COMMENTARY MARP PROJECT

A workflow for causal inference in cross-cultural psychology

Joseph A. Bulbulia

Victoria University of Wellington, New Zealand// https://orcid.org/0000-0002-5861-2056

ARTICLE HISTORY
Compiled March 12, 2022

ABSTRACT
The causal interpretation of a statistical association requires assumptions. Where the data are cross-sectional or cross-cultural these assumptions are even stronger. Here, I leverage a rigorous potential outcomes framework from contemporary epidemiology to (1) sharpen the causal question of whether religious service attendance reduces anxiety, and (2) develop a workflow for addressing this causal question using data from the Multiple Analysts of Religion Project (MARP, $N = 10,535$; 24 countries). This workflow clarifies how we may obtain a counterfactual contrast necessary to infer an average causal effect that is subject to (very) strong assumptions that causal inference requires in this setting. A sensitivity analysis helps to assess the robustness of the causal effect estimate to unmeasured confounding. This study is interesting because it clarifies the conditions under which the statistical association between religious service attendance and lower anxiety in the MARP dataset may have a causal interpretation. More generally, this study presents the potential outcomes framework as a rigorous method for investigating causal questions. I hope the potential outcomes framework will interest psychological scientists who seek better causal hygiene in their own research.

KEYWORDS
Causality, Causality Crisis, Culture, DAG, Distress, E-values, Marginal Structural Model, Potential outcomes, Religion, Sensitivity analysis, Target Trial, Well-being

1. Introduction

1.1. Statistical associations do not quantify causal effects

Statistical associations may be useful for predicting the future. Despite the prevalence of the term "predict" in psychology journals, however, psychological scientists are rarely interested in prediction. Rather, we are interested in how cognition and behaviour works. We have causal interests. Controlled experiments are the gold standard for causal identification. However, many fundamental questions can only be addressed using observational data.

Psychological scientists typically use a framework of "prediction" to interpret statistical associations in observational data. The terms "predict" and "prediction" are meant to convey caution about causation. Absent a causal framework, however, the statistical associations upon which observational researchers base their predictions have little meaning. For this reason, reporting "predictions" in the absence of a causal framework is likely to mislead (Bulbulia, Schjoedt, Shaver, Sosis, & Wildman, 2021;

CONTACT Joseph A. Bulbulia. Email: joseph.bulbulia@vuw.ac.nz

Hernán, 2018; McElreath, 2020; Pearl & others, 2009; Rohrer, 2018; VanderWeele, 2015; Westreich & Greenland, 2013).

Here, I use the Many Analysts of Religion Project (MARP) data to develop a workflow for investigating the causal effects of religion on health in a cross-sectional, cross-cultural data-set https://osf.io/v48gc/. This workflow offers strategies for clarifying the causal assumptions required to obtain the counterfactual contrast that is necessary for quantifying a causal effect. Each causal inferential strategy must be developed bespoke for the scientific setting to which it is applied. Casual inference belongs to Savile Row not Sears Roebuck. However, because there are general features of any causal inference strategy – such as obtaining credible counterfactual contrasts that are informed by previous science – I hope that general features of my workflow will be helpful to researchers with different questions, data, and explanatory interests.

1.2. The fundamental problem of causal inference

Suppose we want to learn whether a dichotomous exposure has a causal effect. We must answer three questions: (1) What would happen with exposure? (2) What would happen with no exposure? (3) Do these potential outcomes differ?

Imagine a dichotomous exposure: exposed, $A = 1$; not-exposed, $A = 0$. Y indicates the outcome for an individual who receives one of the two treatments: $Y|(A = 1)$ or $Y|(A = 0)$. We can observe only one exposure. We denote the potential outcome of Y under the exposure $Y|(A = 1)$ as $Y^{a=1}$. We denote the potential outcome of Y under exposure $Y|(A = 0)$ as $Y^{a=0}$. By physics, if we observe $Y^{a=1}$ we cannot observe $Y^{a=0}$ and vice versa. For this reason, we say that $Y^{a=1}$ and $Y^{a=0}$ are counterfactual or potential outcomes, which we write as Y^a, noting $(Y|(A = a) \neq Y^a$. This constraint in which we cannot observe the individuals outcomes that we require to evaluate a response under different exposures is called "the fundamental problem of causal inference" Gelman et al. (2020); Holland (1986); Rubin (1976).

1.3. Assumptions for estimating an average causal effect in a population

Despite the limitations of quantifying a causal effect of two or more exposures in an individual, we may be able to quantify an average causal effect of two or more exposures in a population:

$$E[Y^{a=1}] - E[Y^{a=0}]$$

We read this as the difference in expected mean outcome when the population receives the exposure $E[Y^{a=1}]$ and the expected mean outcome when the population is withheld from the exposure $E[Y^{a=0}]$. Again, we can only obtain the statistical expectation under actual exposures $E[Y|A]$: we do not directly obtain the counterfactual expectation for the entire population. We cannot cheat physics. However, we may consistently estimate the counterfactual average response $E[Y^a]$ from statistical average response $E[Y|A]$ if the following conditions are satisfied:

(1) *Positivity*: there is a non-zero probability of the exposure among the non-exposed and of non-exposure among the exposed. That is, the exposure cannot be deterministic.

(2) *Consistency*: an individual with observed exposure A has observed outcome Y equal to their counterfactual outcome $Y^A = Y^a$.
(3) *Exchangeability*: individuals are exchangeable when counterfactual outcomes are independent of the actual exposures. To describe conditional independence, we use the symbol: $\perp\!\!\!\perp$. When $Y^a \perp\!\!\!\perp A$, exchangeability is satisfied. In a randomised experiment, random allocation of participants to exposures will typically ensure exchangeability. In observational studies, however, exchangeability typically requires conditioning on a set of measured differences ($L = l$) such that $Y^a \perp\!\!\!\perp A|L = l$. We say the counterfactual outcome of Y under exposure $A = a$ is conditionally independent of exposure assignment given a set of measured confounders $L = l$.
(4) *Correct model specification*: which includes that the outcome and the exposure are measured without error. The assumption is strong, unverifiable, and unlikely to be satisfied in observational settings. Failure of this assumption does not imply that causal inference in observational settings is impossible. It rather implies caution about claiming too much.

I have stated the assumptions that we require for causal inference. We are now ready to proceed to the next step: asking a causal question (for introductions to the key concepts in causal inference, see: Gelman & Su, 2020; Hernán & Robins, 2020; VanderWeele, 2015).

1.4. What is our causal question?

In the abstract, I present the causal question: does religious service reduce anxiety? This question is motivated by previous theory and evidence that ritual behaviour reduces anxiety (see: Ejova, Milojev, Worthington Jr, Sibley, & Bulbulia, 2020; Lang, Kratky, Shaver, Jerotijević, & Xygalatas, 2015; Malinowski, 2014; Sosis, 2007). I am unaware of any evidence that religious service attendance, on average, causes a decrease in anxiety. However, it is unclear whether the buffering effect of religious service attendance for anxiety is cross-culturally robust.

If the assumptions for unbiased causal inference that I just considered were met, the MARP dataset could help us to compute the average causal effect of religious service attendance and, therefore, help to assess whether the average effect of religious service attendance for anxiety reduction is cross-culturally robust. However, my causal question has yet to be clarified. The definitions both of the exposure and the outcome remain vague. First, consider the exposure. Which specific interventions in religious service attendance do we hope to counterfactually contrast? Are we interested in contrasting a counterfactual intervention in which all people who did not attend religious service were made to attend? Or are we interested in contrasting a counterfactual intervention in which all people who participated in religious service were prevented from doing so? Or do we imagine that interventions by increments of attendance, up or down, constitute the interventions of interest? If so, what are the units of these increments? Causal inference requires satisfaction of the positivity assumption, that is, of a non-deterministic exposure assignment. The positivity assumption is violated if there is no chance that individuals could receive different exposures from the actual exposures that they received. VanderWeele raises this question for religious service attendance and argues for restriction of analysis to religiously-identified sub-samples VanderWeele (2017a) (For scepticism about comparisons of church attendance between secular and religious samples, see: VanderWeele (2017b), and for a related discussion

focusing on the causal effects of "race" see: VanderWeele and Robinson (2014)). This objection is intuitively plausible. We cannot administer religious service attendance to an atheist like we might administer a medication.

Next, consider the outcome. Suppose we were able to measure anxiety without error. How long after initiating a change in behaviour would we expect the effects of religious service attendance on anxiety to manifest? One week? One month? Many months? A year? I have not specified when the effect of service attendance on anxiety is hypothesised to occur. Imagine that religious service attendance were to cause anxiety, which in turn leads to the attenuation of service attendance among those susceptible to service-induced anxiety. In such a circumstance, we would be left with an observed sample of individuals robust to the anxiety-provoking effect. Such a bias is called "immortality bias" (see: Hernán & Robins, 2016; Hernán, Sauer, Hernández-Díaz, Platt, & Shrier, 2016). The sample that will observe may not be the treatment sample.

In short, we cannot consider a counterfactual contrast without understanding the potential intervention and the potential outcome that we are contrasting and without ensuring that the four conditions necessary for causal inference have been satisfied. I have yet to provide this mission-critical clarity.

1.5. Emulating a "target trial" to answer a causal question

Miguel Hernán and Jamie Robins suggest that observational researchers clarify their causal question by imagining an idealised experiment, or "target trial," in which researchers might ideally evaluate the effect of an intervention (Hernán & Robins, 2016, 2020; Hernán et al., 2016). Such an idealised experiment is typically implausible. If it were not implausible, there would be no point in using observational data rather than running the experiment. However, Hèrnan and Robins argue that if researchers cannot state an idealised experiment for addressing a causal question – one that clearly defines the outcome, the exposure, the contrasts, and the study design, then researchers might not know which causal question their analysis is answering. Additionally, Hèrnan and Robins argue that by emulating an idealised target trial, observational researchers are better able to clarify their causal assumptions. In the present setting, where data collection is cross-sectional, we cannot remotely emulate a target trial. However, by clarifying how we fall short of the experiment we might want to perform to infer a causal effect our assumptions become clearer.

What are the eligibility criteria for participation? All people who responded to the MARP study are eligible for inclusion in this hypothetical target trial. This consists of ($N = 10,595$) individuals across 24 countries. Both religious service non-attenders and religious service attenders are included in the sample. I present the breakdown of individuals by their level of religious identification in Table 1. Notably, among those who state that they are "atheists", only 41.7% report never going to religious service. Among those who state that they are "not religious," only 47.5% report "never" going to religious service (see Table 1). Positivity is actual, and therefore possible, so the analysis I present here uses the full sample. However, I also perform a separate analysis restricted to those who identify as religious ($N = 3,712$) (The inference remains the same, see: Appendix C).

Treatment strategies The MARP data are cross-sectional. We can only condition on the assumption that, after accounting for all measured confounding and performing

Table 1. Cross tabulation of religious service attendance frequency and religious identification

	Never	Rarely	1 × per year	Holy Days	1 × month	1 × week	>1 × per week	Overall
	(N=5061)	(N=1236)	(N=798)	(N=1371)	(N=618)	(N=957)	(N=494)	(N=10535)
Religious								
Atheist	2110 (41.7%)	172 (13.9%)	97 (12.2%)	101 (7.4%)	12 (1.9%)	8 (0.8%)	5 (1.0%)	2505 (23.8%)
Not-Religious	2404 (47.5%)	686 (55.5%)	406 (50.9%)	544 (39.7%)	134 (21.7%)	107 (11.2%)	37 (7.5%)	4318 (41.0%)
Religious	547 (10.8%)	378 (30.6%)	295 (37.0%)	726 (53.0%)	472 (76.4%)	842 (88.0%)	452 (91.5%)	3712 (35.2%)

a sensitivity analysis for unmeasured confounding, the co-variance of religious service attendance and anxiety quantifies a causal effect.

At first glance, this assumption might cast doubt about whether causal inference is sensible. How can we evaluate a causal effect that we assume? However, I have described experimental and national longitudinal evidence that religious service attendance buffers psychological distress. By conditioning my analysis on the causal assumption of potential buffering, cross-sectional data may help to evaluate whether the buffering effect of religious service on anxiety observed in previous research is robust to cultural variation.

Randomisation I generate a pseudo-randomised population using IPTW weighting of the exposure on measured confounders (Cole & Hernán, 2008; Wal & Geskus, 2011). IPTW weighting allows us to estimate a marginal (or population-level mean) effect that is balanced across the pseudo-population. IPTW weights are less efficient than stratified regression, however, standard stratification methods fail in longitudinal settings (Robins, 1986). Because the estimates from the stratified regression models and the IPTW models are nearly identical and do not affect causal inference, I report results from the IPTW model.

I include in the set of all measured confounders L, any variable that might affect both the exposure and the outcome. I must ensure that this set does not include any variable that blocks the path between exposure and the outcome (a mediator) and that is not an effect of both exposure and the outcome (a "collider" confounder) (Bulbulia et al., 2021; McElreath, 2020) [For discussion see: Appendix E].

Start and follow up The data are cross-sectional: there can be neither an initiation of a start nor a termination by the end of follow up. Thus, inference requires the assumption that the causal effects of the exposure are immediate and that prior levels of religious service and prior levels of anxiety do not affect present levels of religious service attendance or current levels of anxiety. Again, my causal estimand will be biased if there are time-dependent dynamics, such as attrition of those for whom religious service is stressful.

Outcomes I assess anxiety using self-reported measures as described in the MARP protocols https://osf.io/vy8z7/. The MARP anxiety measure asks: "How often do you have negative feelings such as blue mood, despair, anxiety, depression?" This item was measured on a Likert scale: (1) "Never" (2) "Seldom" (3) "Quite often" (4) "Very Often" (5) "Always." I standardise and scale this outcome at the sample average. I assume that anxiety is measured consistently and without error. The no-measurement error assumption is strong. However, failure of this assumption will not necessarily lead to bias against the causal null hypothesis. Non differential and non-directed measurement error in which measurement of the exposure and outcome are not correlated

will attenuate the estimate of a true causal effect. By contrast, differential or directed measurement error risks overstating the true causal effect (Hernán & Robins, 2020). I cannot test the no-measurement error assumption in the MARP dataset. However, I can perform a sensitivity analysis as part of my workflow. Sensitivity analysis is important for causal inference in observational research because we cannot generally know whether the assumptions required for causal inference have been met.

Causal contrasts The original measure of religious service attendance asks: "Apart from weddings and funerals, about how often do you attend religious services these days?" Answers are: (1) "More than once a week" (2) "Once a week" (3) "Once a month" (4) "Only on special holy days" (5) "Once a year" (6) "Less often" (7) "Never, or practically never." The MARP team reverse-scored these responses and scaled them to 0-1 https://osf.io/vy8z7/. Given that responses on this scale are ordinal, I model religious service attendance as a monotonic effect (Bürkner, 2021). My focal contrast is "no attendance" versus "attendance more than once per week." (My choice of statistical model here, as elsewhere, makes no practical difference to inference, see: Appendix B1)

I also estimate counterfactual outcomes using a non-linear continuous model of religious service attendance exposure. These models were implemented using regression splines (R Core Team, 2021). This second approach clarifies the robustness of the inference to model choice, see Appendix B.

Analysis plan I examine missing data using the Naniar package (Tierney, Cook, McBain, & Fay, 2021) in R (R Core Team, 2021). High levels of missingness in the denomination variable prevents confounder adjustment for denominational affiliations (n missing $= 5,676$, or 53% of the responses). To address selection bias from missingness in other variables, I multiply-impute ten missing data sets using both the Amelia package (Honaker, King, & Blackwell, 2011) and the Mice package (Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2021).

I estimate a (quasi)-Bayesian mixed effect model using (Bürkner, 2021), see Appendix D. Note my statistical model is not fully Bayesian because I obtain IPTW estimates using the IPW package in R (Wal & Geskus, 2011). I also estimate a marginal structural model using a Generalised Estimating Equations (GEE) using weighted (and stratified) confounder adjustment strategies to ensure that estimation (Halekoh, Højsgaard, & Yan, 2006). GEE models also enable adjustment for country-level dependencies. My inference proved to be practically equivalent (see online materials). However, my preferred method for estimating the marginal effect of religious service attendance on anxiety is a (quasi)-Bayesian mixed-effect model with IP weights. In this model, I marginalise over the country-level variances by integrating out group level effects. Bayesian estimation has many valuable features, including clarity about model assumptions and the recovery of intuitive posterior probabilities.[1]

Following Gelman et al. (2020) and McElreath (2020), I employ strong priors in our Bayesian models, and compare these to weaker priors. (Results were robust to the choice of priors, see Appendix A).

To recover an estimate of a marginal causal effect I must integrate out the country-level variation in intercepts and slopes. For this purpose, I used the emmeans

[1]Note that for time-series models with time-varying confounding Bayesian estimation remains experimental. Although this topic does relate to the questions at hand, I want to caution readers against any simple adaptation of regression based models to panel data.

package in R (Lenth, 2022). The analysis can be found at https://github.com/go-bayes/many_analysts.

2. Results

2.1. Causal Graph

Causal inference requires causal assumptions (Bulbulia et al., 2021; McElreath, 2020; Pearl, 2019; Pearl & others, 2009; VanderWeele, 2015; VanderWeele, Mathur, & Chen, 2020). Following VanderWeele (2015), I include in my list of confounders the full set of demographic variables in the MARP dataset: these were {age, country, education, male, SES}. High levels of missing data prevented me from including ethnicity (n = 219 missing) or denomination (n = 5,446 missing). I then identify the minimal adjustment set required to quantify our causal associations. I do this to avoid introducing confounding by over-conditioning (see Figure 1b/c/d).

Figure 1.a presents a Direct Acyclic Graph that clarifies my causal assumptions. Following Hèrnan, I present a simplified graph that shows as few nodes as are necessary to convey my causal assumptions. I identify the adjustment set of confounders as: L {age, country, education, male, SES}. As indicated in Figure 1.b, conditioning on a confounder that is a collider of two unmeasured confounders will introduce bias. Figure 1.c, describes a special case, in which an unmeasured confounder of the mediator and outcome can artificially increase the effect of exposure on an outcome can introduce bias. I do not "control" for variables that religious service attendance might affect, such as religiosity. To do so would be to condition on a post-treatment variable or mediator and would partially block the effect of religious service attendance. In Appendix F, I simulate post-treatment confounding. On the problem of controlling for post-treatment variables, (see: Westreich & Greenland, 2013)).

2.2. Marginal Effects

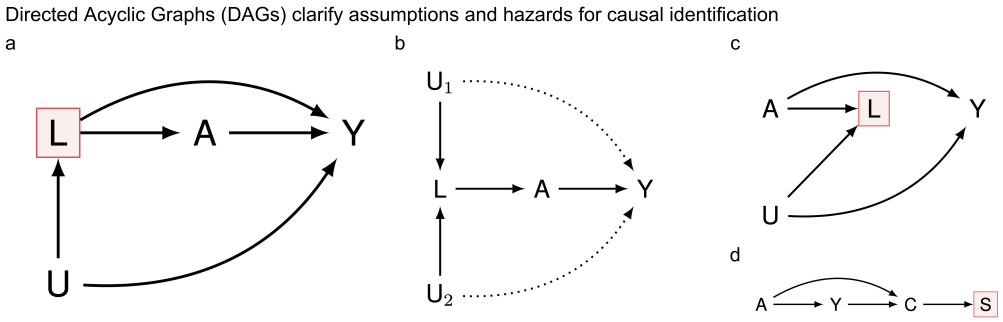
I consider the marginal predictions of the Bayesian estimator to give the most accessible and straightforward estimate of our uncertainty for the marginal effect of church attendance on anxiety. Figure 2.a presents the marginal effect estimate contrasting zero religious service attendance with those who participate more frequently than once a week: (ACE = -0.186, 95% HDP(-0.323, -0.0586)). Figure 2.a graphs average causal effect of the exposure. Figure 2.b graphs the predicted marginal effects of religious service attendance at the extremes.

Conditional on the strong assumptions described above, we find an average causal effect of religious service attendance on anxiety across the populations represented in the MARP dataset.

2.3. Sensitivity Analysis

The EValues package in R provides a user-friendly tool for investigating the sensitivity of a causal estimate to unmeasured confounding (Mathur, Ding, Riddell, & VanderWeele, 2018; R Core Team, 2021; VanderWeele & Ding, 2017). An E-value for unmeasured confounding is defined as "the minimal strength of association on the risk ratio scale that an unmeasured confounder (or confounders) would require in its association with both the exposure and the outcome to explain away a causal effect"

Figure 1. Panel (a) presents my strategy for confounder control. I assume that adjustment for set of measured confounders L {age,country, education, ses} is sufficient to close all backdoor paths between exposure A (service attendance) and the outcome Y (anxiety). This assumption is almost certainly too strong, hence I also perform sensitivity analysis. Panel (b) presents the risk of over-conditioning by adjusting for a measured confounder that opens a closed path between unmeasured confounders liking A and Y. Panel (c) presents the risk of conditioning on a measured confounder that is jointly caused by the exposure and and an unmeasured confounder. For example, imagine distressing experiences were to increase belief in God, and that church attendance also causes belief in God. Suppose belief in God does not affect anxiety, but church attendance does. In this case, conditioning on church attendance may artificially increase the negative statistical association between church attendance and distress. Indeed including belief in God in the regression model has this effect. This DAG illustrates another peril of including variables in a regression model that are not required to ensure exchangeability. I include a simulation in Appendix F that replicates such confounding. Panel (d) presents the threat of confounding by selection bias, a form of collider confounding, in which an association between A and Y is opened because sampling is biased. Such confounding is likely from MARP data, however sensitivity analysis clarifies how much confounding would be required to negate an inferred causal association.



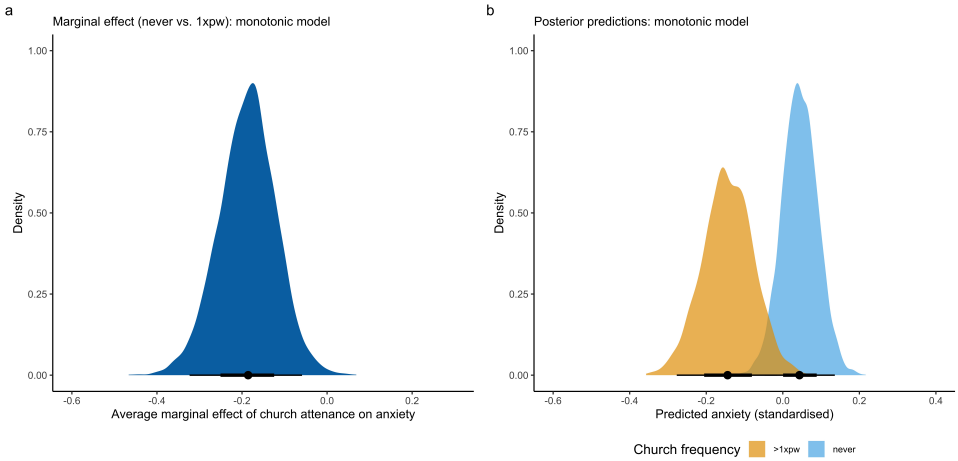
(Mathur et al., 2018). I obtain an E-value for the marginal effect estimate in the monotonic mixed-effects model value (E-value = 1.65, upper limit = 1.28). I infer that an unmeasured confounder associated with both the exposure and the outcome by risk ratios of between 1.65-fold to 1.28-fold each would be sufficient to nullify the causal association obtained from the statistical model. In the context of a cross-sectional comparative study in which we might expect strong demand characteristics, measuring error, and sampling bias, I characterise the evidence for the causal effect of religious service attendance on anxiety reduction as modest. Note that it is possible that the actual causal effect is more substantial than what I have estimated here. A stronger true causal relationship is possible if there is measurement error both in the exposure and in the outcome and if these sources of error are independent.

3. Discussion

Conditional on the strong assumptions I have described, the MARP dataset offers modest evidence for a moderate average causal effect of religious service attendance for reducing anxiety. This effect is robust both to cultural variation and a small amount of unmeasured confounding. What else have we learned?

First, although causal inference resembles ordinary regression, the assumptions re-

Figure 2. Panels (a) average causal effect of religious service on anxiety. Full church attendance may cause a .2 SD average reduction in anxiety; (b) Predicted marginal means for attending "more than once per week" versus attending "never."



quired for causal inference must be considered with care. Absent a potential outcomes causal framework, estimating associations from a regression model is misleading for causal inferential tasks. **Do not report regression coefficients as causal effects unless you have clarified your causal assumptions.**

A corollary: **all regression tables may be safely moved to the supplements. You need only report the coefficients that pertain to your causal interests: typically that is one coefficient.**

Second, causal estimation requires stating one's assumptions about the data and one's model. Formulating such assumptions require expert knowledge about the topic at hand. The data do not generally contain sufficient information to warrant these assumptions. Here, I have written a statistical model whose features rely on previous theory and research. Based on this research I have assumed that habitual religious service attendance is causally antecedent to religious identification and other indicators of religiosity. For this reason I have not included these indicators because to do so would introduce post-treatment confounding. I have also assumed that demographic indicators such as age, education, gender, SES, and country of residence may affect both church attendance and anxiety. All causal assumptions cannot generally be derived entirely from the data alone, it is important to remember that expert knowledge is constantly undergoing revision. Therefore: **state your assumptions explicitly and graphically** and **when expert knowledge is in doubt, perform multiple analyses** (see Bulbulia et al. (2021)).

Third, I have presented a workflow based on Hern'an and Robins's strategy of specifying and emulating a "target trial" (Hernán et al., 2016). In this study, the target trial that I have specified falls well short of any experimental standard. For example, no experiment would enlist people after they have received their treatment. However, by imagining a hypothetical intervention that could test a causal hypothesis we can better appreciate both the strengths and the shortcomings of an observational study. It is easy to focus on the limitations of observational data such as the MARP dataset. However, cultural data need not be discarded because they were not obtained from experiments or because repeated measures are lacking. Rather, what we need is clarity about our causal assumptions. **Clarify your casual question by stating and emulating**

a target trial. How your target trial fails will clarify your assumptions. Remain open to the possibility your question cannot be answered, either because the question is too vague or because the data cannot resolve it.

Fourth, in observational research although we cannot generally guarantee there is no unmeasured confounding we can usually perform a sensitivity analysis to investigate robustness of a causal estimate to unmeasured confounding. The EValues package in R is freely available. Moreover there is an online calculator at: https:www.evalue-calculator.com. **Perform a sensitivity analysis to evaluate the robustness of your causal effect estimate.**

Fifth, causal inference is the holy grail of scientific research. However in the psychological sciences, there has been relatively little discussion about improving workflows for causal inference (although see: Rohrer, 2018). I have argued that the use of non-causal language such as "prediction" and "predicts" may be misleading. A prediction is not a conservative causal estimate. Absent a causal framework, a prediction has no causal implication at all. For this reason, researchers who limit observational inference to "prediction" invite misleading interpretations of their research. They also obscure their theoretical interests, which are generally causal (Hernán, 2018). Help is available. **Do not settle for reporting a "prediction" from a statistical "association." Instead, report your attempt to estimate an average causal effect, and perform a sensitivity analysis.**

A final comment. The development of powerful frameworks for causal inference during the past several decades has resulting in "causal revolution" across many sciences (Pearl & Mackenzie, 2018). During this same period, however, most observational research in psychology has operated in the absence of any explicit causal framework. The widespread adoption of causality-naive practices in observational psychology has resulted in what might be deemed a "causality crisis." The causality crisis in observational psychology has parallels to the replication crisis in experimental psychology – causal naivety casts nearly all previous findings into doubt. Unlike the replication crisis in experimental psychology, however, the magnitude of the causality crisis remains to be fully appreciated. For those unfamiliar with the potential outcomes framework, I hope this study will stimulate your interest in developing a more systematic and rigorous approach to causal inference.

Acknowledgements

I am grateful to the MARP team and their funders for obtaining the data used in this project, and to Templeton Religion Trust grant TRT0196 for support of my time. I am grateful to Dr. Usman Afzali, Professor John Shaver, and an anonymous reviewer for feedback.

Disclosure statement

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This research was supported by Templeton Religion Trust Grant 0196 and Royal Society of New Zealand Grants 19-UOO-090 and Marsden 3721245. A complete description of the data (including missing data) and statistical analysis has been posted to https://github.com/go-bayes/many_analysts.

Notes on contributor

Joseph A. Bulbulia is a professor of Psychology and Director of the Centre for Applied Cross-Cultural Research at Victoria University in Wellington, New Zealand. His research focuses on the evolutionary and longitudinal study of religion, cooperation, and psychological health. He is one of four senior managers for the New Zealand Attitudes and Values Study, a researcher for the Max Plank Institute for the Science of Human History.

References

Bulbulia, J., Schjoedt, U., Shaver, J. H., Sosis, R., & Wildman, W. J. (2021). Causal inference in regression: advice to authors. *Religion Brain & Behaviour*, *11*(4), 353–360.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*(6), 656–664.

Ejova, A., Milojev, P., Worthington Jr, E. L., Sibley, C. G., & Bulbulia, J. (2020). Church attendance buffers against longer-term mental distress. *Religion, Brain & Behavior*, *11*(2), 123–138.

Gelman, A., & Su, Y.-S. (2020). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* Retrieved from https://CRAN.R-project.org/package=arm

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, *15*(2), 1–11.

Hernán, M. A. (2018). The c-word: The more we discuss it, the less dirty it sounds. *American Journal of Public Health*, *108*(5), 625–626. Retrieved from https://doi.org/10.2105/AJPH.2018.304392

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764.

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC. Retrieved from https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, *79*, 70–75.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. Retrieved from http://www.jstatsoft.org/v45/i07/

Lang, M., Kratky, J., Shaver, J. H., Jerotijević, D., & Xygalatas, D. (2015). Effects of anxiety on spontaneous ritualized behavior. *Current Biology*, *25*(14), 1892–1897.

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* Retrieved from `https://CRAN.R-project.org/package=emmeans`

Malinowski, B. (2014). *Magic, science and religion and other essays.* Read Books Ltd.

Mathur, M. B., Ding, P., Riddell, C. A., & VanderWeele, T. J. (2018). Website and R package for computing E-values. *Epidemiology (Cambridge, Mass.)*, *29*(5), e45.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan.* CRC press.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect.* Basic books.

Pearl, J., & others. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Robins, J. (1986). 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*, *7*(9-12), 1393–1512.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42.

Rubin, D. (1976, December). Inference and missing data. *Biometrika*, *63*(3), 581–592. Retrieved from `https://doi.org/10.1093/biomet/63.3.581`

Sosis, R. (2007). Psalms for safety: Magico-religious responses to threats of terror. *Current Anthropology*, *48*(6), 903–911.

Tierney, N., Cook, D., McBain, M., & Fay, C. (2021). *naniar: Data structures, summaries, and visualisations for missing data.* Retrieved from `https://CRAN.R-project.org/package=naniar`

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction.* Oxford University Press.

VanderWeele, T. J. (2017a). Causal effects of religious service attendance? *Social Psychiatry and Psychiatric Epidemiology*, *52*(11), 1331–1336.

VanderWeele, T. J. (2017b). Religion and health: A synthesis. In *Spirituality and religion within the culture of medicine: From evidence to practice.* (pp. 357–401). New York, NY, US: Oxford University Press.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, *167*(4), 268–274.

VanderWeele, T. J., Mathur, M. B., & Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: a new template for empirical studies. *Statistical Science*, *35*(3), 437–466.

VanderWeele, T. J., & Robinson, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, *25*(4), 473–484.

Wal, W. M. v. d., & Geskus, R. B. (2011). ipw: An R package for inverse probability weighting. *Journal of Statistical Software*, *43*(13), 1–23. Retrieved from `http://www.jstatsoft.org/v43/i13/`

Westreich, D., & Greenland, S. (2013). The Table 2 Fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, *177*(4), 292–298.

Appendices

Appendix A. Priors for Bayesian estimation

A.1. Priors for monotonic predictor model

In an ordinal predictor model where there are R indicator responses, there are $C = R-1, c = 1 \dots C$ thresholds to estimate the ordered responses predictors. In this model, b^c locates the direction and size of the effect. That is b^c locates the expected average difference between two adjacent categories of the ordinal predictor. Such models also require the parameter ζ^c, which locates the normalised distances between consecutive predictor categories, producing the shape of each monotonic effect:

$$g(y_i) = \alpha + b^c \zeta^c x_i$$

For the monotonic model I used the following strong priors:

$$\begin{aligned}
y_{ij} &\sim \text{Normal}(\mu_{ij}) \\
\text{Normal}(\mu_{ij}) &\sim \boldsymbol{\alpha} + \boldsymbol{b\zeta} \\
\boldsymbol{\alpha} &\sim \alpha_0 + \alpha_j \\
\alpha_0 &\sim \text{Student T}(3, 0, .25) \\
\boldsymbol{\beta} &\sim \boldsymbol{b} + \beta_j \\
\boldsymbol{b} &\sim\sim \text{Normal}(-.05, .1) \\
\boldsymbol{\zeta} &\sim \text{Dirichlet}(1) \\
\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{S} \right) \\
\boldsymbol{S} &= \begin{pmatrix} \sigma_{\alpha_j} & 0 \\ 0 & \sigma_{\beta_j} \end{pmatrix} \boldsymbol{R} \begin{pmatrix} \sigma_{\alpha_j} & 0 \\ 0 & \sigma_{\beta_j} \end{pmatrix} \\
\sigma_{\alpha_j} &\sim \text{Student T}(3, 0, 2.5) \\
\sigma_{\beta_j} &\sim \text{Student T}(3, 0, 2.5) \\
\boldsymbol{R} &\sim \text{LKJcorr}(1)
\end{aligned}$$

Here, the vector $\boldsymbol{\zeta}$ contains six coefficients for the $c = 1 \dots 6$ simplex parameters estimated for the thresholds.

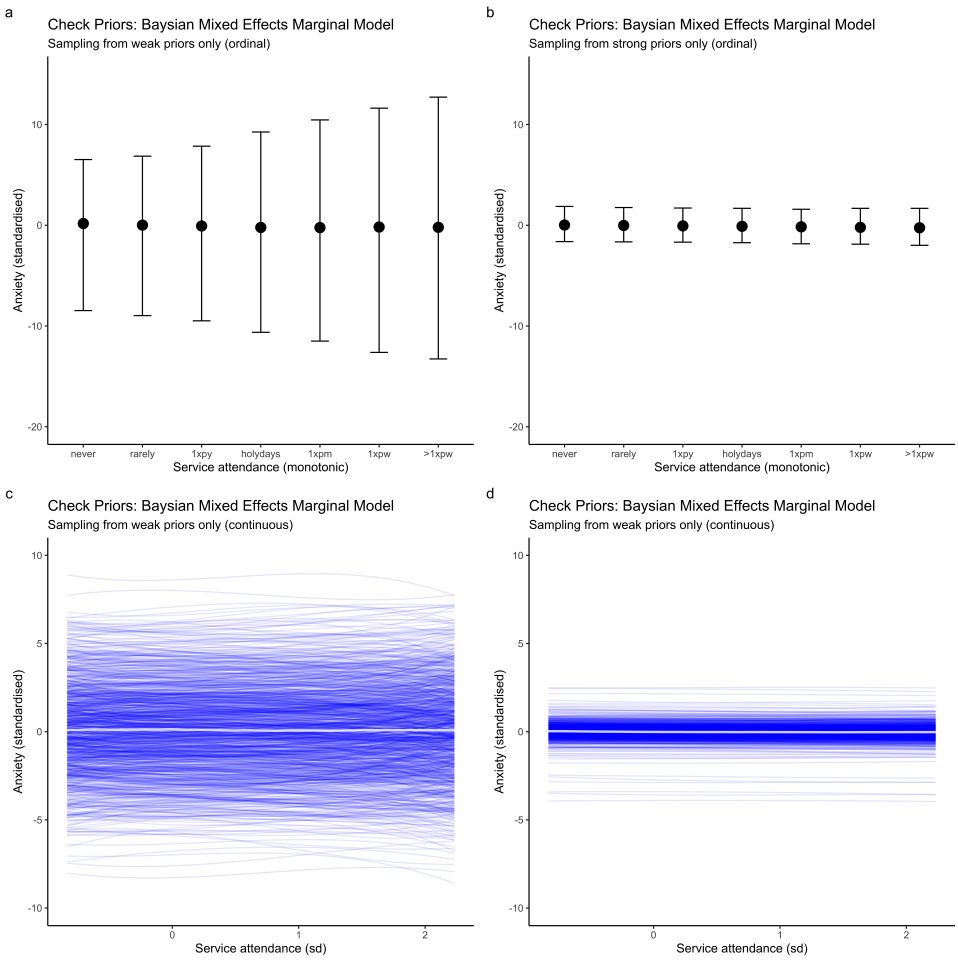
A.2. Priors for continuous predictor model

The strong priors for the continuous predictor model were as follows:

$$
\begin{aligned}
y_{ij} &\sim \text{Normal}(\mu_{ij}) \\
\text{Normal}(\mu_{ij}) &\sim \boldsymbol{\alpha} + \boldsymbol{\beta} \\
\boldsymbol{\alpha} &\sim \alpha_0 + \alpha_j \\
\alpha_0 &\sim \text{Student t}(3, 0, .25) \\
\boldsymbol{\beta} &\sim \boldsymbol{b} + \beta_j \\
\boldsymbol{b} &\sim \text{Normal}(-.05, .1) \\
\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{S}\right) \\
\boldsymbol{S} &= \begin{pmatrix} \sigma_{\alpha_j} & 0 \\ 0 & \sigma_{\beta_j} \end{pmatrix} \boldsymbol{R} \begin{pmatrix} \sigma_{\alpha_j} & 0 \\ 0 & \sigma_{\beta_j} \end{pmatrix} \\
\sigma_{\alpha_j} &\sim \text{Student T}(3, 0, 2.5) \\
\sigma_{\beta_j} &\sim \text{Student T}(3, 0, 2.5) \\
\boldsymbol{R} &\sim \text{LKJcorr}(1)
\end{aligned}
$$

As indicated in Figure A1, choice of priors did not affect the location estimates in either the monotonic or continuous model of religious service attendance on anxiety.

Figure A1. Panel (a) presents posterior predictions sampling only from the weakly default priors for ordinal church attendance model. These priors allow location estimates for each monotonic location with up to 10 standard deviations width to either side of zero. Panel (b) presents posterior predictions sampling only from the stronger priors I used for the ordinal religious attendance model. Panel (c) presents posterior predictions sampling only from the weakly default priors for the continuous religious attendance model. Here too, the default priors allow location estimates of up to 10 standard deviations to either side of zero across the range of church attendance responses. Panel (b) presents posterior predictions sampling only from the stronger priors I used for the continuous church attendance model. Estimates from weakly and strongly informative models in both the monotonic and continuous models converged to within MCMC error. This result indicates the posteriors parameter estimates were informed by the data, not by the priors.



Appendix B. Monotonic and continuous models of religious service: no difference.

Estimates for the monotonic model were:

Table B1. Parameter estimates for the monotonic effect model of religious service on anxiety

Parameter	Component	Median	CI	CI_low	CI_high	pd	Rhat
Intercept	conditional	0.044	0.95	-0.043	0.139	0.834	1.006
bsp_ord	conditional	-0.031	0.95	-0.054	-0.01	0.997	1.004
simo_ord1[1]	simplex	0.27	0.95	0.133	0.411	1	1
simo_ord1[2]	simplex	0.139	0.95	0.04	0.262	1	1
simo_ord1[3]	simplex	0.104	0.95	0.028	0.206	1	1.001
simo_ord1[4]	simplex	0.11	0.95	0.036	0.206	1	0.999
simo_ord1[5]	simplex	0.159	0.95	0.053	0.278	1	0.999
simo_ord1[6]	simplex	0.185	0.95	0.06	0.329	1	1
sigma	sigma	0.971	0.95	0.958	0.983	1	0.999

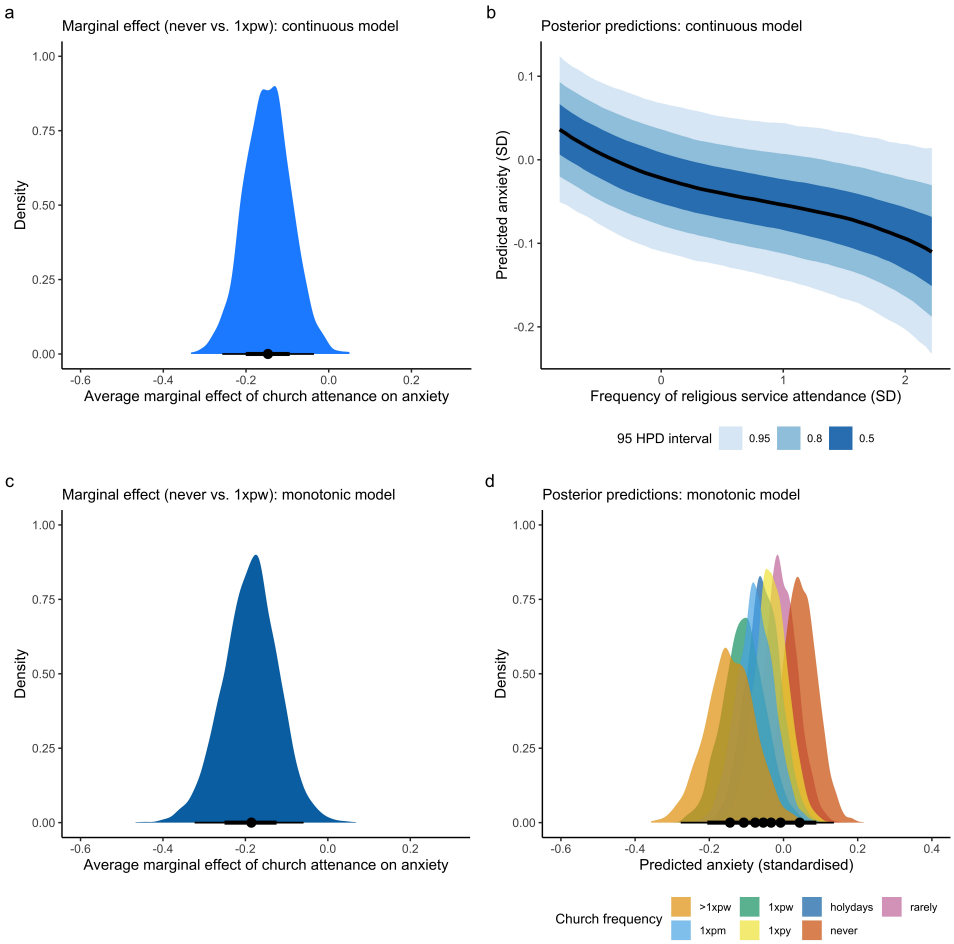
Estimates for the continuous model were:

Table B2. Parameter estimates for continuous (spline) model of religious service on anxiety

Parameter	Component	Median	CI	CI_low	CI_high	pd	Rhat
Intercept	conditional	0.036	0.95	-0.053	0.12	0.791	1.011
ch_s1	conditional	-0.104	0.95	-0.218	0.011	0.958	1
ch_s2	conditional	-0.065	0.95	-0.202	0.061	0.842	1.002
ch_s3	conditional	-0.147	0.95	-0.261	-0.041	0.996	1.004
sigma	sigma	0.971	0.95	0.958	0.984	1	1

As indicated in FigureB1, marginal effects estimates for the monotonic religious service attendance model and the continuous religious service attendance model do not lead to practical difference in the causal inference.

Figure B1. Panels (a) and (b) graph the marginal effect estimates and marginal predictions for a continuous model of church attendance, estimated with splines. In Panel (a) the marginal contrast is taken from the extremes of the range attendance range: more than once per week vs never. Panels (c) and (d) graph the marginal effect estimates and marginal predictions for a monotonic predictor model of church attendance, where again the contrast is taken from the extremes of the range. There is no practical difference between these model choices. I select the monotonic model because the categories of religious service responses are ordinal choices.

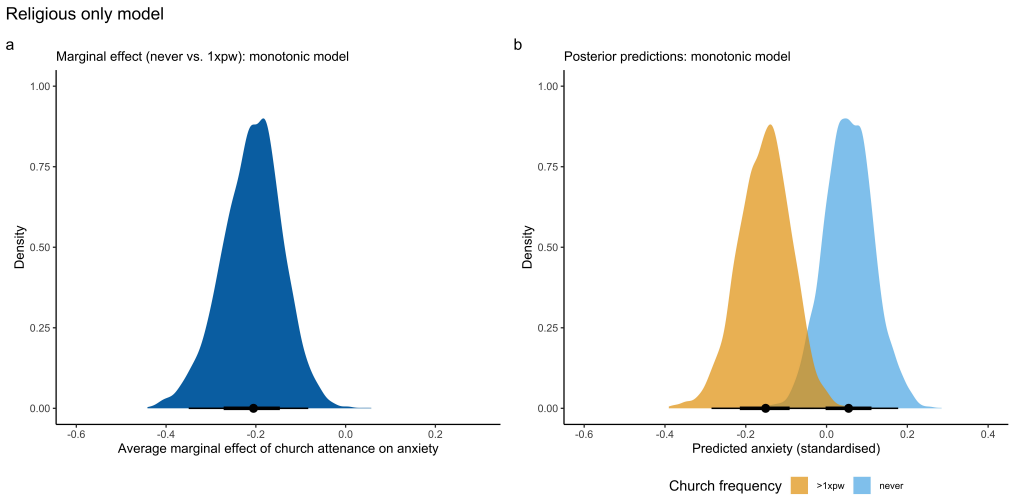


Appendix C. Religious-only sub-sample

Vanderweele observes that positivity might not hold among those who do not identify as religious. I have estimated causal effect estimates over the entire sample because I observe that over half of the non-religious sample attend religious service. However, I do not believe that causal analysis should be restricted to a single option. Some may find it useful to quantify the causal effects in the religious-only sample.

Using the identical procedure to that of the full sample estimate (Bayesian mononotic mixed effect model) we obtain an average marginal effect of (ATE = -0.21, HPD[0.34, -0.08]. we estimate an E-value of 1.71, with a lower bound of 1.35, indicating a stronger causal effect of religious service attendance in diminishing anxiety in the only-religious sample (see: Figure C1).

Figure C1. Causal effect of religious service on anxiety in the religious-only sample



Appendix D. Marginal Structural Model

Our counterfactual contrast of interest is the difference in the expected average response in the (entire) population under exposure and the expected average response in the (entire) population under no exposure. Here, we compare the non-attending exposure with the attending-more-than-once-per-week exposure.

It is essential to realise that estimating a statistical difference in the observed associations in our data:

$$E[Y|A = 1] - E[Y|A = 0]$$

does not provide the contrast that we require for causal inference, in which the whole population does, or does not, receive the exposure:

$$E[Y^{a=1}] - E[Y^{a=0}]$$

Rather than estimating the expectation of the observed exposures, as in standard regression analysis, a marginal structural model estimates:

$$E[Y^a] = \beta_0 + \beta_1 a$$

where:

$$E[Y^{a=0}] = \beta_0$$

and:

$$E[Y^{a=0}] - E[Y^{a=1}] = \beta_1 a$$

A marginal structural model, then, allows us to estimate a difference in average counterfactual expectations (or equivalently, an average difference in counterfactual expectations). The method requires satisfaction of the assumption that those who received an exposure are similar to those who did not. To estimate a marginal structural model, we weight responses by the inverse probability of exposure; that is, we use IPTW weighting. In the simulated population that we obtain from IPTW weighting: $E[Y|A] = \theta_0 + \theta_1 A$, where $\theta_1 A$ is equivalent to $\beta_1 a$ in the counterfactual model (Hernán & Robins, 2020).

Appendix E. Graphical methods for confounder control

There are four primary conditions for confounder control: (1) There is no unopened or "backdoor" path leading from exposure to outcomes: we block a backdoor path by conditioning on a variable along the path; (2) There is no intervening mediator between an exposure and an outcome has been conditioned on. Unless we are interested in mediation, we open a path by omitting this mediator from analysis; (3) There is no common outcome of the exposure and outcome that has been conditioned on. Conditioning on a "collider" will induce a spurious association that is removed by omitting the collider from the analysis; (4) No descendants of a collider have been conditioned on, which we address by omitting any descendants from the analysis.

Appendix F. Simulation of mediation confounding

The simulation (in R) reported below clarifies the bias of conditioning on a mediator where there is unmeasured mediator-outcome confounding. Specifically, A mediator affected by an exposure that is only associated with the outcome by a hidden confounder may attenuate the estimate of an actual causal effect for the exposure on the outcome. I present a DAG for such confounding in Figure 1.c. The solution to the problem is straightforward. Do not condition on a mediator unless your causal question pertains to mediation. And if your causal question pertains to mediation, it is essential to assess the robustness of the indirect effect to mediator outcome confounding. For advice on mediation see: VanderWeele (2015).

```
set.seed(123)

```{r}
sim_fun_A = function() {
  n <- 1000
  U <- rnorm(n, 1) # unmeasured distress
  A <- rnorm(n, 1) # religious service attendance,
  M <- rnorm(n , A + .2 * U)# religious identification
  Y <- rnorm(n , - A *.2 + U *.4) #  distress reduced by attendance.

# Religious id, caused by attendance + distress

# simulate data
simdat_A <- data.frame(
  M = M, # rel identification
  A = A, # rel service
  U = U, # distressing events (unmeasured)
  Y = Y )# outcome

sim_A <- lm(Y ~ A + M, data = simdat_A)
sim_A
}

# Replicate including the mediator
r_lm_A <- NA
r_lm_A = replicate(100, sim_fun_A(), simplify = FALSE )


# mediator attenuates true effect ~ .1 instead of ~.1
parameters::pool_parameters(r_lm_A)

## Repeat but with no mediator in model
set.seed(123)
sim_fun_B = function() {
  n <- 1000
  U <- rnorm(n, 1) # unmeasured distress
  A <- rnorm(n, 1) # religious service attendance,
  M <- rnorm(n , A + .2 * U)# religious identification
```
```

```
  Y <- rnorm(n , - A *.2 + .4 * U) #  distress reduced by attendance.

# Simulate data
simdat_B <- data.frame(
  M = M, # rel identification
  A = A, # rel service attendance
  U = U, # distressing events (unmeasured)
  Y = Y ) # outcome

# Only model exposure
sim_B <- lm(Y ~ A, data = simdat_B)
sim_B
}

# Replication 100xs
r_lm_B <- NA
r_lm_B = replicate(100, sim_fun_B(), simplify = FALSE )

# diminished recover true effect ~ .2
parameters::pool_parameters(r_lm_B)
```