ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2019)
182, *Part* 3, *pp.* 983–1002

Bayesian modelling for binary outcomes in the regression discontinuity design

Sara Geneletti

London School of Economics and Political Science, UK

and Federico Ricciardi, Aidan G. O'Keeffe and Gianluca Baio

University College London, UK

[Received October 2017. Revised January 2019]

Summary. The regression discontinuity (RD) design is a quasi-experimental design which emulates a randomized study by exploiting situations where treatment is assigned according to a continuous variable as is common in many drug treatment guidelines. The RD design literature focuses principally on continuous outcomes. We exploit the link between the RD design and instrumental variables to obtain an estimate for the causal risk ratio for the treated when the outcome is binary. Occasionally this risk ratio for the treated estimator can give negative lower confidence bounds. In the Bayesian framework we impose prior constraints that prevent this from happening. This is novel and cannot be easily reproduced in a frequentist framework. We compare our estimators with those based on estimating equation and generalized methods-of-moments methods. On the basis of extensive simulations our methods compare favourably with both methods and we apply our method to a real example to estimate the effect of statins on the probability of low density lipoprotein cholesterol levels reaching recommended levels.

Keywords: Bayesian inference; Binary outcomes; Causal inference; Instrumental variables; Prior constraints

1. Introduction

A regression discontinuity (RD) design is a quasi-experimental method for treatment effect estimation, which was introduced in the 1960s in Thistlethwaite and Campbell (1960) and is widely used in economics and related social sciences (Imbens and Lemieux, 2008) and more recently in the medical sciences (Geneletti *et al.*, 2015; Bor *et al.*, 2014; Smith *et al.*, 2015; Moscoe *et al.*, 2015; Linden *et al.*, 2006). The RD design has become of interest in the context of public health as it enables the use of routinely gathered medical data to evaluate the causal effects of drugs when these are prescribed according to well-defined decision rules. This can be very useful as government agencies such as the Federal Drug Administration in the USA and the National Institute for Health and Care Excellence (NIHCE) in the UK are increasingly issuing guidelines for drug prescription. Currently the fact that we can use the guidelines to estimate causal effects is an unintended side product of the guidelines. However, we can imagine a future where RD designs will be planned to understand where, in the range of a continuous health outcome, administering a drug is most beneficial to patients.

Address for correspondence: Sara Geneletti, Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: s.geneletti@lse.ac.uk

Unlike in most of the RD design literature, which focuses on continuous outcomes, our motivating example is the estimation of the effect of statins (a class of cholesterol lowering drugs) on the probability of low density lipoprotein (LDL) cholesterol levels reaching two separate targets as recommended by the NIHCE. We thus develop a Bayesian approach for binary outcomes, which are frequently of primary interest in healthcare contexts. Our model aims to estimate a risk ratio directly, rather than through the commonly used odds ratio approximation. These two measures are of course similar only in the case where the underlying event of interest has a low probability of occurrence. However, in several applications (including ours), the 'rare disease' assumption does not hold, which means that the output of standard methods (e.g. logistic regression) does not allow the assessment of the actual problem in a principled way.

Especially in the case where the guidelines are not strictly followed by practitioners (a situation termed 'fuzzy RD', which we describe formally in Section 2), the RD design naturally leads to an instrumental variable (IV) analysis. Indeed, causal effects from fuzzy RD designs are usually estimated by using methods from the IV literature for both continuous and binary outcomes (Lee and Lemieux, 2010; Angrist *et al.*, 1996; Clarke and Windmeijer, 2012). In particular, for a binary outcome, an IV-based multiplicative structural mean model (MSMM) is often used.

In the context of the IV-based MSMM, several estimators are available to estimate the risk ratio for the treated (RRT), which is a measure of the change in risk for those who received a treatment—the binary analogue of the effect of treatment on the treated (Clarke and Windmeijer, 2010, 2012; Hernan and Robins, 2006; Abadie, 2002). However, these are all developed under a frequentist approach to statistical inference. In this paper, we develop a Bayesian approach to the estimation of the RRT, within the context of a fuzzy RD design. The use of a Bayesian approach has several benefits when compared with frequentist methods. Firstly, we obtain the variances of our estimates from our posterior samples directly without having to use bootstrapping or other variance approximation approaches (e.g. the delta method). The Bayesian MSMM RRT estimator is very flexible as we can estimate its components by using a large number of models. Finally we can impose prior constraints on the MSMM estimator for the RRT as it is known to misbehave occasionally (Clarke and Windmeijer, 2010) in that the lower limits of 95% confidence intervals may be negative. Our prior constraints may prevent the posterior Markov chain Monte Carlo (MCMC) sample from dropping below zero. This is a novel implementation of prior constraints and, although possible, is more difficult to reproduce in a frequentist context. Using these prior constraints also helps to produce intuitive and stable uncertainty intervals. We compare our estimators with those based on the MSMM (Clarke *et al.*, 2015) and the logistic structural mean model (LSMM) (Vansteelandt *et al.*, 2011; Vansteelandt and Goetghebeur, 2003; van der Laan *et al.*, 2007) among others. In simulations our methods compare favourably with these frequentist estimators. This is especially true in contexts where confounding is strong.

An open question in the RD design literature is how small the distance—bandwidth—between the units and the threshold must be for the RD design to be valid (Imbens and Kalyanaraman, 2012; Calonico *et al.*, 2015). We do not directly tackle this question here. Instead we adopt the 'local' regression approach (Imbens and Lemieux, 2008) for four bandwidths and assess the sensitivity of our results to these changes.

The paper is organized as follows: in Section 2 we briefly describe the RD design and introduce our example. Section 2.3 lays out our notation and assumptions. In Section 3 we describe in detail our models as well as giving a short overview of the most commonly used competing methods. We present the results of a simulation study in Section 4. Section 5 follows with the results of the real application. We finish with some discussion in Section 6. Additional analyses and plots are available in the on-line supplementary material.

2. Background and example

An analysis based on the RD design is appropriate for public health interventions that are implemented according to pre-established guidelines. Specifically, when a decision rule is based on whether a continuous variable exceeds a certain threshold, it becomes possible to implement the RD design. In our running example we use data from the Health Improvement Network (THIN), which is a UK data set which consists of routine patient data from clinical practice. We investigate the effect of prescribing statins, a class of cholesterol lowering drugs, on reaching the NIHCE recommended LDL cholesterol targets of below 2 or 3 mmol l^{-1} for healthy and high risk patients respectively. Between 2008 and 2014 NIHCE guidelines recommended that statins should be prescribed to patients whose 10-year cardiovascular risk score exceeded 20% provided that they had no history of cardiovascular disease. If we are willing to assume that individuals just above and below the 20% threshold are exchangeable—an essential condition for causal inference in the RD design—then we have a quasi-randomized trial with those just below the threshold randomized to the no-treatment 'arm' and those just above randomized to the treatment arm. Thus any jump or discontinuity in the values of the outcome across the threshold can be interpreted as a local causal effect or risk ratio in the case of binary outcomes.

There are two types of RD design. The first is termed *sharp* and refers to the situation where the guidelines are adhered to strictly. In our application we would encounter a sharp RD design if all the doctors complied with the NIHCE guidelines and prescribed statins exclusively to patients with a 10-year cardiovascular risk score above 20%. In practice this is often not so, and our application makes no exception. There is some 'contamination' whereby individuals whose risk score lies below the threshold are prescribed statins whereas others whose risk score lies above the threshold receive no prescription. This situation is termed a *fuzzy* RD design. These concepts are easier to understand with the help of plots as shown in the next section.

2.1. Data

The data that we consider are a subset of THIN data. The THIN data are primary care data for over 500 practices in the UK and include a large number of individual patient, diagnostic and prescription information. We focus on a subset of 1386 male patients between the ages of 50 and 70 years who did not smoke or have diabetes in the year 2008.

2.2. Exploratory plots for the regression discontinuity design

We present examples of sharp and fuzzy designs as well as both continuous and binary outcomes from our application. In all figures the circles are individuals who do not have statin prescriptions whereas crosses represent those who do have statin prescriptions. We plot the mean of the outcome (continuous or binary) within bins of the risk score (x-axis) against the risk score and fit a cubic spline. The reason for using this plot is that in particular when the design is fuzzy the spline will show a jump at the threshold indicating the presence of a discontinuity and thus a potential causal effect around the threshold.

Fig. 1(a) shows data that are obtained from the real data by removing all the individuals who have a statin prescription below the threshold and do not have a statin prescription above the threshold—a sharp design situation. The y-axis is a continuous measure of LDL cholesterol levels in millimoles per litre—the outcome. The x-axis is the risk score centred at 0.2—the so-called running variable. There is a small but noticeable downward jump at the threshold. This can be seen as evidence of a discontinuity around the threshold and the jump can be seen as representing the effect of the statin prescription. Fig. 1(b) is a similar plot, but with the raw probability of treatment above and below the threshold on the y-axis. Again there is a clear

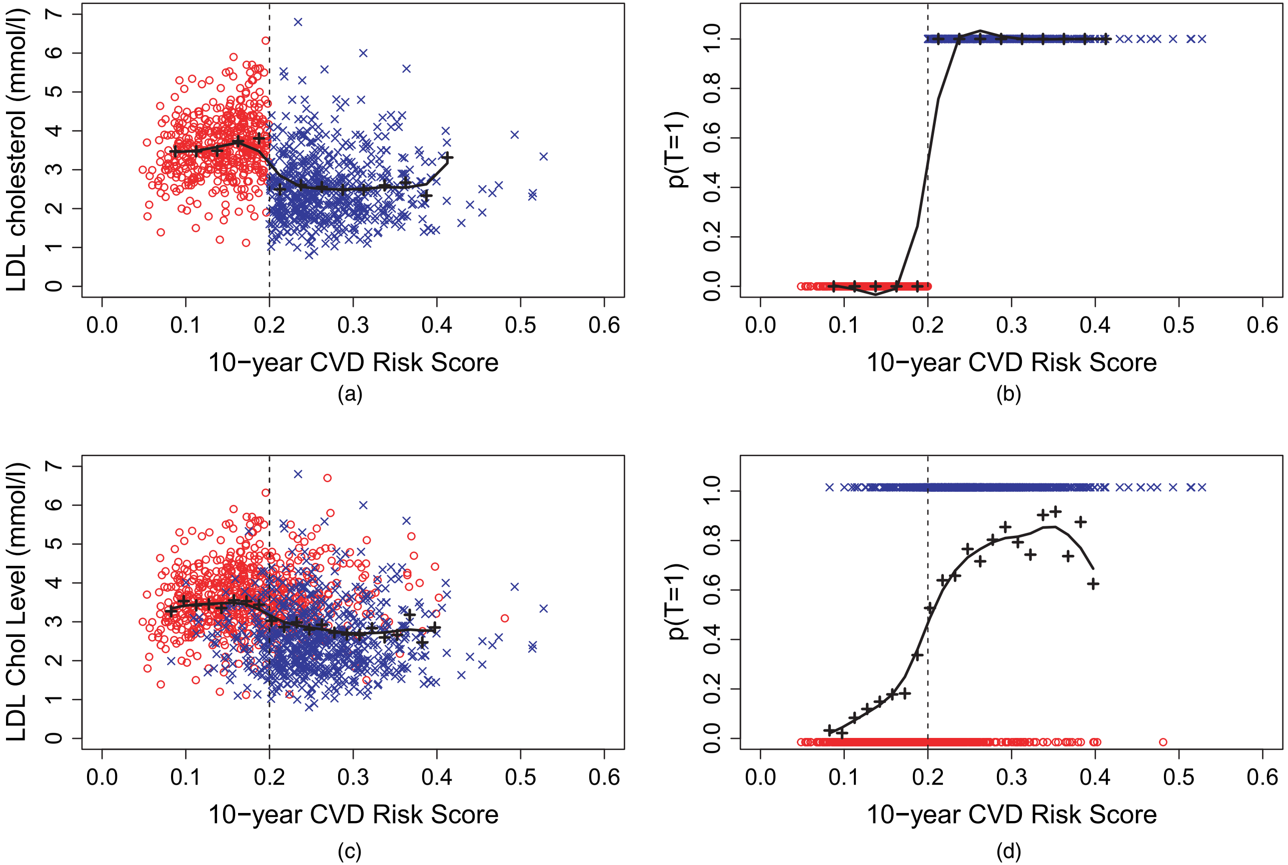


Fig. 1. (a), (b) The sharp design and (c), (d) the real (fuzzy) design (in both cases means within bins and corresponding fitted cubic spline are overlayed): (a), (c) risk *versus* LDL cholesterol level; (b), (d) risk score *versus* raw probability of treatment

jump from 0 to 1 in probability of receiving a statin prescription. Figs 1(c) and 1(d) are the real data from our application. It is clear that the design is fuzzy as there are circles above the threshold and crosses below. Despite the fuzziness there is still a small downward jump at the threshold as shown in Fig. 1(c). Fig. 1(d) is the corresponding raw probability plot. The increase in probability is more gradual but there is a distinct jump at the threshold. On the basis of these plots we would be happy to proceed to an RD design analysis of the data with LDL cholesterol level as a continuous outcome variable (Geneletti *et al.*, 2015).

We now show the plots for the two binary outcomes of interest: LDL cholesterol dropping below 2 mmol l^{-1} for high risk patients and below 3 mmol l^{-1} for healthy patients. High risk patients were those with one additional risk factor such as smoking or diabetes (see the on-line supplementary materials for further details). Looking at Fig. 2, a small jump can be seen for LDL cholesterol level going below 3 mmol l^{-1} in healthy patients (Fig. 2(a)) and a larger jump for the probability of treatment (Fig. 2(b)). For the high risk patients there is no discernible jump in LDL cholesterol level going below 2 mmol l^{-1} (Fig. 2(c)): just a steady increase. However, there is evidence of a jump in the probability of being treated for such patients (Fig. 2(d)). Taken together, there appears to be sufficient support for an RD design analysis for healthy and high risk patients. The lack of a clear jump in the outcome plot might be because high risk individuals are more likely to have received statins for other conditions (e.g. high blood pressure) before being diagnosed with high cholesterol. If we had failed to see evidence of a jump in both outcome and treatment plots then we would not have proceeded with an RD design analysis. This would indicate that the general practitioners (GPs) are not complying with the guidelines sufficiently to be able to make reliable inference.

2.3. Assumptions and notation

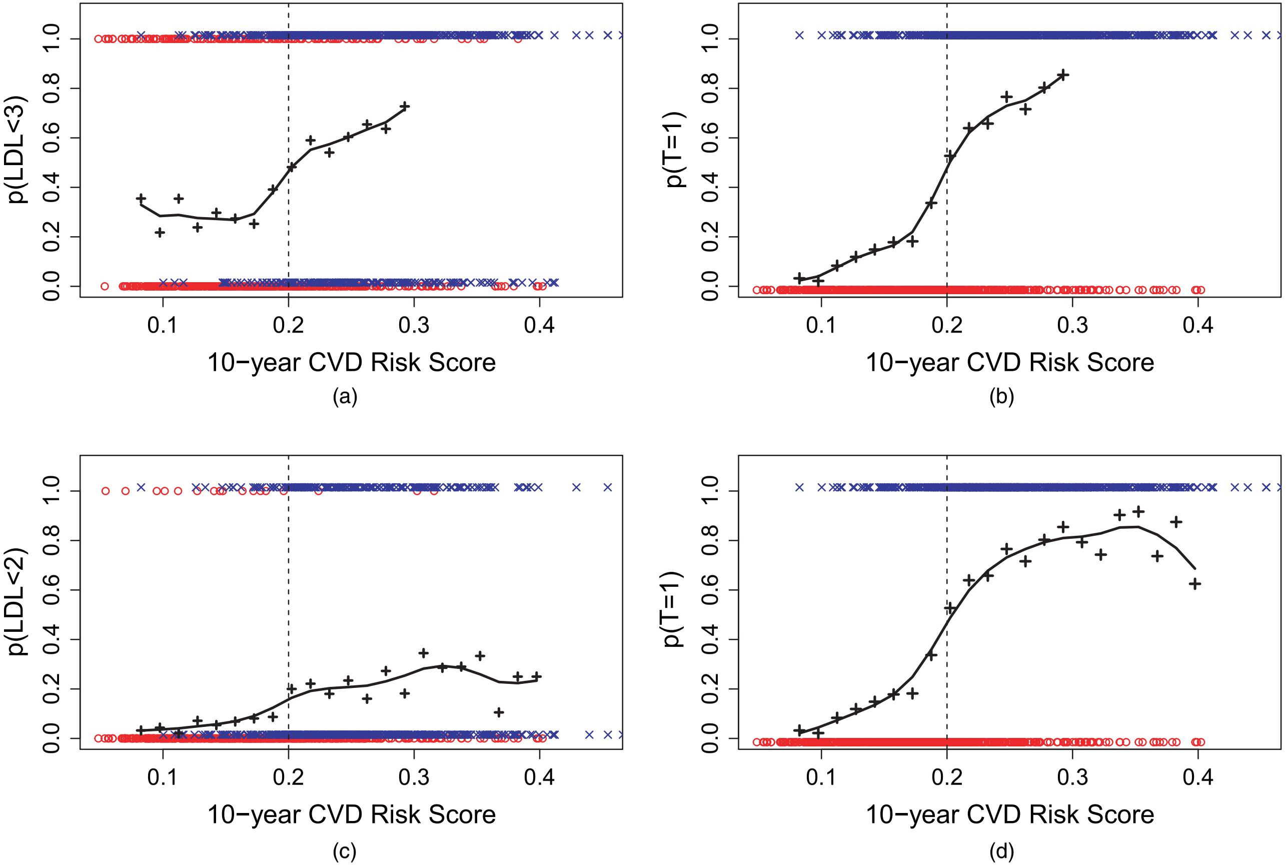
For the RD design to be appropriate, some formal assumptions also must hold. These assumptions can be formulated in different ways (Hahn *et al.*, 2001; Imbens and Lemieux, 2008; van der Klaauw, 2008; Lee, 2008) and we give a brief overview of them, as described in more detail in Geneletti *et al.* (2015). In the binary outcome case we need to make additional assumptions to identify the RRT (Didelez *et al.*, 2010; Hernan and Robins, 2006). We express our assumptions in the language of conditional independence following Dawid (1979). We refer to our example to anchor the theoretical arguments, although generalizing to other contexts is straightforward.

Let X be the 10-year cardiovascular risk score. The threshold indicator Z represents the NIHCE treatment guidelines and is such that $Z=1$ if $X \geqslant 0.2$ and $Z=0$ if $X<0.2$. Let T represent statin prescription (not whether the patient takes the treatment); $T=1$ means that statins are prescribed and $T=0$ means that they are not. Also, let $\mathbf{C}=\{\mathbf{O} \cup \mathbf{U}\}$ be the set of relevant pretreatment confounders, where $\mathbf{O}$ and $\mathbf{U}$ indicate fully observed and partially or fully unobserved variables respectively. Confounders of interest include age, ethnicity and socio-economic status where typically age and ethnicity are observed but socio-economic status is not. Y is the binary outcome variable where $Y=1$ if LDL cholesterol is below 2 (or 3) mmol l^{-1} and $y=0$ otherwise.

To reflect the fact that these assumptions are valid only around the threshold, we assume throughout the paper an additional conditioning on $X \in [0.2, 0.2+h]$ if above the threshold and $X \in [0.2-h, 0.2]$ below the threshold for some suitably small h. We do not explicitly write this conditioning except where necessary.

2.4. Regression discontinuity design assumptions

The first three RD design assumptions are essentially IV assumptions—where Z is the



(a)

(b)

(c)

(d)

Fig. 2. (a), (b) Mean and probability plots for the healthy subpopulation with outcome LDL cholesterol going below 3 mmol l^{-1} and (c), (d) mean and probability plots for the high risk subpopulation with outcome LDL cholesterol going below 2 mmol l^{-1}

instrument—whereas the fourth is specific to the RD design. See Geneletti *et al.* (2015) for more details and interpretation.

Assumption 1 (association of treatment with threshold indicator: $Z \not\perp\!\!\!\perp T$). Statin prescription (T) must be associated with the NIHCE treatment guidelines (Z).

Assumption 2 (independence of guidelines: $Z \perp\!\!\!\perp \mathbf{C} \mid (X)$). The NIHCE guidelines (Z) cannot depend on any of the characteristics of the patient (C) excluding X. A weaker version (i.e. within strata of $\mathbf{O}$) is $Z \perp\!\!\!\perp \mathbf{U} \mid (X, \mathbf{O})$. We can think of this as the RD design applied within strata of the observed confounders, e.g. by considering statin prescription for men only.

Assumption 3 (unconfoundedness: $Y \perp\!\!\!\perp Z \mid (T, \mathbf{C}, X)$). For the RD design to be a valid randomization device, whether the LDL cholesterol level (Y) is below 2 (or 3) mmol l^{-1} must be independent of the NIHCE guidelines (Z) conditionally on the other variables (Lee, 2008).

Assumption 4 (continuity). The expectation of the outcome $E(Y|X, C, Z, T=t)$ is continuous at $X = x_0$ for $t \in \{0, 1\}$.

2.4.1 Weak instruments

Broadly speaking when the threshold indicator Z is a poor predictor of the treatment T, i.e. the correlation between them is low, Z becomes a weak instrument. Typically the ‘fuzzier’ the RD design, the weaker the IV is and the smaller the bandwidth (and thus the sample size) the more the weakness of the IV becomes problematic. A weak IV can lead to causal effect estimates that are overly biased towards the unadjusted observational effect (Burgess and Thompson, 2012). It can also lead to a lack of identification for a class of semiparametric models, e.g. the generalized method of moments (Clarke and Windmeijer, 2010; Burgess *et al.*, 2014).

In our simulations (Section 4) we consider two levels of instrument strength which we term weak and strong. Our definition of IV strength does not correspond directly to that in Stock and Yogo (2005). However, their definitions are based on the linear case which do not have an obvious translation into the binary case that we are tackling. Thus we propose an ‘operational definition’ that is unique to the RD design based on looking at plots like those in Fig. 2. If there is no indication of a jump—in either outcome or treatment plots—we do not proceed with an RD design analysis. When we say that an IV is weak in our simulations we mean that a jump can still be discerned in the plots but that any additional weakening of the T–X-relationship results in no visible jump. In our analysis (Section 5.1) we perform a number of weak IV tests to ensure that the threshold indicator is not weak (Stock and Yogo, 2005; Burgess *et al.*, 2014). We appreciate that these may not be exactly suited to binary outcomes but they can be informative.

2.5. Risk ratio for the treated estimator and associated assumptions

We estimate risk ratios in our analysis as these are of primary interest and most suited to our method of analysis. In the epidemiological literature odds ratios are usually estimated as they are easily obtained from logistic regressions. However, simple estimators of causal odds ratios are typically more biased than risk ratio estimators (Didelez *et al.*, 2010), and in our case the rare disease assumption does not hold, making the odds ratio approximation less than ideal.

We focus on estimating the RRT, which is defined as follows:

$$\log\left\{\frac{E_1(Y \mid Z, T=t)}{E_0(Y \mid Z, T=t)}\right\} = \rho_t \tag{1}$$

where E_t is the expectation under the interventional regime where $T=t$. Following Geneletti and Dawid (2010) we can think of ρ_1 (where conditioning on $T=1$ indicates the effect on the treated) as a comparison between the GP–patient pairs randomized to statins *versus* those randomized to no treatment for those GP–patient pairs where the GP wanted to treat the patients. Although equation (1) is written in terms of unobserved interventional regimes ρ_1 can be identified by using an observational RD design when Y, Z and T are binary if in addition to assumptions 1–4 the following assumptions hold (Didelez *et al.*, 2010; Clarke *et al.*, 2015).

Assumption 5 (log-linear in t). $\log\{E_1(Y \mid T=t, Z=z)\} - \log\{E_0(Y \mid T=t, Z=z)\}$ is linear in the treatment.

Assumption 6 (no T–Z-interaction in Y on the multiplicative scale). This assumption is known as the no effect modification assumption.

Assumption 5 is trivial when Y, Z and T are binary: however, we list it here for completeness. Assumption 6 requires that whether the LDL cholesterol level is below 2 (or 3) mmol l^{-1} does not depend on an interaction term between T (statin prescription) and Z (whether the risk score exceeds 20% on the log-scale). If there were an interaction term it would mean that the GPs above and below the threshold would be different with respect to their ability to predict the outcome. As we are looking at individuals around the threshold whom we already consider to be exchangeable in some way, we are willing to assume that there is no interaction. When assumptions 1–4 and 5 and 6 hold, we can directly apply estimators that have been derived in the binary IVs literature in the context of the RD design. Thus we obtain the following formula (Hernan and Robins, 2006; Clarke and Windmeijer, 2010):

$$\widehat{\exp(\rho_1)} = \mathrm{RRT} = 1 - \frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(Y\bar{T} \mid Z=1) - E(Y\bar{T} \mid Z=0)}, \tag{2}$$

where $\bar{T} = 1 - T$. Equivalent expressions and details of their derivation and interpretation can be found in Didelez *et al.* (2010), Abadie (2003) and Angrist (2001). Note that equation (2) is based on an approximation of the true causal model. Here, the term 'true causal model' refers to the model from which data originate. Since the true model is assumed to be log-linear in t, this implies that the RRT estimator satisfies

$$E(Y|T=t, Z) = \exp(\alpha + \beta t).$$

3. Models

The models in Sections 3.2 and 3.3 are embedded in a Bayesian framework, and code that was used to run the models by using JAGS (Plummer, 2003) is available in the on-line supplementary material. We first obtain a full posterior MCMC sample for each of the relevant parameters in the models described in Section 3.2 and combine these to induce a posterior sample for the RRT. When prior constraints are added in Section 3.3, we sample the RRT directly. From the posterior samples we easily obtain variances and interval estimates without having to rely on bootstrap methods or asymptotic arguments, as is required with the frequentist estimators.

We note here that our focus is on obtaining the posterior distribution of the RRT, rather than on the joint posterior of the parameters that are used to estimate the RRT (i.e. the posterior distributions for $E(Y|Z)$ and $E(Y\bar{T}|Z)$, constructed by using a full specification of the joint distribution for Y, T and Z and the confounders U). Naturally, our chosen model relies on fewer parametric assumptions when compared with a fully joint model. In addition, our models are, in general, easier to fit and interpret by using standard MCMC algorithms.

We present some possible models to estimate the components. We mostly use the same type of model in the numerator and the denominator. However, we do mix different models in the numerator and denominator where we consider this necessary. Generally speaking we write the estimates for the RRT as $\text{RRT}_{\text{num.denom}}$ where 'num' indicates the form of the numerator and 'denom' the denominator in the fraction in equation (2).

3.1. Interaction versus product models

The denominator of the fraction in the expression for the RRT (henceforth only 'the denominator') is given by $E(Y\bar{T} \mid Z=1) - E(Y\bar{T} \mid Z=0)$. We can break up the individual terms further as follows:

$$E(Y\bar{T} \mid Z=z) = E(Y \mid \bar{T}, Z=z) E(\bar{T} \mid Z=z). \tag{3}$$

In our analysis, we produce estimates for both of these models. We call *interaction* models those which use the formulation on the left-hand side of equation (3) and *product* models those that use the formulation on the right-hand side. Our motivation for including analyses with the product of two conditional probabilities as in equation (3) is that the data for the product term are sparse (see the on-line supplementary material). By using Y alone this is mitigated. We also consider zero-inflated Poisson regression models (Lambert, 1992) to address this but the results are not substantially different.

3.2. Poisson regression models

In the first set of models that we considered, all components in the RRT are estimated by using Poisson regression models, in line with assumption 5. Assuming a Poisson sampling distribution for the outcome naturally accounts for the fact that the observed discrete counts are positive, as well as the log-linearity of the mean and the interpretation of the treatment effect as a risk ratio. One potential limitation is using the Poisson model, which implies equality of the mean and variance. However, particularly within the Bayesian approach, this can be easily overcome by simply including structured ('random') effects, to account for overdispersion.

It is easy to verify that if the same parametric form can be assumed to hold for experimental and observational regimes then a log-linear relationship in t follows for each of the components of equation (2).

We set $X^* = X - 0.2$ and fit Poisson regressions in both the numerator and the denominator:

$$\text{numerator} := \begin{cases} y_{i_z} \sim \text{Poisson}(\mu_{i_z}), \\ \log(\mu_{i_z}) = \alpha_z + \beta_z x^*_{i_z}; \end{cases}$$

$$\text{denominator} := \begin{cases} y\bar{t}_{i_z} \sim \text{Poisson}(\nu_{i_z}), \\ \log(\nu_{i_z}) = \delta_z + \gamma_z x^*_{i_z} \end{cases}$$

with priors $\alpha_z, \beta_z, \delta_z, \gamma_z \sim^{\text{IID}} N(0, 100)$, $i_z = 1, \ldots, n_z$ and $z \in \{0, 1\}$ throughout.

We assign relatively vague priors on the regression coefficients. Tighter priors such as those suggested in Gelman *et al.* (2008) have been considered, but the results are not very sensitive to the choice of prior, at least for the specific data at hand. As we centre the risk score at the threshold, the parameter of interest in all the regressions is the exponential of the intercept term. The posterior MCMC samples of the parameters α_1, α_0, δ_1 and δ_0 can be used to characterize $E(Y \mid Z=1), E(Y \mid Z=0), E(Y\bar{T} \mid Z=1)$ and $E(Y\bar{T} \mid Z=0)$ respectively and then combined to obtain the posterior sample for the RRT. The model described above has the interaction model denominator:

$$\mathrm{RRT}_{\mathrm{pois.pois}} = 1 - \frac{\Pi_{\mathrm{pois}}}{\Psi_{\mathrm{pois}}}$$

where

$$\Pi_{\mathrm{pois}} = \exp(\alpha_1) - \exp(\alpha_0)$$

and

$$\Psi_{\mathrm{pois}} = \exp(\delta_1) - \exp(\delta_0).$$

We also consider an interaction model where the denominator is based on a flexible binomial model as used in Geneletti *et al.* (2015). In this model the prior information is used to create distance between the two elements in the denominator of the fraction in the RRT in equation (2). This often stabilizes the results because it pushes the difference in the probability of treatment at the threshold away from zero and thus inflates the fraction in equation (2). In this case the denominator is defined as

$$y\bar{t}_{i_z} \sim \mathrm{binomial}(q_z, n_z),$$

priors

$$\mathrm{logit}(q_1) \sim N(-3, 1)$$

and

$$\mathrm{logit}(q_0) \sim N(3, 1)$$

so that $\Psi_{\mathrm{flex}} = q_1 - q_0$.

This results in the interaction model

$$\mathrm{RRT}_{\mathrm{pois.flex}} = 1 - \frac{\Pi_{\mathrm{pois}}}{\Psi_{\mathrm{flex}}}.$$

We now consider the product denominator as follows:

$$\mathrm{Denominator.prod} = \begin{cases} y_{i_z} \sim \mathrm{Poisson}(\theta_{iz}), \\ \log(\theta_{i_z}) = \delta_z + \gamma_z x_{i_z} + \kappa_z \bar{t}_{i_z}, \\ t_{i_z} \sim \mathrm{binomial}(n_z, r_z), \end{cases}$$

priors

$$\mathrm{logit}(r_1) \sim N(-3, 1)$$

and

$$\mathrm{logit}(r_0) \sim N(3, 1).$$

We then combine the conditional probabilities as follows:

$$E(Y\bar{T} \mid Z = z) = (\delta_z + \kappa_z) r_z,$$

as we are interested in the case where both Y and $\bar{T}$ are equal to 1. Note that we use the binomial flex model again for the probability of $\bar{T}$ as this was less variable than regression-based models, in this case. Thus we obtain $\mathrm{RRT}_{\mathrm{pois.prod.flex}}$ as follows:

$$\mathrm{RRT}_{\mathrm{pois.prod.flex}} = 1 - \frac{\Pi_{\mathrm{pois}}}{\Psi_{\mathrm{prod}}}$$

where

$$\Pi_{\text{pois}} = \exp(\alpha_1) - \exp(\alpha_0)$$

and

$$\Psi_{\text{prod}} = \exp(\delta_1 + \kappa_1) r_1 - \exp(\delta_0 + \kappa_0) r_0.$$

3.3. *Constraints*

The RRT can drop below zero if the fraction in equation (2) exceeds 1 (Clarke and Windmeijer, 2010). We avoid this problem by imposing *a priori* constraints on the distribution of the RRT which force the RRT to remain within acceptable bounds.

Imposing prior constraints is straightforward in the Bayesian framework. We place a gamma prior on the RRT with most of the mass close to 1, as we do not want to encourage an arbitrarily large risk ratio. The most straightforward constraint was to make α_1 a function of the other variables above the threshold so that we could place a prior on the RRT. We could equally have chosen α_0. We write out the changes that the model implies to the priors below:

$$\text{RRT}_{\text{pois.pois}} \sim \text{gamma}(3, 1)$$

and

$$\alpha_1 = \log\{(1 - \text{RRT}_{\text{pois.pois}}) \Psi_{\text{pois}} + \exp(\alpha_0)\}$$

where $\exp(\alpha_1) - \exp(\alpha_0) = \Pi_{\text{pois}}$ and there is of course no prior on α_1 (whose distribution is induced by the logical relationships that were specified above). Similar changes can be made to all the RRTs with any of the other models presented in Section 3. It is also possible to impose constraints on logistic-regression-based estimates.

We also tried to impose constraints on the intercept in the denominator δ_1 for $\text{RRT}_{\text{pois.pois}}$. However, the results were more variable even though point estimates remained in the same region.

3.4. *Frequentist estimators: multiplicative and logistic structural mean models*

To assess the performance of our Bayesian estimators, we compared them with some of the most common estimators for binary outcomes in the IV literature. These include the MSMM (Clarke *et al.*, 2015), the double LSMM (Vansteelandt *et al.*, 2011; Vansteelandt and Goetghebeur, 2003; van der Laan *et al.*, 2007), the Wald risk ratio (Didelez *et al.*, 2010) and a final method based on a single estimating equation (Burgess *et al.*, 2014). We give an overview of the MSMM and LSMM below. Details of all the frequentist estimators including moment conditions and additional assumptions for identification can be found for all listed methods in the on-line supplementary material.

Clarke *et al.* (2015) showed that the semiparametric MSMM estimator in equation (2) can be estimated efficiently by the generalized method of moments and we also relied on this method to obtain the estimates. However, generalized method-of-moments estimators can lead to a lack of identification in binary outcome situations. This is amplified when the instrument is weak (Burgess *et al.*, 2014). This could potentially be the reason for the erratic behaviour that we see in small bandwidths and when the threshold instrument is weak in our simulation studies.

The LSMM as defined by Vansteelandt and Goetghebeur (2003) can be used to estimate the causal odds ratio for a binary outcome in the presence of an instrument. A potential problem is that the odds ratio is non-collapsible which means that logistic regression equivalents of assumptions 5 and 6 are not sufficient to identify the causal odds ratio. It can be identified if we specify $\text{logit}\{E(Y \mid Z, T)\} = \text{lp}(X, T)$ where lp is a linear predictor. However, unless the lp-model is saturated, it will be uncongenial (i.e. correspond to a different data-generating mechanism) to the LSMM. Vansteelandt *et al.* (2011) argued that this is not a problem in practice. Our results show that the LSMM produces highly variable results when the instrument is weak and/or bandwidths are small and is typically close to the MSMM. It also dramatically overestimates the true effect when it is high. This may be a consequence of the violation of the rare disease assumption.

4. Simulation study

We set up a simple simulation study aimed at examining the properties of the models that were presented in Section 3. We considered two levels of unobserved confounding U, two levels of IV strength Z and three causal effects resulting in 12 simulation scenarios. We based our simulated data set on the larger data set from which the set that was described in Section 2 and is analysed in Section 5 was obtained. We used the original values for the risk score X and the standardized high density lipoprotein cholesterol level as unobserved confounder U. The threshold indicator Z was defined deterministically as $Z_i = 0$ if $X_i < 0.2$ and $Z_i = 1$ otherwise. For each simulation we run the analyses in each of four bandwidths identified by values $h = \{0.025, 0.05, 0.075, 0.1\}$ and assess the sensitivity of the results to these changes.

The simulated data-generating process for the treatment or outcome is as follows: first we linked the probability of receiving treatment $p_1 = P(T = 1)$ and outcome $p_2 = P(Y = 1)$ with a confounding variable U and the threshold indicator Z. We then used the observed U and Z and generated the treatment status T and the outcome Y respectively based on p_1 and p_2 respectively. Specifically the treatment was simulated as

$$p_{1,i} = \min\{1, \exp(-2.5 + \xi_1 Z_i + \xi_2 U_i)\}$$

and

$$T_i \sim \text{Bernoulli}(p_{1,i}).$$

For the values of $\xi_1 \in \{1.3, 2.3\}$ (weak and strong IV) and $\xi_2 \in \{0.4, 2\}$ (low and high confounding) we obtain different settings of strength of instrument and unobserved confounding.

The outcome was then simulated as

$$p_{2,i} = \min\{1, \exp(-2 + \zeta_1 Z_i + \zeta_2 U_i)\}$$

and

$$Y_i \sim \text{Bernoulli}(p_{2,i})$$

where $\zeta_2 = \xi_2/4$ and $\zeta_1 \in \{0, 0.75, 1.5\}$, so that we obtain three different risk ratios: 1, 2.11 and 4.48 corresponding to no, low and high causal effect. We ran 100 replications each with 7500 MCMC runs for burn-in and a further 15000 iterations of which the last 1000 were used for estimation. We ran diagnostics on a random selection of simulation runs and were satisfied that convergence was reached on all relevant parameters (see the on-line supplementary material for diagnostics).

We note here that the risk ratios corresponding to no, low and high casual effects depend on the level of truncation owing to the specifications of $p_{1,i}$ and $p_{2,i}$ and may not be 'exactly' 1, 2.11 and 4.48 depending on the level of truncation within this specification. This is not likely to be problematic in our simulations, where the level of truncation is negligible. However, for situations in which the level of truncation is thought to be high, Monte Carlo methods could be used on a large sample to estimate the ratio for those for whom $T = 1$. In addition, the change in bandwidth within the simulation set-up does not imply a change in the exchangeability assumption. Rather, as the bandwidth decreases, the sample size reduces and, as such, a change in bandwidth should be considered as analogous to a change in sample size in the simulation set-up.

We produced exploratory plots like those introduced in Section 2 for all simulated scenarios. Briefly, for high causal effects all scenarios showed a clear jump at the threshold although it was smaller in the weak IV–high confounding scenario. For the low causal effects the results were variable. No jump was discernible in the weak IV–high confounding scenario in the outcome probabilities but a small jump was discernible in the probability of treatment. A clear jump was visible for both outcome and probability in the strong IV scenarios. Finally for the no-effect scenarios no jump was visible as expected. A selection of plots can be seen in the the on-line supplementary material section 2.

We obtained results for Bayesian constrained and unconstrained models as well as a number of frequentist estimates and the Balke–Pearl bounds which are available from the authors on request. In the body of the paper we present only results for the constrained models as in many scenarios (in particular for small bandwidths, high confounding and weak instruments) the posterior MCMC sample for the unconstrained models contained values below 0. In addition, although convergence was reached for the relevant parameters, there were some extreme results due to small values in the denominator that occasionally led to inflated mean estimates. Medians for the unconstrained models and means for the constrained models were generally close although the constrained model estimates were typically smaller (see the supplementary material). We thought that this might be due to the gamma prior pulling the RRT in the constrained models towards 1. However, on investigation we saw that results were not sensitive to the choice of prior.

Fig. 3 shows the results over 100 replications of four estimators in three out of 12 simulated scenarios.

In Fig. 3(a) the risk ratio is 4.48, the confounding is low and the instrument is strong. It can be seen as the ‘best case’ scenario. The Bayesian and MSMM estimators perform well with the latter performing slightly better. The LSMM overestimates the true effect dramatically. This is echoed in the real data analysis in the next section.

In Fig. 3(b) the risk ratio is 2.11, the confounding is high and the instrument is weak. This is the ‘worst-case’ scenario. Overall the Bayesian estimators outperform the frequentist approaches. They are stable if non-significant across all bandwidths with point estimates close to the true value. Both the MSMMs and the LSMMs are very sensitive to bandwidth.

Results in Fig. 3(c) are for the low confounding and weak instrument scenario when the true risk ratio is 2.11. The Bayesian estimators are close to the true value throughout with good mean-squared error (MSE) and coverage although they remain non-significant (albeit barely) for all bandwidths. In the smallest bandwidth both the MSMM and the LSMM have very high means (over 31000 and 1 million respectively) and large confidence intervals. Medians are in line with the Bayesian means. Point estimates improve for the frequentist estimators and become (barely) significant for the highest bandwidth.

Finally Fig. 3(d) represents the scenario with a high level of confounding, weak instrument and no effect. The MSMM and LSMM perform well in the 0.05-bandwidth and the latter performs better for the larger bandwidths. The Bayesian estimators overestimate the true effect—in fact they are close to the observational unadjusted risk ratio, which is 2.18. This is not surprising or particularly worrying. The RD design analysis relies on being able to estimate a jump in the

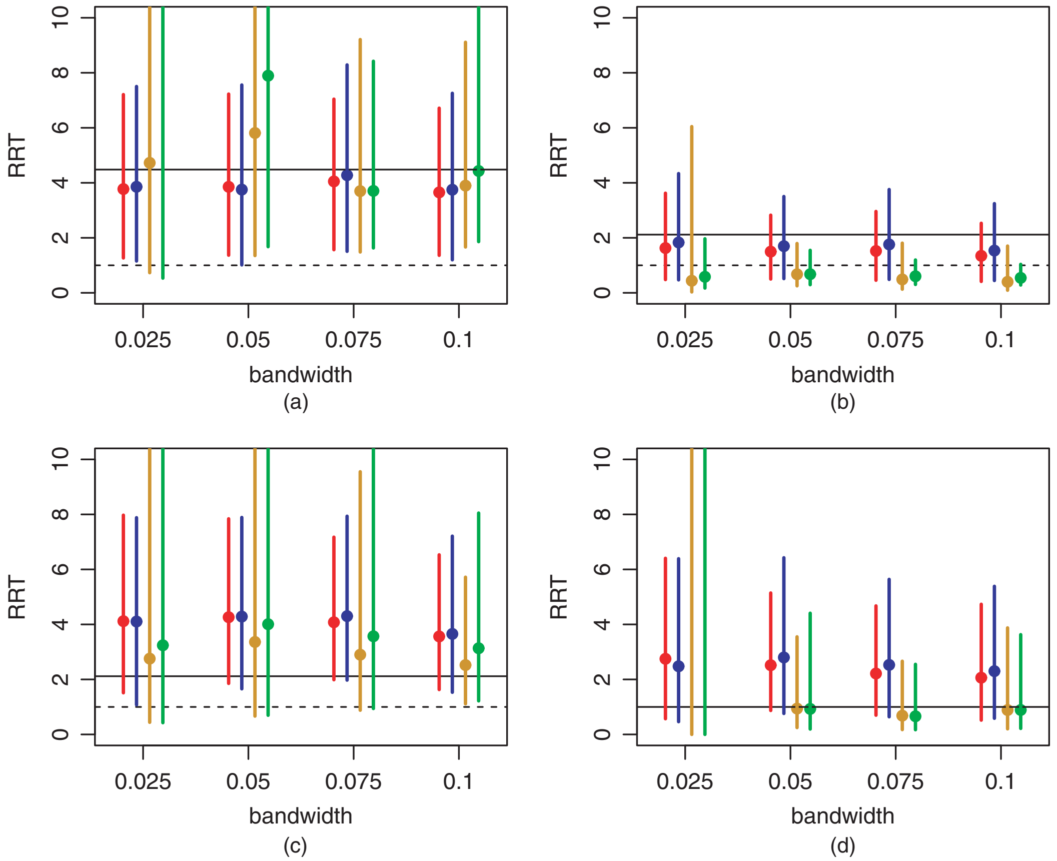


Fig. 3. Comparison of the results of pois.flex (|), pois.prod.flex (|), MSMM (|) and LSMM (|) in four out of 12 simulated scenarios for the constrained models (for each estimator, the coloured line represents the upper and lower 95% quantiles over the 100 replicates whereas the dots represent the mean value; the true risk ratio is shown by the horizontal line and, in case the true effect differs from 1, the broken line indicates the absence of effect): (a) low confounding, strong IV, RR = 4.48; (b) high confounding, weak IV, RR = 2.11; (c) low confounding, weak IV, RR = 2.11; (d) high confounding, weak IV, RR = 1

outcome at the threshold. When this is not present as in this scenario it makes sense that it would revert to the observational estimate. In this case we would not recommend an RD design because the RD plots (like those in Fig. 2) do not exhibit a jump. It is encouraging that over the replicates they cover the true null effect.

More details of the simulation studies corresponding to Fig. 3 (including the MSEs and coverage) can be found in the on-line supplementary materials.

Overall the Bayesian estimators are robust to changes in IV strength, confounding levels and size of causal effect. For small bandwidths in particular the Bayesian estimator compares favourably with the competing methods in the borderline cases: the credible intervals always include the true value, and the simulation MSE and coverage are above 90% in all cases. The frequentist approaches are not as consistent. For example, the MSMM is unreliable when the instrument is weak. The LSMM performs better than the MSMM but is also sensitive to bandwidth and size of effect, something which had also been noted by Clarke and Windmeijer (2010). The MSE and coverage were also more variable. We show the results for three scenarios here, the 'worst' and 'best case' and an intermediate case. For the other scenarios the results fall between the two extremes. The Bayesian estimators consistently produce stable and reliable estimates.

A possible reason for the robustness of Bayesian estimators in the extreme scenarios is that continuous information is used in estimating the components of the RD design whereas the frequentist estimates are based on binary data only and treat the RD design as a standard IV problem. Although our approach is robust, it does not perform well in the worst-case scenario when the effect is null. We are not overly concerned with this, however, as in these cases no RD design analysis is recommended.

Our prior constraints are very strong and we assessed how much results were affected. Counter to our expectation the results from our simulations and applied example constraining the models to be above zero did not result in higher RRT estimates; nor were results very sensitive to prior specification. Specifically varying the values of the gamma distribution on the RRT (e.g. by moving the mass further from 1) or even using a different prior with positive support (i.e. log-normal) did not alter the results substantially. Instead the constrained models stabilized the posterior MCMC sample.

5. Example: statin prescription

Using the THIN data (which were introduced in Section 2.1), we shall investigate whether or not statin prescription lowers the LDL cholesterol to below 2 and 3 mmol l^{-1}, recommended levels for high and low risk individuals respectively. We have performed analyses for the two risk groups separately but the results do not differ substantially and we focus here on all the patients.

From trials (Ward *et al.*, 2007) we know that statins are effective in lowering cholesterol. As LDL cholesterol tends to decrease quickly within a month of uptake and our data span the 6 months around the cholesterol measurement we can use our binary outcomes RD design to determine whether statins result in people achieving LDL cholesterol targets within a small time window. Our approach is also useful when we are interested in whether a drug acts on a relevant marker of a disease which is easier to measure and is affected quickly by treatment.

5.1. Preliminary analyses

Before estimating the RRT we investigated whether a Poisson regression was an appropriate model for the data. The model fit was good overall and there was no evidence for overdispersion.

In line with recent recommendations regarding what should be presented in IV analyses (Swanson and Hernan, 2013) we performed F-tests to determine IV weakness for non-linear situations (Windmeijer and Didelez, 2016) for both binary outcomes (LDL level below 2 and below 3 mmol l^{-1}). The F-values ranged from 10 (for bandwidth 0.025) to 211 (for bandwidth 0.1) with p-values significant at the 5% level throughout. We also produced estimating functions (Burgess *et al.*, 2014) that are shown in the on-line supplementary materials. These indicated that for our data it was possible to find unique solutions to the MSMM and LSMM moment conditions. The Balke–Pearl bounds (Balke and Pearl, 1997; Palmer *et al.*, 2011) were also in line with our results.

Finally, we performed the McCrary density test (McCrary, 2008) which detects whether there is a discontinuity at the threshold in the running variable (in our case the risk score). If the result is significant any effect that we observe might be due to the discontinuity in the running variable. The result was non-significant at the 5% level.

5.2. Main analysis

We fitted our models by using `JAGS` (Plummer, 2003) with two chains, a burn-in of 10000

Table 1. Results for estimates of the constrained RRTs, as well as the MSMM and LSMM estimator†

Model	*Results for LDL < 3 mmol* l^{-1}			*Results for LDL < 2 mmol* l^{-1}		
	Mean	*L95*	*U95*	*Mean*	*L95*	*U95*
$h=0.025$						
pois.flex	3.12	1.32	5.62	2.59	0.86	5.57
pois.pois	3.98	1.32	8.16	2.96	0.89	6.42
pois.prod.flex	3.89	1.11	6.16	2.77	0.77	5.81
MSMM	3.99	1.30	12.21	8.39	1.49	47.37
LSMM	21.3	4.42	102.6	13.7	2.12	87.9
$h=0.05$						
pois.flex	3.95	1.51	7.39	2.59	0.99	4.91
pois.pois	4.01	1.12	7.93	2.81	0.85	5.92
pois.prod.flex	4.17	1.08	8.14	2.64	0.85	5.34
MSMM	4.53	2.29	8.93	4.49	3.04	29.62
LSMM	17.2	6.66	44.2	13.9	4.12	47.3
$h=0.075$						
pois.flex	3.76	1.48	7.84	2.63	1.02	4.76
pois.pois	4.33	1.47	8.60	3.19	0.95	6.13
pois.prod.flex	4.60	1.57	8.60	2.80	1.01	5.33
MSMM	4.22	2.55	7.03	6.68	3.07	14.56
LSMM	15.7	7.60	32.6	9.30	3.98	21.8
$h=0.1$						
pois.flex	3.69	1.42	6.82	2.32	1.30	3.70
pois.pois	4.02	1.37	7.60	2.68	1.02	7.13
pois.prod.flex	3.99	1.36	7.58	2.70	1.31	4.86
MSMM	3.76	2.60	5.44	7.36	3.73	14.55
LSMM	13.7	7.89	23.8	10.3	4.92	21.6

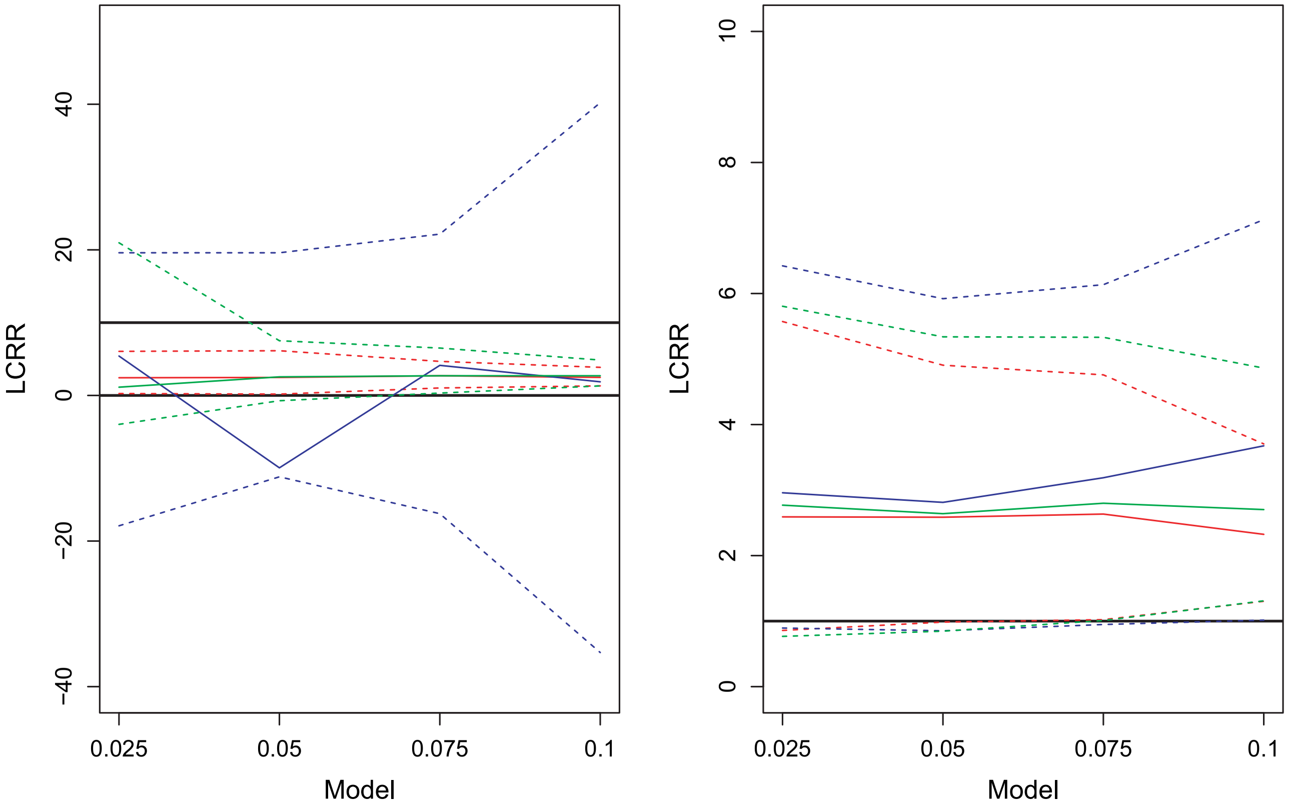
†Means and 95% credible or confidence intervals are provided. The second to fourth columns are for the case where the target value of the outcome is 3 mmol l^{-1} whereas the last three are for an outcome threshold equal to 2 mmol l^{-1}.

iterations and a further 50000 iterations. Our posterior samples were based on the last 1000 iterations. On average each model took 5 min to run on a standard personal computer. Convergence was reached for all relevant parameters and mixing was good. See the on-line supplementary material.

Overall the results indicate a positive effect of statin treatment on LDL cholesterol levels for patients in our sample with a large (if not universally significant) twofold to threefold increase in the probability of achieving the target LDL cholesterol level within 6 months of prescription for those prescribed relative to those eligible. This is especially true for the target of reducing the LDL cholesterol level to below 3 mmol l^{-1}.

Table 1 shows for each bandwidth the mean and lower and upper 95% interval estimates for the RRT with constrained models the MSMM and the LSMM.

Generally, constrained Bayesian estimates are similar and increase slightly as the bandwidth increases. The point estimates for LDL less than 3 mmol l^{-1} are slightly larger (3.1–4.6) than those for LDL less than 2 mmol l^{-1} (2.5–3.2 except $h=0.1$) and are significant. The median LDL cholesterol level for the untreated is 4 mmol l^{-1}, so a drop of 2 mmol l^{-1} LDL will be less frequent and lead to less significant results. The results are also very stable across bandwidths.



(a)

(b)

Fig. 4. Means (—, pois.flex; —, pois.pois; —, pois.prod.flex) and 95% credible intervals (⋯, pois.flex; ⋯, pois.pois; ⋯, pois.prod.flex) for three Bayesian estimators and four different bandwidths (the *y*-axes have different scales): (a) no constraint, LDL level less than 3 mmol l^{-1}; (b) constrained models, LDL level less than 3 mmol l^{-1}.

The plots in Fig. 4 show the RRT estimates for the unconstrained models in Fig. 4(a) and the constrained models in Fig. 4(b) for LDL level below 3 mmol l^{-1}. We can see that in particular for the low bandwidths where the data are sparse the estimates by using the unconstrained models are very variable. Similar plots displaying a similar pattern for the LDL level below 2 mmol l^{-1} are given in the supplementary material. Overall the constrained estimates both reduce variability and remain above 0 and the means for the constrained and medians for unconstrained modes are similar.

When we compare the results of the MSMM estimator with ours we see that for the LDL level below 3 mmol l^{-1} the scenarios are broadly similar. For the outcome LDL below 2 mmol l^{-1}, however, the results are different. For all the bandwidths the means of the MSMM are high and the intervals are wide. The LSMM consistently overestimates the effect and this is aggravated for small bandwidths and higher effects. We see similar behaviour in the simulation studies.

6. Conclusions

In this paper we use the RD design to develop a flexible Bayesian method to estimate the causal effect of treatments for binary outcomes. Using our proposed method routinely gathered medical data can be exploited to estimate the effects of government drug administration guidelines. We focus on the causal risk ratio for the treated as this is of primary interest to medical practitioners.

Our RRT estimates are based on the structural mean model assumptions (Clarke and Windmeijer, 2010; Hernan and Robins, 2006). This estimator is known to produce values below 0. We avoid this by imposing prior constraints. The fact that the RRT as identified by equation (2) has no built-in safeguard to dropping below 0 raises some questions about its appropriateness as a risk ratio estimator. It is not easy to identify causal quantities when all elements are binary and as a consequence some strong assumptions must be met. It is likely that some will hold only approximately. If we suspect that assumptions 1–4 do not hold then there is no point in attempting an RD design analysis or estimating the RRT from these data; however, if we think of assumptions 5 and 6 as holding only approximately around the threshold then we can view this as a model misspecification problem and the RRT as an approximation to the true underlying effect. In this paper, we chose to focus on Bayesian estimation of the RRT. We note that a similar approach could be applied to other risk ratio estimators (e.g. the Wald risk ratio estimator).

Our results compared favourably with the MSMMs and the SMMs as well as other frequentist estimators. In particular they were more robust to weak instruments, high levels of confounding and low effects in the simulation studies. They also produced more credible results in our application. On the basis of our simulation study they also outperform the frequentist estimators in terms of the MSE and coverage especially in small bandwidths.

We preferred Poisson regression-based models on theoretical grounds and because in the RD design we need to use a continuous threshold variable. However, we also implemented models that are identical to those we propose where logarithms are replaced by logits and exponentials with expits. Further we tried zero-inflated Poisson regression models and binomial models. The results are broadly similar especially for the regression-based models. `JAGS` code for some of these models is given in the on-line supplementary materials and results are available on request. It is also feasible to use spline or other semiparametric models.

We included only the running variable in our regression; however, it is possible to add more covariates if data are available. The resulting effect estimates would then be conditional on these covariates.

We found that in the simulation studies, and indeed in the application, the method proposed gave the most reliable and realistic results. However, we recommend that anyone attempting an

RD design analysis, especially for binary outcomes, follow a procedure that is similar to ours. This includes some preliminary analyses from plots to tests of IV weakness and then using a range of numerator–denominator combinations within our framework. This will ensure that if there is an effect it can be identified as lying within a reliable range.

Acknowledgements

We thank the Joint Editor and two referees for their comments and suggestions. The paper has greatly benefitted from their input. This research has been funded by UK Medical Research Council grant MR/K014838/1. Approval for this study was obtained from the THIN Scientific Review Committee in August 2014.

References

Abadie, A. (2002) Bootstrap tests for distributional treatment effects in instrumental variable models. *J. Am. Statist. Ass.*, **97**, 284–292.

Abadie, A. (2003) Semiparametric instrumental variable estimation of treatment response models. *J. Econmetr.*, **113**, 231–263.

Angrist, J. (2001) Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *J. Bus. Econ. Statist.*, **19**, 2–16.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of casual effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Balke, A. and Pearl, J. (1997) Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Ass.*, **92**, 1171–1176.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M. L. and Barnighausen, T. (2014) Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology*, **25**, 729–737.

Burgess, S., Granell, R., Palmer, T. M., Sterne, J. A. C. and Didelez, V. (2014) Lack of identification in semiparametric instrumental variable models with binary outcomes. *Am. J. Epidem.*, **180**, 111–119.

Burgess, S. and Thompson, S. G. (2012) Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statist. Med.*, **31**, 1582–1600.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2015) Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica*, **82**, 2295–2326.

Clarke, P. S., Palmer, T. M. and Windmeijer, F. (2015) Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Statist. Sci.*, **30**, 96–117.

Clarke, P. and Windmeijer, F. (2012) Instrumental variable estimators for binary outcomes. *J. Am. Statist. Ass.*, **107**, 1638–1652.

Clarke, P. S. and Windmeijer, F. (2010) Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, **11**, 756–770.

Dawid, A. P. (1979) Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc.* B, **41**, 1–31.

Didelez, V., Meng, S. and Sheehan, N. A. (2010) Assumptions of IV methods for observational epidemiology. *Statist. Sci.*, **25**, 22–40.

Gelman, A., Aleks, J., Pittau, M. and Su, Y. (2008) A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.*, **2**, 1360–1383.

Geneletti, S. and Dawid A. (2010) The effect of treatment on the treated: a decision theoretic perspective. In *Causality in the Sciences* (eds M. Ilari, F. Russo and J. Williamson). Oxford: Oxford University Press.

Geneletti, S., O'Keeffe, A. G., Sharples, L. D., Richardson, S. and Baio, G. (2015) Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statist. Med.*, **34**, 2334–2352.

Hahn, J., Todd, P. and van der Klaauw, W. (2001) Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69**, 201–209.

Hernan, M. and Robins, J. (2006) Instruments for causal inference—an epidemiologist's dream? *Epidemiology*, **17**, 360–372.

Imbens, G. and Kalyanaraman, K. (2012) Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econ. Stud.*, **79**, 933–959.

Imbens, G. W. and Lemieux, T. (2008) Regression discontinuity designs: a guide to practice. *J. Econmetr.*, **142**, 615–635.

van der Klaauw, G. (2008) Regression-discontinuity analysis: a survey of recent developments in economics. *Labour*, **22**, 219–245.

van der Laan, M. J., Hubbard, A. and Jewell, N. P. (2007) Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *J. R. Statist. Soc.* B, **69**, 463–482.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lee, D. S. (2008) Randomized experiments from non-random selection in US House elections. *J. Econmetr.*, **142**, 675–697.

Lee, D. S. and Lemieux, T. (2010) Regression discontinuity designs in economics. *J. Econ. Lit.*, **48**, 281–355.

Linden, A., Adams, J. and Roberts, N. (2006) Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *J. Evaln Clin. Pract.*, **12**, 124–131.

McCrary, J. (2008) Manipulation of the running variable in the regression discontinuity design: a density test. *J. Econmetr.*, **142**, 698–714.

Moscoe, E., Bor, J. and Baernighausen, T. (2015) Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *J. Clin. Epidem.*, **68**, 132–143.

Palmer, T. M., Ramsahai, R. R., Didelez, V. and Sheehan, N. A. (2011) Nonparametric bounds for the causal effect in a binary instrumental-variable model. *Stata J.*, **11**, 345–367.

Plummer, M. (2003) Jags: a program for analysis of Bayesian graphical models using Gibbs sampling.

Smith, L. M., Kaufman, J. S., Strumpf, E. C. and Levesque, L. E. (2015) Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study. *Can. Med. Ass. J.*, **187**, E74–E81.

Stock, J. and Yogo, M. (2005) *Testing for Weak Instruments in Linear IV Regression*, pp. 80–108. New York: Cambridge University Press.

Swanson, S. and Hernan, M. A. (2013) How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, **24**, 1044–3983.

Thistlethwaite, D. and Campbell, D. (1960) regression-discontinuity analysis—an alternative to the ex-post-facto experiment. *J. Educ. Psychol.*, **51**, 309–317.

Vansteelandt, S., Bowden, J., Babanezhad, M. and Goetghebeur, E. (2011) On instrumental variables estimation of causal odds ratios. *Statist. Sci.*, **26**, 403–422.

Vansteelandt, S. and Goetghebeur, E. (2003) Causal inference with generalized structural mean models. *J. R. Statist. Soc.* B, **65**, 817–835.

Ward, S., Jones, L., Pandor, A., Holmes, M., Ara, R., Ryan, A., Yeo, W. and Payne, N. (2007) A systematic review and economic evaluation of statins for the prevention of coronary events. *Hlth Technol. Assessmnt*, **11**, 1–160.

Windmeijer, F. and Didelez, V. (2016) Methods for binary outcomes. In *Mendelian Randomization: How Genes Can Reveal the Biological and Environmental Causes of Disease* (ed. G. Davey-Smith). Oxford: Oxford University Press. To be published.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Bayesian modelling for binary outcomes in the regression discontinuity design supplementary materials'.