Colpe, E. Hiripe, D. Mroczek, S. Normand, E. Walters, and A. Zaslavsky, “Short screening scales to monitor population prevalences and trends in non-specific psychological distress,” *Psychological Medicine*, vol. 32, pp. 959–956, 2002.

[14] G. Andrews and T. Slade, “Interpreting scores on the kessler psychological distress scale (k10),” *Australian and New Zealand Journal of Public Health*, pp. 494–497, 2001.

[15] A. Dawid, “Causal Inference without Counterfactuals,” *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 407–448, 2000.

[16] M. Hernán, A. Alonso, R. Logan, F. Grodstein, K. Michels, W. Willett, J. Manson, and R. J.M., “Observational studies analyzed like randomized experiments: an application to
postmenopausal hormone therapy and coronary heart disease,” *Epidemiology*, vol. 19, no. 6, pp. 766–779, 2008.

[17] S. Greenland and J. Robins, “Identifiability, Exchangeability, and Epidemiological Confounding,” *International Journal of Epidemiology*, vol. 15, no. 3, pp. 413–419, 1986.

[18] R. Koenker and G. Bassett, “Regression quantiles,” *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[19] R. Koenker, *quantreg: Quantile Regression*, 2019. R package version 5.54.

[20] D. Horvitz and D. Thompson, “A generalization of sampling without replacement from a finite universe,” *Journal of the American Statistical Association*, vol. 47, pp. 663–685, 1952.

[21] J. Robins, M. Hernán, and B. Brumback, “Marginal structural models and causal inference in epidemiology,” *Epidemiology*, vol. 11, no. 5, pp. 550–560, 2000.

[22] J. Robins, A. Rotnitzky, and L. Zhao, “Estimation of regression coefficients when some regressors are not always observed,” *Journal of the American Statistical Association*, vol. 89, pp. 846–866, 1994.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[24] J. Robins, “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect,” *Mathematical Modelling*, vol. 7, no. 9, pp. 1393–1512, 1986.

[25] T. Morris, I. White, and M. Crowther, “Using simulation studies to evaluate statistical methods,” *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, 2019.

[26] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press, 1997. ISBN 0-521-57391-2.

[27] A. Canty and B. D. Ripley, *boot: Bootstrap R (S-Plus) Functions*, 2022. R package version 1.3-28.1.

[28] M. Moreno-Betancur, J. J. Koplin, A.-L. Ponsonby, J. Lynch, and J. B. Carlin, “Measuring the impact of differences in risk factor distributions on cross-population differences in disease occurrence: a causal approach,” *International Journal of Epidemiology*, vol. 47, no. 1, pp. 217–225, 2017.
[29] Y. Xie, C. Cotton, and Y. Zhu, “Multiply robust estimation of causal quantile treatment effects,” *Statistics in Medicine*, pp. 21–14, 2020.

[30] D. Xu, M. Daniels, and A. Winterstein, “A bayesian nonparametric approach to causal inference on quantiles,” *Biometrics*, vol. 74, pp. 986–996, 2018.

[31] I. Díaz, “Efficient estimation of quantiles in missing data models,” *Journal of Statistical Planning and Inference*, vol. 190, pp. 39–51, 2004.
Confounding-adjustment methods for the difference in medians

Supplementary Material

Daisy A. Shepherd*1,2, Benjamin R. Baer3 & Margarita Moreno-Betancur1,2

1Clinical Epidemiology & Biostatistics Unit, Department of Paediatrics, The University of Melbourne, VIC, Australia
2Clinical Epidemiology & Biostatistics Unit, The Murdoch Children’s Research Institute, The Royal Children’s Hospital, VIC, Australia
3Department of Biostatistics and Computational Biology, The University of Rochester, New York, USA

*Corresponding author: Daisy A. Shepherd. Email: daisy.shepherd@mcri.edu.au

Abbreviations: LSAC, The Longitudinal Study of Australian Children; SDQ, Strengths and Difficulties Questionnaire; QR, quantile regression; WQR; weighted quantile regression; IPW, inverse probability weighted; PS, propensity score

Contents

S1 The LSAC case study
S2 Additional simulation study details
  S2.1 Data generation models
  S2.2 Distribution of the outcome variable
  S2.3 True difference in medians
  S2.4 Monte Carlo standard errors

References
S1 The LSAC case study

Figure S1: Distribution of the Strengths & Difficulties questionnaire (SDQ) scores at wave 3 for the Longitudinal Study Of Australian Children (LSAC) case study.[11, 12]

S2 Additional simulation study details

S2.1 Data generation models

Table S1: Detail on the models used to generate variables for the simulation study, with the structure based on the LSAC dataset unless otherwise specified in Section 4 of the main text.

| Variable | LSAC variable | Generating distribution | Additional details |
|----------|---------------|-------------------------|--------------------|
| C₁       | Sex           | $C_1 \sim \text{Binomial}(1, 0.51)$ |                     |
| C₂       | Maternal age  | $C_2 \sim \text{Normal}(35.17, 5.47)$ |                     |
| C₃       | Maternal education | $C_3 \sim \text{Binomial}(1, Pr(C_3 = 1))$ | $\text{logit}[Pr(C_3 = 1)] = -1.41 + 0.78C_1 + 0.04C_2$ |
| C₄       | Financial hardship$^a$ | $C_4 \sim \text{Binomial}(1, Pr(C_4 = 1))$ | $\text{logit}[Pr(C_4 = 1)] = -1.55 + 0.47C_1 + 0.03C_2 + 0.80C_3$ |
| C₅       | Baseline SDQ (logged) | $C_5 \sim \text{Normal}(\mu_{C_5}, 0.63)$ | $\mu_{C_5} = 1.91 + 0.03C_1 + 0.01C_2 + 0.05C_3 + 0.12C_4$ |
| A        | Maternal mental health | $A \sim \text{Binomial}(1, Pr(A = 1))$ | $\text{logit}[Pr(A = 1)] = -2.39 + 0.04C_1 - 0.05C_2 - 0.09C_3 + 0.51C_4 + 1.07C_5$ |
| log(Y)   | SDQ score     | $\text{log}(Y) \sim \text{Normal}(\mu_Y, \sigma)$ | $\mu_Y = 1.40 + 0.49A + 0.03C_1 - 0.01C_2 + 0.01C_3 + 0.03C_4 + 0.26C_5 + 0.12AC_1 - 0.01AC_2$ |
|          |               |                         | $\sigma = 0.75, 1, 1.25, 1.5$ |

$^a$Dichotomised: No (score = 0 on original scale), Yes (score > 0 on original scale)
S2.2 Distribution of the outcome variable

![Image](image.png)

**Figure S2:** Distribution of the outcome variable \( (Y) \) under each skewness scenario and confounding bias strength used within the simulation study.

S2.3 True difference in medians

Within the simulation study, the true difference in medians \( \delta \), with respect to which assess bias, was computed by empirical methods outlined in the supplementary material of Sun et al. (2021).\[9\] Specific details of our implementation are as follows. Initially we generated a large dataset (1,000,000 observations) per skewness scenario and strength of confounding bias (weak or strong). For each large dataset, a quantile regression model (including all main effects and exposure-confounder interactions) was fitted under different quantiles \( \tau \); we used 200 different values for \( \tau \) equally distributed across the range \([0.05, 0.95]\). For our vector of \( m_a \) support values, we covered the range \([3,8]\) in increments of 0.005. Different values of \( \delta \) were obtained for each skewness scenario and under the two different strengths of confounding bias, as outlined below in Table S2.

| Confounding bias strength | Skewness scenario 1 | 2 | 3 | 4 |
|---------------------------|---------------------|---|---|---|
| Weak                      | 0.895               | 1.220 | 1.600 | 1.910 |
| Strong                    | 0.850               | 1.195 | 1.525 | 2.100 |

**Table S2:** True values for the causal difference in medians \( \delta \) obtained under each skewness scenario and confounding setting.
S2.4 Monte Carlo standard errors

**Table S3:** Monte Carlo standard errors of performance estimates calculated over the 1000 simulated datasets per skewness scenario and confounding bias strength under each of the confounding-adjustment methods.

| Confounding scenario | Method          | Bias  | Relative Bias | Empirical SE | Model SE | Relative error SE (%) | Coverage (%) |
|----------------------|----------------|-------|---------------|--------------|----------|------------------------|--------------|
| Weak 1               | Unadjusted     | 0.012 | 0.013         | 0.008        | 0.002    | 2.387                  | 0.715        |
|                      | QR             | 0.012 | 0.014         | 0.009        | 0.002    | 2.234                  | 0.702        |
|                      | IPW estimator  | 0.014 | 0.016         | 0.010        | 0.003    | 2.414                  | 0.696        |
|                      | Weighted QR    | 0.014 | 0.016         | 0.010        | 0.003    | 2.420                  | 0.676        |
|                      | G-comp (MC)    | 0.010 | 0.011         | 0.007        | 0.001    | 2.179                  | 0.768        |
|                      | G-comp (approx)| 0.010 | 0.011         | 0.007        | 0.001    | 2.179                  | 0.774        |
| Weak 2               | Unadjusted     | 0.017 | 0.014         | 0.008        | 0.004    | 2.607                  | 0.715        |
|                      | QR             | 0.017 | 0.014         | 0.009        | 0.003    | 2.364                  | 0.702        |
|                      | IPW estimator  | 0.020 | 0.016         | 0.010        | 0.007    | 2.768                  | 0.696        |
|                      | Weighted QR    | 0.020 | 0.016         | 0.010        | 0.007    | 2.750                  | 0.676        |
|                      | G-comp (MC)    | 0.014 | 0.011         | 0.007        | 0.002    | 2.234                  | 0.768        |
|                      | G-comp (approx)| 0.014 | 0.011         | 0.007        | 0.002    | 2.234                  | 0.774        |
| Weak 3               | Unadjusted     | 0.022 | 0.014         | 0.008        | 0.009    | 3.352                  | 0.715        |
|                      | QR             | 0.022 | 0.014         | 0.009        | 0.007    | 2.783                  | 0.702        |
|                      | IPW estimator  | 0.025 | 0.016         | 0.010        | 0.013    | 3.735                  | 0.696        |
|                      | Weighted QR    | 0.025 | 0.016         | 0.010        | 0.013    | 3.744                  | 0.676        |
|                      | G-comp (MC)    | 0.017 | 0.011         | 0.007        | 0.003    | 2.426                  | 0.768        |
|                      | G-comp (approx)| 0.017 | 0.011         | 0.007        | 0.003    | 2.427                  | 0.774        |
| Weak 4               | Unadjusted     | 0.026 | 0.014         | 0.008        | 0.014    | 4.112                  | 0.715        |
|                      | QR             | 0.027 | 0.014         | 0.009        | 0.010    | 3.405                  | 0.702        |
|                      | IPW estimator  | 0.033 | 0.017         | 0.010        | 0.022    | 5.563                  | 0.696        |
|                      | Weighted QR    | 0.032 | 0.017         | 0.010        | 0.022    | 5.590                  | 0.676        |
|                      | G-comp (MC)    | 0.022 | 0.012         | 0.007        | 0.006    | 2.947                  | 0.768        |
|                      | G-comp (approx)| 0.022 | 0.012         | 0.007        | 0.006    | 2.872                  | 0.774        |
| Strong 1             | Unadjusted     | 0.012 | 0.015         | 0.009        | 0.002    | 2.357                  | 0.796        |
|                      | QR             | 0.013 | 0.015         | 0.009        | 0.002    | 2.242                  | 0.733        |
|                      | IPW estimator  | 0.014 | 0.016         | 0.010        | 0.003    | 2.491                  | 0.662        |
|                      | Weighted QR    | 0.014 | 0.016         | 0.010        | 0.003    | 2.480                  | 0.669        |
|                      | G-comp (MC)    | 0.010 | 0.011         | 0.007        | 0.001    | 2.228                  | 0.745        |
|                      | G-comp (approx)| 0.010 | 0.011         | 0.007        | 0.001    | 2.230                  | 0.745        |
| Strong 2             | Unadjusted     | 0.017 | 0.014         | 0.009        | 0.005    | 2.582                  | 0.796        |
|                      | QR             | 0.018 | 0.015         | 0.009        | 0.004    | 2.397                  | 0.733        |
|                      | IPW estimator  | 0.020 | 0.016         | 0.010        | 0.007    | 2.879                  | 0.662        |
|                      | Weighted QR    | 0.020 | 0.016         | 0.010        | 0.007    | 2.869                  | 0.669        |
|                      | G-comp (MC)    | 0.014 | 0.012         | 0.007        | 0.002    | 2.289                  | 0.745        |
|                      | G-comp (approx)| 0.014 | 0.012         | 0.007        | 0.002    | 2.291                  | 0.745        |
| Strong 3             | Unadjusted     | 0.023 | 0.015         | 0.009        | 0.008    | 3.106                  | 0.796        |
|                      | QR             | 0.023 | 0.015         | 0.009        | 0.007    | 2.734                  | 0.733        |
|                      | IPW estimator  | 0.027 | 0.018         | 0.010        | 0.017    | 4.559                  | 0.662        |
|                      | Weighted QR    | 0.027 | 0.017         | 0.010        | 0.017    | 4.532                  | 0.669        |
|                      | G-comp (MC)    | 0.018 | 0.012         | 0.007        | 0.003    | 2.476                  | 0.745        |
|                      | G-comp (approx)| 0.018 | 0.012         | 0.007        | 0.003    | 2.471                  | 0.745        |
| Strong 4             | Unadjusted     | 0.032 | 0.015         | 0.009        | 0.020    | 5.633                  | 0.796        |
|                      | QR             | 0.032 | 0.015         | 0.009        | 0.014    | 4.172                  | 0.733        |
|                      | IPW estimator  | 0.038 | 0.018         | 0.010        | 0.029    | 6.956                  | 0.662        |
|                      | Weighted QR    | 0.037 | 0.018         | 0.010        | 0.028    | 6.831                  | 0.669        |
|                      | G-comp (MC)    | 0.024 | 0.011         | 0.007        | 0.008    | 3.304                  | 0.745        |
|                      | G-comp (approx)| 0.024 | 0.011         | 0.007        | 0.006    | 2.915                  | 0.745        |