Published in final edited form as:

Int J Psychol. 2017 April ; 52(2): 96–105. doi:10.1002/ijop.12275.

Squeezing observational data for better causal inference: Methods and examples for prevention research

Diego Garcia-Huidobro^{1,2,3} and J. Michael Oakes⁴

¹Department of Family Medicine, School of Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

²Department of Family Social Science, University of Minnesota, Saint Paul, MN, USA

³Department of Pediatrics, University of Minnesota Medical School, Minneapolis, MN, USA

⁴Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA

Abstract

Randomised controlled trials (RCTs) are typically viewed as the gold standard for causal inference. This is because effects of interest can be identified with the fewest assumptions, especially imbalance in background characteristics. Yet because conducting RCTs are expensive, time consuming and sometimes unethical, observational studies are frequently used to study causal associations. In these studies, imbalance, or confounding, is usually controlled with multiple regression, which entails strong assumptions. The purpose of this manuscript is to describe strengths and weaknesses of several methods to control for confounding in observational studies, and to demonstrate their use in cross-sectional dataset that use patient registration data from the Juan Pablo II Primary Care Clinic in La Pintana-Chile. The dataset contains responses from 5855 families who provided complete information on family socio-demographics, family functioning and health problems among their family members. We employ regression adjustment, stratification, restriction, matching, propensity score matching, standardisation and inverse probability weighting to illustrate the approaches to better causal inference in non-experimental data and compare results. By applying study design and data analysis techniques that control for confounding in different ways than regression adjustment, researchers may strengthen the scientific relevance of observational studies.

Keywords

Causal inference; Methodology; Methods; Observational studies; Randomised trials

Despite limitations and controversy (Oakes, 2013; Victora, Habicht, & Bryce, 2004), there is substantial agreement that randomised controlled trials (RCTs) are the gold standard for causal inferences (Friedman, Furberg, & DeMets, 2010; Hernan & Robins, 2016; Rothman, Greenland, & Lash, 2008). In RCTs, participants have exactly the same probability of being

exposed or unexposed to a certain intervention, and with large numbers and minimal loss to follow-up, randomisation ensures that participants in a treatment group are exchangeable with those in a control group, save for the exposure to the intervention. Yet because RCTs are expensive, time consuming and sometimes ethically problematic, observational studies are frequently used to study causal associations. In this manuscript, basic concepts of effect identification will be reviewed, strategies to design and analyse observational studies “as-if” these were RCTs will be described, and the data analysis techniques will be illustrated using a cross-sectional dataset evaluating the effect of a hypothetical intervention to improve family functioning and reduce attention deficit hyperactivity disorder.

APPROACHES TO CAUSATION AND CAUSAL INFERENCE

Inferring whether an association is causal or not is fundamental in research and in the development of preventive interventions that truly prevent undesired health outcomes in a given population. *Inference* is the process of extrapolating the findings of a study in a given sample to a larger population, and *causal inference* is the process of extrapolating a causal relationship between an exposure and an outcome observed in a sample, to the larger population (Hernan & Robins, 2016).

There are multiple philosophical and practical approaches to causality and causal inference (Babbie, 2012; Rothman et al., 2008; Szklo & Nieto, 2012). Reviewing all approaches to causation is beyond the scope of this manuscript, and we refer readers to Heckmann (2008); Pearl (2009) and Spirtes, Glymour, and Scheines (2011) for in-depth discussions about them. Among these approaches, we would like to highlight the Counterfactual or Potential-Outcome Model, which is now widely regarded as key to understanding causal effects.

Potential-outcome model for causal inference

In simplest terms, this model proposes that a causal effect is the difference between an observed outcome when a person is exposed or treated and the hypothetical outcome of the very same person had they not been exposed or treated. This is called the counterfactual (counter-to-the fact) comparison. Since it is obviously impossible to observe the hypothetical or counterfactual outcomes, researchers seek the best observable comparison to substitute for the unobservable counterfactuals. Differences between desired but unobservable counterfactuals and observable counterfactual substitutes make up the bulk of methodological obstacles in research.

In order to identify, or tease out, a cause, competing hypotheses or explanations must be ruled out. For example, to state that a child’s ADHD is caused by problematic family functioning, all other potential causes of that child’s condition must be excluded, including genetic, co-morbid mental health conditions, learning disabilities and other environmental factors besides family functioning. If these potential explanations cannot be discarded, there is an identification problem, because no single explanation accounts for the observed phenomena.

According to this model, three conditions are central in the identification of causal effects: positivity, consistency and exchangeability (Hernan & Robins, 2016). *Positivity* requires that

any person has a probability greater than zero to be exposed to the intervention (or hypothetical intervention). This means that causal effects can only be identified amongst people who might be exposed to the causal agent. If certain populations cannot be exposed it is not possible to estimate causal effects among them, and their data is not informative.

Consistency refers to the requirement that all individuals in the population experience the same intervention or exposure (equal type, intensity and duration). For example, to study the causal effect of family functioning in ADHD every child in the sample should live in families with similar functioning. This way, it is possible to isolate the effect of specific levels of family functioning. The presence of dissimilar treatments or exposures is particularly problematic when the causal effect varies according to the levels of exposure.

The very important idea of *exchangeability* comes next. Here, the idea is that observable counterfactual substitutes are the same as desired but unobservable counterfactuals but for the intervention or exposure itself. Setting aside measurement and related errors, when study participants are exchangeable, the only remaining explanation for the appearance (or not) of the outcomes is the exposure. It is in this way that the counterfactual explanation allows identifying causal associations between an exposure and an outcome. Because in real life there are no perfect substitutes for another person, there is always the possibility of differences that are related to the outcome of interest. When this happens, we say there is *confounding*, meaning that the relationship between an exposure and an outcome is biased because of a common cause, or a *confounder* (Hernan & Robins, 2016). In addition to confounding, there are other types of biases (e.g., selection and measurement bias), that can lead to erroneous estimation of causal effects. Discussing how these biases affect causal inference is beyond the scope of this manuscript, and we refer interested readers to Hernan and Robins (2016) for in-depth discussions about these topics.

Besides positivity, consistency and exchangeability, the identification of causal effects requires *temporality*. This means that the outcome must be preceded by the exposure. This implies that the identification of causal associations requires longitudinal data collection or assumptions about temporal sequences.

Even though there is controversy about the importance of these conditions (Cole & Frangakis, 2009; Westreich & Cole, 2010), the larger the violation of these principles, the more biased the study findings and the conclusions of a study may be. Prevention researchers should be familiar with these principles and aim to fulfil them in their research with sound study designs and/or data analysis techniques.

STRATEGIES TO REDUCE CONFOUNDING IN OBSERVATIONAL STUDIES

Reducing confounding through the study design

At the study design level, strategies to minimise confounding are randomisation, restriction and matching.

Randomisation—Although it is not possible to ensure exchangeability at the individual level, randomisation of large numbers with no or negligible lost to follow-up ensures

exchange-ability at the group level. This means that in expectation, randomisation produces comparable groups in measured and unmeasured confounders, and therefore excellent counterfactual substitutes (Oakes, 2013). In other words, the control group serves to substitute for the unobservable counterfactuals of the treatment group, at least with large samples. Ensuring exchangeability among research groups is what makes RCTs the gold standard for causal effect identification. In addition, because all participants have the same probability of receiving the same intervention, positivity and consistency principles are satisfied. However, because randomisation is expensive, evaluates one effect at a time, requires follow-up and is not possible in many situations (e.g., randomising children to live in different families according to their family functioning), observational and quasi-experimental studies are the rule rather the exception (Rosenbaum, 2002), and other strategies to limit confounding must be implemented to identify causal effects.

When randomisation does not exist, the so-called treatment assignment mechanism is rarely independent of the outcomes and some subjects may have a zero probability of being exposed. Thus, we are often better off comparing exposed subjects to the best counterfactuals we can find for them. The best ones are the non-exposed subjects that have the same probability of being exposed as the exposed subjects being studied (or the set of them). This theoretically satisfies the critical assumption about independence of the treatment assignment mechanism. Therefore, responding the questions *How can positivity, consistency, and exchangeability principles be satisfied in this study?* and *How can this study be designed and analysed to simulate a RCT?*, is fundamental for strong causal inference with non-randomised studies.

Restriction—This strategy consists of limiting the studied sample to only a certain group of participants. For example, if having a genetic predisposition to ADHD is a known confounder of the association between family functioning and the incidence of this condition, the study sample could be composed only by participants who are either genetically predisposed (e.g., have a relative that was diagnosed with ADHD) or not. By just restricting the sample to only one of the groups, the confounding effect is removed. Although this approach reduces the generalisation (or external validity) of the findings (e.g., the association between family functioning and ADHD was studied in a limited sample and not in the entire population), the *causal inference* (or internal validity) is stronger, because of less confounding. However, this can only be performed with known confounders.

Matching—In this strategy, confounding variables are identified a-priori, and when participants enrol, they are paired to counterparts who share those same characteristics. In the example of genetic predisposition, family functioning and ADHD, participants with high and low levels of family functioning (exposure of interest) can be matched according to their genetic predisposition (exposure to the confounder). By design, this procedure ensures exchangeability of participants in known confounding variables, and therefore satisfies this requirement of effect identification. In addition to only be possible with known confounders, and that as the number of these increase, it becomes harder to find participants to match each other.

Reducing Confounding through Data Analysis

Although it is preferable to include strategies to minimise confounding at the study design phase, these can also be done at the data analysis phase. At this stage, strategies to reduce confounding include regression adjustment, restriction, stratification, matching, propensity score matching, standardisation, inverse probability weighting (IPW) of marginal structural models, among others (e.g., g-estimation and using instrumental variables). Strengths and limitations of these techniques are presented in Table 1, but it is important to highlight that only standardisation, IPW and g-estimation can be used to estimate effects of time-varying exposures (Hernan & Robins, 2016).

Regression adjustment—This is the most commonly used method to control for confounding (Rothman et al., 2008). In essence, all different types of regression techniques estimate values of an outcome conditional on some exposure or treatment of interest, and perhaps other control or confounding variables. For example, when estimating the effect of family functioning on ADHD, regression adjustment allows “controlling” for the effect of potential confounders (e.g., genetic predisposition, school problems), by including them as covariates in the statistical model. Although this approach is simple and the interpretation very appealing, it is problematic because it is unclear what the meaning of “holding all else constant” is: the same score?, an average?, a reference value? In addition, the effect of all other variables cannot be held constant in real life, and maybe the adjusted values might not represent any person in the sample (Cochran, 1957). Moreover, unless the potential confounders are specified in advance, researchers often end up capitalising on chance (Berk, 2004). Finally, even this was conceivable and the interpretation understandable, it is extremely difficult to remove the complete effect of a confounder by just adding it into a statistical procedure. Because of measurement error, residual confounding is always possible, and therefore regression does not end up “adjusting” completely for the confounder of interest.

Stratification—This procedure removes confounding by conducting the same analysis within different strata of exposures to a confounder. For example, the association of family functioning and ADHD can be analysed according levels of genetic susceptibility, presence of a learning disability, and levels of school problems. If a researcher wants to control for multiple confounders at the same time, analyses could be conducted within different strata of multiple variables. If the results are consistent across strata, it means that there was no statistically significant confounding, and some authors suggest combining the analysis using weighted averages of the strata-specific estimates of effect (Szklo & Nieto, 2012). However, no statistically significant confounding does not mean no confounding. On the other side, if the results are different between strata (statistically or not), the effects of an exposure on the outcome should be reported according to the levels of the confounder.

Restriction and matching—These strategies reviewed at the study design phase, can also be applied when analysing data, by limiting the analysis to sub-samples of the data, either by focusing in certain groups (restriction) or by pairing observations according to potential confounders (matching). The main limitation of using them at this stage is that as data has already been collected, these methods reduce sample sizes considerably. However,

having a more homogeneous sample allows for better causal inference, which could outweigh this disadvantage.

Propensity score matching—This modern technique to control for confounding aims to mimic a RCT by matching participants in their probability of being exposed given the subjects' background characteristics. The most important assumption of this method is that treatment assignment is independent of potential outcomes conditional on covariates. As with randomised experiments, the expected result of an independent treatment assignment mechanism is balance in confounders between treated and untreated groups, thus yielding exchangeable counterfactual substitutes. To do this, first the outcome variable needs to be set aside. Then, use a statistical model to determine the propensity or probability of being exposed (frequently using logistic regression). Once the predicted probabilities of exposure (propensity scores) are calculated, randomly match exposed and unexposed participants on these probabilities, and remove those who did not have matches. Then assess balance between groups. If groups are unbalanced, re-estimate the propensity score until balance is achieved. If groups are balanced, analyse the outcome using the matched sample. Although this is not the same as a prospective randomisation, by matching participants on their probability of exposure this technique mimics the random assignment of a RCT and ensures exchangeability amongst the matched subjects.

Standardisation—This technique matches the distribution of confounders to the larger population that each individual represents or is being compared to (*the standard population*). This procedure could be easily conducted one-at-a-time for categorical variables according to each stratum of the variable, but gets complicated as the number of confounders is greater than one, these have multiple levels, or are continuous. Fortunately, software packages have implemented the standardisation formula (general- or g-formula), which allows computing marginal probabilities of an outcome, standardising that result to the probability of that outcome among the different levels of exposure according to the distribution of their covariates in the larger population of the study. Assuming exchangeability, this result could be interpreted as the probability of an outcome in the exposed and unexposed given the distributions of the confounders in the exposed and unexposed populations.

Inverse probability weighting—This statistical method estimates exposure effects on an outcome by weighting the contribution of each observation on the inverse probability of their exposure. Weights are estimated based the conditional probability of exposure given the distribution of the measured confounders in the observed cases. In other words, weights are assigned based on the propensity score of each observation. Observations with lower probabilities of exposure have larger weights and observations with higher probabilities of exposure have smaller weights. This creates a “pseudo-population”, in which exposure is independent of the measured confounders, and therefore confounding and selection bias are eliminated. This is an important step that needs to be confirmed, because it assumes that censored and uncensored observations are exchangeable conditional on the covariates. In addition, as the pseudo-population represents two identical people for each case, one exposed and one unexposed, the mean of the estimated weights should average 2.00. This should also be confirmed, as deviations from this value represent misspecification of the

exposure estimation or possible violations to positivity (Hernan & Robins, 2016). Once inverse probability weights are estimated, traditional regression models can be used in the weighted population without the addition of any covariates, as confounders are represented in the weight given to each observation. Because IPW estimates expected values of an outcome in the exposed and unexposed without covariates, these are called *marginal structural models* (Hernan & Robins, 2016).

EXAMPLES OF ANALYTIC METHODS TO REDUCE CONFOUNDING

The analytic techniques just described will now be illustrated using a cross-sectional dataset. This study design was selected because it is the most common type of research study (Rosenbaum, 2002). The research question guiding the data analyses is: What is the effect of family functioning on ADHD? Analyses will aim to ensure the exchangeability as in a RCT, evaluating the causal association of a hypothetical intervention that improves family functioning to reduce ADHD (e.g., a family-strengthening intervention).

Dataset

The dataset includes de-identified information from the electronic records of families registered at the Juan Pablo II Primary Care Clinic in La Pintana, Chile, from its opening in September 2006 until March 2010. This clinic is located in an underserved district of the South-East Metropolitan area of Santiago (Diaz et al., 2010) and serves approximately 25,000 individuals.

Participants

When people registered at the clinic, they were asked to group themselves in families according to their own definition of their family group, and to select a family member to answer a three-section questionnaire for the entire group.

A total of 6202 families completed the registration survey. They represented 25,037 individual family members (average 3.3 ± 1.9 (mean $\pm$ SD) members per family, ranging from 1 to 22). Most family members were female (12,929, 51.6%), and children under 20 included 44.1% of the sample (11,036). Most families had a monthly income equal or lower to the Chilean minimum wage (about \$300 US dollars) and 8.5% identified having an illiterate family member ($n = 526$). Survey responders were mostly female ($n = 4630$, 74.7%), with an average age of 43.3 years. Of the total number of families, 5855 completed the family functioning evaluation (94.7%). This is the total sample included in the analyses.

Measures

The questionnaire used for registration at the clinic had 67 questions. The first part of the survey included sociodemographic information about all family members, the second section assessed health problems within the family, and the third section evaluated the family functioning style. An example of the questions used to evaluate health problems in the family is: *Is there a child or an adolescent in your family, meaning someone 20 years or younger, that has any of the following?* Among children and adolescents (aged 20 years or younger), health problems evaluated included: delayed child development, concentration

difficulties or being extremely active, conduct problems at school, learning disabilities, among others. Response options were: yes, no and don't know. If survey responders answered positive to any problem, they were also asked to report how many family members had that issue.

Family functioning was measured using the 22-item *Family Functioning Style Scale* (FFSS), which has been validated (Larrain, Zegers, Diez, Trapp, & Polaino-Lorente, 2003) and used in Chilean primary care (García-Huidobro, Puschel, & Soto, 2012). This instrument appraises seven family areas using a 5-point likert scale (almost never to almost always). Assessed areas of family functioning include agreement, cohesion, support, strengths, commitment, problem solving strategies and resources. Examples of included items are: *How often in my family: we try to look at the positive side of things?, we talk about different ways to treat and solve problems?, we support each other when something unexpected happens?* A global score was calculated by adding the scores of all items, and higher scores reflect better family functioning ($\alpha = 0.90$).

Dependent variable—Self-identified having a child with concentration difficulties or extremely active, or having conduct problems at school was the outcome of interest. Families with these children were classified as potentially having a child with ADHD. In total, 1451 families (24.8%) self-reported this condition.

Independent variable—In a randomised experiment, this would be the intervention allocation. For the purpose of this example, we will model the effects of a hypothetical family-strengthening intervention that modified family functioning in the sample. To determine the participant's group in the hypothetical intervention, families were divided in two groups according to their scores on the FFSS: below or equal, and above the median (91 points). By definition, this division assumes that all families in the simulated intervention group had higher scores in the FFSS compared to the families in the control group (100.7 ± 5.6 vs. 76.3 ± 14.0 , $p < 0.001$), mimicking an intervention that would have improved family functioning in all families. Table 2 summarises participant and household characteristics according to the family's family functioning group.

Covariates—These included monthly family income, self-reported having a family member with depression, alcohol or illicit substance use problems, or reporting the experience of family violence. Because assigned groups were statistically different in all household characteristics (Table 2), which were also statistically associated with the outcome of interest (data not shown), these household characteristics confound the association between family functioning and ADHD (Rothman et al., 2008).

Data analysis

All described strategies to minimise confounding are exemplified. Before conducting the analyses, a crude or unadjusted model without covariates was fit. All calculations were conducted using STATA 14 (StataCorp, College Station, TX).

Regression adjustment—A Generalised Linear Model (GLM) was used to model the probability of having ADHD adjusting for all potential household confounders. We

estimated risk differences by using a GLM with Bernoulli error distribution and an identity link (Diggle, Heagerty, Liang, & Zeger, 2002).

Stratification—Because all household variables were related to family income and identifying a family member with depression (data not shown), all analyses were stratified by income and depression of a family member, and adjusted for the other covariates. Therefore, four GLMs (higher income and no depression, higher income and depression, lower income and no-depression, and lower income and depression) were fit.

Restriction—Analyses were limited to the a sub-sample of families who reported having the minimum wage as their monthly family income and not having family members with depression, alcohol or substance use problems, or experiencing family violence. The new sample included 1900 families. Because all confounding was managed by restricting the sample, the statistical relationship between family functioning and ADHD was assessed with a GLM with no covariates.

Matching—Participants from families with high and low FFSS scores were randomly matched 1:1 based on all covariates. This procedure identified 2526 matches ($n = 5,052$, 86.3% of the population). Among the matched cases, a GLM was used to determine the association between family functioning and ADHD in each dataset. Because all confounding was removed by matching, the model did not include covariates.

Propensity score matching—All covariates were entered into a logistic regression model to estimate the propensity of living in a family with high or low FFSS scores. Once these propensity scores were estimated, participants from families with high and low functioning were randomly matched based on their propensity to be in one of these types of families. Matches were specified to pair within a range of 0.01 (matching caliper). Participants who were not matched were excluded from the analysis ($n = 1$). Balance on covariates between groups was confirmed and the average treatment effect of family functioning on ADHD was estimated using a t -test. Then, 1000 replications were used to bootstrap the correct standard errors and confidence intervals.

Standardisation—The predicted probabilities calculated from the adjusted GLM were incorporated into the standardisation formula and standardised to the entire population of the study using the *margins* post-estimation command of STATA.

Inverse probability weighting—The propensity score calculated previously was used to estimate the weights. Among exposed participants final weights were the inverse values of the propensity score ($1/\text{propensity score}$), and among unexposed weights were the inverse values of the probability of not being exposed ($1/(1-\text{propensity score})$). These weights were used in the marginal structural GLM model (without covariates) to estimate the exposure effect. Standard errors were bootstrapped using 1000 replications.

RESULTS

Table 3 shows risk differences of the hypothetical intervention on ADHD, according to the different data analysis strategy. Unadjusted results showed that families in which the survey responder rated their family with higher levels of functioning on average had 8.5% lower likelihood of self-reporting that at least one child in their household could possibly have ADHD. All analytic techniques reduced the risk difference of ADHD by levels of family functioning between 0.7 and 7.2%. This confirms that household variables confounded the association between these two variables.

The stratified analysis demonstrated an important interaction between household income and if the survey responder reported that at least one family member had depression. Second to the stratified the results of the group with lower income and that had a family member with depression, the matched analysis reported the most conservative effect estimates. Propensity score matching, standardisation and IPW reported results that were close to the ones obtained with traditional regression adjustment.

DISCUSSION

What results should I report?

All analyses used different techniques to reduce confounding. Therefore, each method is estimating different conditional causal effects, which could be one reason why the estimates are not identical between techniques. In addition, each method has its own set of modelling assumptions and that is another possibility of why the estimated effects differ between approaches.

Although the effect estimates of the different analytic techniques are different, we consider that these do not differ much from each other. However, what truly differs between the presented methods is the causal inference that can be claimed from each of them, as all of them approach to the exchangeability condition of effect identification in different ways. Thus, the quality of the causal inference is better by using alternatives with that ensure greater exchangeability (as with restriction, case matching, propensity score matching, standardisation and IPW) compared with traditional regression. This is the most important message of this manuscript: the effect estimates might not change, but the quality of the scientific inference often will.

If a researcher is comfortable with the assumptions implied with regression adjustment, s/he should continue using that analytic method. We simply suggest that there are analytic alternatives that impose different assumptions to the computational modelling. At the data analysis phase, stratification divides the analysis according to potential confounders, restriction limits the analysis to a more homogeneous sub-sample, matching pairs subjects who have the exact same value on a variable (or within a determined caliper), propensity score matching pairs subjects according to their likelihood to be assigned to the exposure of interest, standardisation homogenises confounders among the whole population, and IPW ponders data based on their likelihood of exposure. Additional methods not covered in this manuscript are g-estimation and using instrumental variables. The first approach estimates

marginal structural models but without assumptions of a direct effect of covariates (hence estimating a *structural nested model*). The latter, estimates causal effects by using a variable (the *instrumental variable*) that is not directly related to the exposure nor the outcome, as a way to control for measured and unmeasured confounding. More descriptions about these methods are offered at Hernan and Robins (2016), but all of them make different assumptions about ways of approximating to the exchangeability requirement for effect identification that is expected of a large RCT with minimal loss to follow-up.

Another consideration in the selection of an analytical technique at the data analysis phase is the amount of available data. Stratification, restriction and matching are the methods that require the larger amounts of data. An example of this limitation was observed in the presented example. Although the dataset was large, it did not allow us to stratify the analysis with more than two variables.

Limitations of this example

As any other study, our report has limitations. First, we conducted secondary data analyses using a dataset with limited number of variables. There are numerous unmeasured confounders that constrain our capacity to ensure we have eliminated all confounding, either in the direct analyses (regression, matching, stratification or restriction) or through the estimation of the exposure probability (propensity score matching, standardisation or IPW). This means that effect estimates are probably biased. Second, the dataset was cross-sectional. Therefore, there is no way to establish a temporal direction between the variables that we have designed as exposure and outcome: we assumed temporality. It is completely possible that having a child with ADHD affects the functioning style of a family, which limits our ability to make a causal claim to the findings of the study. Third, this example has important amounts of measurement bias in dependent, independent and covariate variables, as all these were self-reported by a single family member. Fourth, this exercise also has important amounts of selection bias: we defined that families with better family relationships and mental health were going to be treated by the hypothetical intervention (Table 2). Finally, the hypothetical intervention by definition was effective in *all* exposed families (all families with a score higher than the median were classified as exposed) and modified family functioning in a magnitude that is just unbelievable (about 25 points of the FFSS, which reflects an effect size of about 1.5 standard deviations). These limitations are very important for causal inference, as no analytic technique will overcome the limitations imposed by poor study designs and important measurement errors. Regardless of these limitations for the causal inference of the findings, this example allowed us to discuss the strengths and weaknesses of multiple techniques to manage confounding and exemplify them using a Chilean dataset that presents common challenges for researchers working in contexts with limited resources.

Recommendations for researchers

Prevention researchers are encouraged to conduct RCTs testing the effect of specific interventions. However, because it is known that this type of research is uncommon, we encourage preventionists to think in ways to either design or analyse observational studies “as-if” these were the ideal experiments testing different hypotheses. This means, to fully

envision how the ideal RCT would be, what would be the competing hypothesis or explanations that are needed to rule out and to think deeply about ways to approach the ideal RCT with an observational study. When doing this, think about the assumptions of each decision, and describe the requisites that are necessary to infer causal conclusions.

Doing this will improve the quality of our science and ultimately the public's welfare. Sometimes, researchers will notice that certain effects will not be identifiable. This will encourage them either to conduct the ideal RCT, or to not even bother spending time and other resources understanding causal effects that could not be fully identifiable with a given dataset. Other times, researchers will notice that although the effects could not be identifiable, taking that approach will highlight the limitations of their research. This will improve the quality of the scientific reports, if these weaknesses are clearly described in the discussion sections. Recognising methodological short-coming is an important issue to advance any scientific field and makes prime targets for further research.

CONCLUSION

We have reviewed the basic conditions to identify causal effects and emphasised strategies to control for confounding at the study design and data analysis phases, offering illustrative examples. Although most research is conducted to answer a causal explanation, the core elements of effect identification are not universally understood and rarely used in practice (Oakes, 2013). We encourage prevention researchers to think about study design and data analysis approaches in observational studies that preserve the positivity, consistency and exchangeability properties of RCTs and rule out competing explanations. This way, preventionists who have limited access to conduct RCTs will better exploit observational studies to maximise the quality of the inference given their resources.

Acknowledgments

This research was supported in part by National Research Service Award in Primary Medical Care, grant no. T32HP22239 (PI: Borowsky), Bureau of Health Professions, Health Resources and Services Administration and U.S. Department of Health and Human Services.

References

- Babbie, E. The practice of social research. Belmont, CA: Wadsworth; 2012.
- Berk, RA. Regression analysis: A constructive critique. Thousand Oaks, CA: Sage; 2004.
- Cochran WG. Analysis of covariance: Its nature and uses. *Biometrics*. 1957; 13(3):261–281.
- Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*. 2009; 20(1):3–5. [PubMed: 19234395]
- Díaz, F., Gatica, S., González, P., Labra, Y., Labrín, R., Morales, L., Vera, R. Diagnóstico Epidemiológico Servicio de Salud Sur Oriente. Chile: Departamento de Control de Gestión; 2010. Retrieved from: <http://www.ssmso.cl/tmparchivos/portal/destacado/Diag.pdf>
- Diggle, P., Heagerty, P., Liang, KY., Zeger, S. Analysis of longitudinal data. Oxford: Oxford University Press; 2002.
- Friedman, LM., Furberg, C., DeMets, DL. Fundamentals of clinical trials. New York, NY: Springer; 2010.
- García-Huidobro D, Puschel K, Soto G. Family functioning style and health: Opportunities for health prevention in primary care. *British Journal of General Practice*. 2012; 62(596):e198–e203. [PubMed: 22429437]

- Heckman JJ. Econometric causality. *International Statistical Review*. 2008; 76(1):1–27.
- Hernan, MA., Robins, JM. Causal inference. Boca Raton, FL: Chapman & Hall/CRC; 2016.
- Larain M, Zegers B, Diez I, Trapp A, Polaino-Lorente A. Validity and reliability of the Spanish version the Dunst, Trivette & Deal's Styles of Family Functioning Scale (SFF), for the diagnosis of family functioning in the Chilean population. *Psyche*. 2003; 12(1):195–211.
- Oakes M. Effect identification in comparative effectiveness research. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2013; 1(1, 4):1–8.
- Pearl, J. Causality: models, reasoning and inference. Cambridge, MA: Cambridge University Press; 2009.
- Rosenbaum, PR. Observational studies. New York, NY: Springer; 2002.
- Rothman, KJ. Greenland, S., Lash, TL., editors. Modern epidemiology. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
- Spirtes, P., Glymour, CN., Scheines, R. Causation, prediction, and search. New York, NY: Springer; 2011.
- Szklo, M., Nieto, FJ. Epidemiology: beyond the basics. New York, NY: Jones & Bartlett Publishers; 2012.
- Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *American Journal of Public Health*. 2004; 94(3):400–405. [PubMed: 14998803]
- Westreich D, Cole SR. Invited commentary: positivity in practice. *American Journal of Epidemiology*. 2010; 171(6):674–677. [PubMed: 20139125]

TABLE 1

Strengths and limitations of data analysis methods to control for confounding

<i>Method</i>	<i>Strengths</i>	<i>Limitations</i>
Regression adjustment	<ul style="list-style-type: none"> • Easy to perform • Easy result interpretation 	<ul style="list-style-type: none"> • Definition and number of variables in a model • Potential for overadjustment (adjusting for a mediator) and residual confounding • Not useful with time-varying exposures and covariates
Stratification	<ul style="list-style-type: none"> • Easy to perform • Easy result interpretation • Does not impose assumptions in the analytic method • Allows for the evaluation of effect modification (moderation) 	<ul style="list-style-type: none"> • Only for categorical covariates • Continuous variables require categorization • Difficult with large number of variables • Not useful with time-varying exposures and covariates
Restriction	<ul style="list-style-type: none"> • Confounding is removed by definition 	<ul style="list-style-type: none"> • Reduction in sample size • Continuous variables require categorization • Not useful with time-varying exposures and covariates
Matching	<ul style="list-style-type: none"> • Ensures exchangeability among measured variables • High statistical efficiency 	<ul style="list-style-type: none"> • Number of variables • Precision of the matching • Reduction in sample size • Potential for overmatching (matching on a mediator or on a variable that is not a confounder) • Not useful with time-varying exposures and covariates
Propensity score matching	<ul style="list-style-type: none"> • Strong causal assumptions • Ensures exchangeability of exposed and unexposed 	<ul style="list-style-type: none"> • Best for categorical exposures • Precision of the matching (caliper) • Requires bootstrapping standard errors • Not useful with time-varying exposures and covariates
Standardisation	<ul style="list-style-type: none"> • Strong causal assumptions • Ensures exchangeability of exposed and unexposed • Useful with time-varying exposures 	<ul style="list-style-type: none"> • Requires complex statistical computation
Inverse probability weighting	<ul style="list-style-type: none"> • Strong causal assumptions • Ensures exchangeability of exposed and unexposed • Useful with time-varying exposures 	<ul style="list-style-type: none"> • Requires bootstrapping standard errors

TABLE 2

Summary of survey responder and household characteristics according to hypothetical intervention group

	<i>Lower FFSS ("control group")</i> n = 3030	<i>Higher FFSS ("intervention group")</i> n = 2825	p-value
Survey responder			
Gender, female	2271 (75.0)	2125 (75.2)	0.862
Age, years, mean $\pm$ SD *	44.6 $\pm$ 14.4	44.6 $\pm$ 13.8	0.959
Marital status, married *	1268 (41.9)	1199 (42.4)	0.645
Household			
Child with potential ADHD	875 (28.9)	576 (20.4)	< 0.001
People living in the household, mean $\pm$ SD *	3.3 $\pm$ 2.0	3.2 $\pm$ 1.9	0.005
Minimum wage as monthly family income *	1742 (57.5)	1471 (52.1)	< 0.001
Adult family members with depression *	554 (18.3)	375 (13.3)	< 0.001
Adult family member with drinking problem *	627 (20.7)	321 (11.4)	< 0.001
Adult family member with substance use problem *	499 (16.5)	264 (9.4)	< 0.001
Family violence *	426 (14.1)	219 (7.8)	< 0.001

Note: Values are *n* (%), unless otherwise stated;

* denotes variables that were statistically associated with ADHD.

TABLE 3

Risk differences and 95% confidence intervals of the effect of a hypothetical intervention improving family functioning on ADHD using different data analysis techniques

	<i>Risk difference</i>	<i>95% confidence interval</i>
Unadjusted regression ($n = 5855$)	-0.085	-0.107, -0.063
Adjusted regression ($n = 5585$)	-0.064	-0.087, -0.042
Stratification		
Higher income, no depression ($n = 2268$)	-0.060	-0.095, -0.025
Higher income, depression ($n = 374$)	-0.072	-0.173, 0.028
Lower income, no depression ($n = 2658$)	-0.075	-0.108, -0.043
Lower income, depression ($n = 555$)	-0.013	-0.095, 0.070
Restriction ($n = 1900$)	-0.078	-0.114, -0.042
Matching ($n = 2526$ pairs)	-0.059	-0.082, -0.035
Propensity score matching ($n = 5854$)	-0.065	-0.097, -0.032
Standardisation ($n = 5585$)	-0.063	-0.085, -0.041
Inverse probability weighting ($n = 5855$)	-0.060	-0.084, -0.037