Bayesian shrinkage estimation of credible subgroups for count data with excess zeros

Duy Ngo
Department of Statistics, Western Michigan University
and
Daniel Quartey
Structural Heart & Aortic, Medtronic PLC.
and
Patrick M Schnell
Division of Biostatistics, The Ohio State University College of Public Health
and
Richard Baumgartner
BARDS (Biostatistics and Research Decision Sciences), Merck & Co., Inc.
and
Shahrul Mt-Isa
BARDS (Biostatistics and Research Decision Sciences), MSD.
and
Dai Feng
Medical Affairs and Health Technology Assessment Statistics, AbbVie Inc.

July 16, 2024

Abstract

Heterogeneity of treatment effects due to heterogeneous patient characteristics often arises in clinical trials. Subgroup analysis and the analysis of interactions are the most common approaches for evaluating such heterogeneous effects but do not explicitly address multiplicity issues. Another common challenge of analyzing treatment effect heterogeneity is the large number of possible covariates which inevitably causes problems related to multiplicity and lack of power. In this article, we develop a Bayesian credible subgroups method using continuous shrinkage priors to assess heterogeneity in treatment effects and multiplicity–adjusted benefiting subgroup identification for zero–inflated count data, which are often encountered in medical and public health studies. Our proposed method provides two bounding subgroups for the true benefiting subgroup: one that is probably contained by the true benefiting subgroup and one that probably contains the true benefiting subgroup. A simulation study has been conducted to compare the performance of the proposed method with other methods through frequentist properties. We apply our method to a clinical bladder tumor trial studying the effect of thiotepa treatment on the reduction of the recurrence of bladder tumor.

Keywords: Bayesian credible subgroups, continuous shrinkage, conditional average treatment effect, zero–inflated regression.

1 Introduction

Randomized clinical trials are primarily designed to draw inferences about a potential causal relationship between patient outcomes and a particular treatment. The effectiveness of a treatment has typically been measured by the average treatment effect (ATE) as the difference in average outcomes between two treatment groups. The ATE in the study population is assumed to be adequately reflective of the effect in any subject within this population, so the ATE may oversimplify the heterogeneity of each patient or similar subgroups of patients, which is known as heterogeneous treatment effect (HTE). The HTE often arises in clinical trials and observational studies when a treatment, that has a positive effect on a majority of patients, may have no effect on a subset of patients with certain characteristics due to variation in patient characteristics. For example in a study of antiretroviral therapy (ART), timing of ART varies in individuals with tuberculosis and newly infected with human immunodeficiency virus type 1 (Havlir et al. 2011). Among individuals with CD4+ T–cell counts less than 50 per cubic millimeter, those with earlier ART have a lower rate of new AIDS–defining illnesses and deaths than those with later ART, while patients with higher CD4+ T–cell counts did not significantly benefit from earlier ART. Therefore, the success of precision medicine depends on the correct identification of subgroups of patients who benefit from a treatment.

One research direction in precision medicine is subgroup analysis, where the patients are grouped based on the estimated individual–level treatment differences (Cai et al. 2011, Foster et al. 2011). Alternatively, Qian et al. (Qian & Murphy 2011) proposed a regression model for the response, and recommended the treatment achieving the best prediction. Ballarini et al. (Ballarini et al. 2018) introduced pointwise confidence intervals around predicted individual treatment effects for continuous, survival and binary endpoints. Machine learning techniques, such as the tree–based methods built on the idea of counterfactuals, have been used to identify subgroups with differential treatment responses (Su et al. 2009, 2011, Foster et al. 2011). In Bayesian framework, Schnell et al. (Schnell et al. 2016, Schnell 2017) developed simultaneous credible bands for conditional average treatment effects for continuous endpoints, and Ngo et al. extended their approach to survival endpoints (Ngo et al. 2020). However, a method to identify benefiting subgroups in the credible subgroups framework for count endpoints is desirable, but not yet available.

In many clinical trial applications, the outcome of interest is counting the occurrence of an event, such as the number of hospitalizations, doctor visits (Wagner et al. 2007) or adverse events related to a vaccine (Rose et al. 2006). Such count data are typically very skewed and exhibit overdispersion. A classical approach is to use Poisson regression with an overdispersion parameter or the negative binomial distribution (Hilbe 2011). When the overdispersion is a result of a bimodal distribution, such as the observed number of zeros exceeds the expected number of zeros from the corresponding Poisson regression, zero–inflated Poisson (ZIP) (Lambert 1992) or zero–inflated negative binomial models are common choices. In the ZIP model, a binary process will determine whether the observations are always zero or realizations from a Poisson distribution, so zeros can either arise from the binary process or from a Poisson distribution. In contrast to the ZIP model, the zeros in hurdle models (Mullahy 1986) are disjointed from the non-zeros modelling with a truncated Poisson distribution and are commonly used in econometrics (Cameron & Trivedi 2013). The choice between hurdle and ZIP models depend on the type of data, goals of study and statistical grounds (see Neelon et al. (2016) for more details).

For a particular choice of zero–inflated model, researchers measure the ATE by the simple difference–in–means estimator for the full sample, but ATE cannot explain how a treatment varies across the patient population. The conditional average treatment effect (CATE) of a binary treatment within the potential outcomes framework (Rubin 2005), is an alternative method to the ATE for identifying a target subgroup of subjects who are expected to gain substantial benefit from a given treatment. The CATEs are often estimated at each predictive covariate point, that is, a set of baseline characteristics that predicts the patient's response to a particular treatment. Then researchers can perform a null hypothesis significance testing for CATE at each covariate point. The drawback to this approach is that there are too many potential characteristics that can influence treatment effect, and this leads to low power and false positive findings due to multiple testings (Berry 1990, Cui et al. 2002, Lagakos et al. 2006).

To overcome these limitations, we develop a Bayesian credible subgroups approach for zero–inflated count data. Our proposed method is an extension of subgroup identification methodology proposed by Schnell et al. (Schnell et al. 2016, Schnell 2017) for count endpoints. Particularly, to estimate a true benefiting subgroup for count data with excess zeros, our approach is based on a two–stage process: (1) fit a ZIP regression in a Bayesian model setting for computing

CATE; (2) construct bounding subgroups based on the posterior distribution of CATE, resulting in a pair of credible subgroups: one that is probably contained by the true benefiting subgroup and one that probably contains the true benefiting subgroup. This means that patients within the former subgroup benefit from a particular treatment; while those are outside the later subgroup, they cannot benefit from a particular treatment, irrespective of their characteristics. The key feature of the proposed approach is that it enables controlling for multiplicity and providing simultaneous inference, which means that all covariate points corresponding to a specific subgroup simultaneously have a treatment effect exceeding a specified threshold.

Moreover, current advances in technology allow researchers to collect information from many biomarkers in a form of multidimensinal data, but many of the collected biomarkers may not have significant impacts on an experimental treatment. To improve the estimation and interpretation of CATE in the case where many potential covariates are observed, we have developed zero–inflated regression model by incorporating a Bayesian variable selection approach. There are numerous traditional variable selection methods, such as stepwise procedures in linear regression models, but they are unstable because the selection and estimation steps are performed separately (Breiman 1996). In the non–Bayesian approaches, penalization procedures such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) and its extensions have been used in many applications since the variable selection and parameter estimation can be handled simultaneously. In a Bayesian framework, variable selection can be performed by assuming shrinkage priors, such as spike–and–slab priors, on the model coefficients (Mitchell & Beauchamp 1988, George & McCulloch 1993, Geweke 1996). These types of priors are characterized by density functions that are concentrated at zero and have a large probability mass in a wide range of non–zero values. This structure tends to shrink the posterior mean of truly zero coefficients towards zero, but it less affects the posterior mean of non–zero coefficients. In this work, we extend the spike–and–slab priors (George & McCulloch 1993) to the context of zero–inflated data, and investigate the frequentist coverage properties of these priors on our proposed Bayesian credible subgroups.

This article proposes two major contributions that are not being addressed in the current literature: (1) our proposed method is an extension of the Bayesian credible subgroups (Schnell et al. 2016) method to identify benefiting subgroups for count data with excess zeros. This is important in many areas of benefit and risk analysis and in the field of personalized medicine. Moreover, our method is amenable to the confirmatory setting in a post hoc manner because it can address the multiplicity issues in late–stage clinical trials with multiple subgroups of patients, and it aims to evaluate treatment effect heterogeneity across subgroups of patients defined by the baseline or demographic covariates; (2) our Bayesian shrinkage approach improves interpretability of credible subgroups when there is a large collection of potential covariates. Finally, we provide the R codes and data which are available on GITHUB repository for reproducibility.

The remainder of this article is structured as follows. In Section 2, we present an overview of the ZIP model for count data with excess zeros. We introduce the CATE in Section 3, and use those to present the methodological approaches to construct Bayesian credible subgroups in Section 4. We examine the performance of our proposed approaches from extensive simulation studies in Section 5. In Section 6, we use a randomized controlled trial in patients with tumor bladder (Baetschmann & Winkelmann 2013) to illustrate our proposed methods. We conclude with a brief discussion in Section 7.

2 Zero–inflated count data regression

As discussed above, the first stage of our two–stage approach is to fit a ZIP regression in a Bayesian model setting. For zero–inflated count data, ZIP regression provides a convenient framework to model two subpopulations: a not–at–risk group for which the outcome is always zero and an at–risk group for which the outcome is realized from a count data distribution, such as Poisson or negative binomial distribution (Winkelmann 2008). A negative binomial model is often chosen over Poisson model when the at–risk group distribution exhibits overdispersion. Although the focus of this paper is to develop Bayesian credible subgroups for ZIP models, the zero–inflated negative binomial or other methods can be extended in the same manner. We now present an overview of ZIP regression in Bayesian model setting.

For subjects $i = 1, \dots, n$, let y_i be the response variable taking on only non–negative integers. The response y_i is assumed to be independent, with the density defined as

$$f(y_i) = \theta_i \mathbb{1}_0(y_i) + (1 - \theta_i) g(y_i), \tag{1}$$

where $0 \leq \theta_i \leq 1$ is the mixture proportion, and $\mathbb{1}_0(y_i)$ is an indicator function equaling 1 when $y_i = 0$ and 0 otherwise. Equation 1 directly represents the responses y_i as a product of two independent processes u_i and v_i, i.e. $y_i = (1 - u_i)v_i$. The Bernoulli process $u_i \sim$ Bernoulli(θ_i) determines whether the observed y_i is zero or not. If the observed outcome is nonzero, v_i is drawn from a count data model g. When g is a Poisson distribution, the ZIP model can be expressed as

$$f(y_i) = \begin{cases} \theta_i + (1-\theta_i)e^{-\mu_i} & y_i = 0 \\ (1-\theta_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & y_i > 0, \end{cases} \tag{2}$$

where $\mu_i = E(v_i)$. From Equation 2, the response y_i takes zero value either for $v_i = 0$ or $u_i = 1$. Therefore, the ZIP model can handle the extra zeros compared to the traditional generalized linear models, and the amount of extra zeros from the Poisson component v_i is determined by the mixture proportion θ_i. The mean and variance of the ZIP model are given by $E\,(y_i) = (1-\theta_i)\mu_i$ and $Var\,(y_i) = (1-\theta_i)(1+\mu_i\theta_i)\mu_i$. Moreover, the parameters θ_i and μ_i are modelled through canonical link generalized linear models (Lambert 1992), i.e.

$$\begin{aligned} \log\left(\frac{\theta_i}{1-\theta_i}\right) &= \boldsymbol{Z}_i^\intercal \boldsymbol{\tau} + \psi_i \boldsymbol{S}_i^\intercal \boldsymbol{\phi}, \\ \log(\mu_i) &= \boldsymbol{X}_i^\intercal \boldsymbol{\beta} + \psi_i \boldsymbol{W}_i^\intercal \boldsymbol{\gamma}. \end{aligned} \tag{3}$$

where $\boldsymbol{Z}_i$ and $\boldsymbol{S}_i$ are q and m dimensional vectors of prognostic and predictive covariates for the $i-$th subject in the zero component, respectively, and $\intercal$ denotes the transpose operator. The covariates $\boldsymbol{Z}_i$ is regarded as prognostic (also known as main effects) as they directly influence the outcome y_i, whereas the covariates $\boldsymbol{S}_i$ is considered as predictive (also known as moderation effects) as they influence the outcome y_i only through an interaction with the treatment variable ψ_i. Therefore, the predictive covariates are useful to identify the characteristics of patients that benefit from a treatment, and the prognostic covariates improve the precision of the estimates of treatment benefits. Similarly, $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ denote p and k dimensional vectors of prognostic and predictive covariates in the Poisson component, respectively. In Equation 3, it is generally applicable to other link functions (probit, complementary log-log, cauchit or log link) for modeling the zero component, but will not be considered here. Note that the terms predictive and prognostic are often used in clinical trials literature for precision medicine (Beckman et al. 2011, Ziegler et al. 2012, Zhao et al. 2022), and $\boldsymbol{Z}_i, \boldsymbol{S}_i, \boldsymbol{X}_i$ and $\boldsymbol{W}_i$ may have some common terms or be distinct. A practical aspect of our model is distinguishing prognostic and predictive effects, which might result in reducing the bias in the predictive effect. Moreover, identifying a covariate having both predictive and prognostic functions provides practical insights for subject matter experts.

In this article, our focus is on two treatment arms, such as $\psi_i = 1$ if the subject i is assigned to treatment and $\psi_i = 0$ otherwise. The parameters $\{\boldsymbol{\tau}, \boldsymbol{\phi}\}$ and $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ are corresponding vectors of regression coefficients, which provide separate inference for the zero component and Poisson component respectively. The interpretation of $\{\boldsymbol{\tau}, \boldsymbol{\phi}\}$ are the covariate effects on the probability of the treatment being fully effective (not–at–risk subpopulation), whereas $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ are the effect on the average count when the treatment is less than fully effective (at–risk subpopulation).

A common approach to assess the treatment effect is performing a likelihood ratio test for testing the significance of the unknown regression parameters in a ZIP model, but it is difficult to evaluate the overall treatment effect from separate treatment effect estimates of the two components. An average predicted value approach (Albert et al. 2014) and marginalized ZIP model (Todem et al. 2016) were proposed to do inference on the overall mean count while adjusting for covariates, in particular, to compare the means between treatment groups. However, these approaches provide inference for the overall treatment effect on the entire study population, which cannot be used to identify which patient benefits from a treatment. The CATE was introduced to determine the subgroups of subjects for which the treatment is the most beneficial (or most harmful) within the context of experimental data. In the following section, we precisely define the CATE based on the ZIP model as a device to account for extra zeros in the data.

3 The Conditional Average Treatment Effect (CATE)

We frame CATE for a ZIP model described in Equation 3 using the Neymann–Rubin potential outcomes framework (Rubin 2005), as follows. Under the stable unit treatment value assumption, let $y_i(0), y_i(1)$ denote the potential outcome for subject i receiving treatment assignment $\psi_i = 0$ and $\psi_i = 1$, respectively, and we cannot simultaneously observe a subject in both treatment arms. Under randomized experiments, we assume that these outcomes $y_i(0), y_i(1)$ are independent of the treatment assignment ψ_i. The observed data consist of $\boldsymbol{\mathcal{D}} = \{y_i, \psi_i, \mathcal{X}_i = (\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i, \boldsymbol{S}_i)\}$, for $i = 1, \ldots, n$, which are assumed to be independent and identically distributed draws from a superpopulation, which is the data generating process for a finite target population (Imbens & Rubin 2015, Ding et al. 2017). The average of the treatment effect is defined as

$$\Delta_{ATE} = E\left[y_i(1) - y_i(0)\right] = E\left[y_i|\psi_i = 1\right] - E\left[y_i|\psi_i = 0\right], \tag{4}$$

which is constant for all subjects, so this one-size-fits-all phenomenon cannot address the heterogeneity of the treatment effect. The expectations in Equation 4 refer to the distribution of the target population induced by the random sampling or by the (conditional) random assignment of the treatment.

The CATE, as opposed to the ATE, for the $i-$th subject is defined as the conditional average treatment difference in potential outcomes, i.e.

$$\Delta_{CATE}(\mathcal{X}_i) = E\left[y_i(1)|\mathcal{X}_i\right] - E\left[y_i(0)|\mathcal{X}_i\right]. \tag{5}$$

Since the characteristics of subjects are often defined by genetic or biomarker differences, we focus on biomarkers that are potential effect modifiers measured prior to the intervention to reduce the risk of confounding bias. Therefore, under the unconfoundeness assumption, i.e. $\{y_i(1), y_i(0)\} \perp \psi_i|\mathcal{X}_i$, we have

$$\Delta(\mathcal{X}_i) = \Delta_{CATE}(\mathcal{X}_i) = E\left[y_i|\mathcal{X}_i, \psi_i = 1\right] - E\left[y_i|\mathcal{X}_i, \psi_i = 0\right], \tag{6}$$

which measures the causal treatment effect for subjects with baseline covariates $\mathcal{X}_i$. Consistency and positivity are two other identifiability principles that receive less attention than the unconfoundeness assumption but are equally important in causal inference. The consistency assumption states that the treatment is sufficiently well-defined and does not have multiple versions with different effects on outcomes (Hernán 2016, VanderWeele 2009, Shiba & Kawahara 2021). While the positivity assumption states that all subpopulations have positive probability of being assigned to either of the treatments (Westreich & Cole 2010, Shiba & Kawahara 2021). In the context of ZIP regression, the expected outcome of subject i given $\mathcal{X}_i$ and ψ_i is

$$E(y_i|\mathcal{X}_i, \psi_i) = (1-\theta_i)\mu_i = \frac{\exp(\boldsymbol{X}_i^{\intercal}\boldsymbol{\beta} + \psi_i\boldsymbol{W}_i^{\intercal}\boldsymbol{\gamma})}{1+\exp(\boldsymbol{Z}_i^{\intercal}\boldsymbol{\tau} + \psi_i\boldsymbol{S}_i^{\intercal}\boldsymbol{\phi})}. \tag{7}$$

Consequently, the CATE for each subject i can be expressed as:

$$\begin{aligned}\Delta(\mathcal{X}_i) &= E(y_i|\mathcal{X}_i, \mathcal{Z}_i, \psi_i = 1) - E(y_i|\mathcal{X}_i\mathcal{Z}_i, \psi_i = 0), \\ &= \frac{\exp(\boldsymbol{X}_i^\intercal\boldsymbol{\beta} + \boldsymbol{W}_i^\intercal\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{Z}_i^\intercal\boldsymbol{\tau} + \boldsymbol{S}_i^\intercal\boldsymbol{\phi})} - \frac{\exp(\boldsymbol{X}_i^\intercal\boldsymbol{\beta})}{1 + \exp(\boldsymbol{Z}_i^\intercal\boldsymbol{\tau})}, \\ &= \frac{(1 + \exp(\boldsymbol{Z}_i^\intercal\boldsymbol{\tau})(\exp(\boldsymbol{X}_i^\intercal\boldsymbol{\beta} + \boldsymbol{W}_i^\intercal\boldsymbol{\gamma})) - (\exp(\boldsymbol{X}_i^\intercal\boldsymbol{\beta}))(1 + \exp(\boldsymbol{Z}_i^\intercal\boldsymbol{\tau} + \boldsymbol{S}_i^\intercal\boldsymbol{\phi}))}{(1 + \exp(\boldsymbol{Z}_i^\intercal\boldsymbol{\tau}))(1 + \exp(\boldsymbol{Z}_i^\intercal\boldsymbol{\tau} + \boldsymbol{S}_i^\intercal\boldsymbol{\phi}))}, \end{aligned} \tag{8}$$

which is a function of all covariates and parameters from both components of the ZIP model, thus making it possible to estimate $\Delta(\mathcal{X}_i)$. The interpretation of $\Delta(\mathcal{X}_i)$ is the difference in conditional expected counts between treatment and control for a patient with covariates $\mathcal{X}_i$.

Following the literature (Qian & Murphy 2011, Zhao et al. 2012), we define a treatment rule $f(\boldsymbol{\mathcal{X}})$ as a deterministic map from the covariate space $\boldsymbol{\mathcal{X}}$ to the binary treatment assignment, $f(\boldsymbol{\mathcal{X}}) : \boldsymbol{\mathcal{X}} \rightarrow \{0, 1\}$. For each subject, we observe the outcome y_i and the corresponding covariates $\mathcal{X}_i \in \boldsymbol{\mathcal{X}}$, and an optimal treatment rule of such subject is the one that maximizes the population average value $E\left[y_i\left(f\left(\mathcal{X}_i\right)\right)\right]$. Let $\mathcal{I}(A)$ be the indicator function for the event A. Since $\arg\max_f E\left[y_i\left(f\left(\mathcal{X}_i\right)\right)\right] = 1$ when $\mathcal{I}\left(\Delta(\mathcal{X}_i) > 0\right)$, and 0 otherwise, $\Delta(\mathcal{X}_i) > 0$ is a straightforward solution to obtain the optimal treatment rule for a subject with covariates $\mathcal{X}_i$. Note that when the alternative treatments have unequal cost, the decision rule can simply be replaced by $\Delta(\mathcal{X}_i) > \delta$ for some constant threshold δ. In other words, we can identify the benefiting subgroups of the population for which their CATEs exceed a pre–determined threshold δ representing the level at which the treatment is deemed effective.

4 Bayesian credible subgroups for zero–inflated count data

In the second stage, we focus on identifying subgroups in which every covariate point for a subject has an expected benefit from a treatment via CATE measurement $\Delta(\mathcal{X}_i)$. In recent years, a number of novel methods for subgroup identification have been developed in the arena of CATE assessment. (Berger et al. 2014) proposed a Bayesian model selection approach using tree–based priors for subgroup effects, in which the subgroups were considered as terminal nodes of the trees used to construct models for treatment effects and baseline covariates. (Zhang & Zhang 2022) introduced personalized modeling providing an optimal treatment regime. However, these approaches do not provide simultaneous inferences. The simultaneous inferences mean that the treatment effect of all covariate points in a specific region will exceed a specified threshold simultaneously. Under a counterfactual framework, (Weisberg & Pontes 2015) and (Lamont et al. 2018) discussed predicted individual treatment effect (PITE) for heterogeneous treatment effects. (Ballarini et al. 2018) considered the maximum likelihood and LASSO approaches for estimating PITE and constructing confidence intervals for the individual effects, but these approaches are susceptible to multiplicity issues. In the following sections, we adapt the Bayesian credible subgroups method (Schnell et al. 2016) for zero–inflated count data which can handle multiplicity and provide simultaneous inferences.

4.1 Defining Bayesian credible subgroups (BCSs)

The BCSs method (Schnell et al. 2016) was developed for simultaneous inference regarding who benefits from treatment in the context of a hierarchical linear model in a Bayesian framework, and its concept is to divide the samples according to cross-sectional subject characteristics by using CATE. Specifically, The BCSs method searches for the set of covariate points from a covariate space $\boldsymbol{\mathcal{C}}$ such that $B = \{\mathcal{X}_i \in \boldsymbol{\mathcal{C}} : \Delta(\mathcal{X}_i) > \delta\}$. In a Bayesian framework, we can estimate B by searching the covariate points $\mathcal{X}_i \in \boldsymbol{\mathcal{C}}$ in which the posterior probability of having $\Delta(\mathcal{X}_i)$ greater than δ given the observed data is greater than $(1-\alpha)$, where $1-\alpha$ is a credible level, i.e. $\hat{B}_\alpha = \{\mathcal{X}_i \in \boldsymbol{\mathcal{C}} : P\left(\Delta(\mathcal{X}_i) > \delta \mid \boldsymbol{\mathcal{D}}\right) > 1-\alpha\}$. A natural approach to identify such covariate points $\mathcal{X}_i$ is to perform hypothesis testing $\Delta(\mathcal{X}_i)$ at every covariate point, so the corrections for multiple testing should be used. However, as the number of tests for all possible covariate points is frequently large, it is difficult to find an appropriate multiple test adjustment.

To control for multiplicity issues, the BCSs approach constructs the credible subgroup pair (D, S) such that $P(D \subseteq B \subseteq S \mid \boldsymbol{\mathcal{D}}) > 1-\alpha$. Thus, the subgroup D (referred to as the exclusive credible subgroup) consists of covariate points $\mathcal{X}_i$ for which the types of subjects benefit from the treatment with posterior probability $1-\alpha$, whereas the subgroup S (referred to as the inclusive credible subgroup) consists of all types of subjects who benefit and also contains many non–benefit subjects. Figure 1 illustrate the BCSs method which divides the covariates space $\boldsymbol{\mathcal{C}}$ into three regions. The green region D represents profile of patients who benefit from treatment while the blue region S^c represents profile of patients who do not benefit from treatment. The orange region $S \backslash D$ represents an uncertainty in which we do not have enough information to

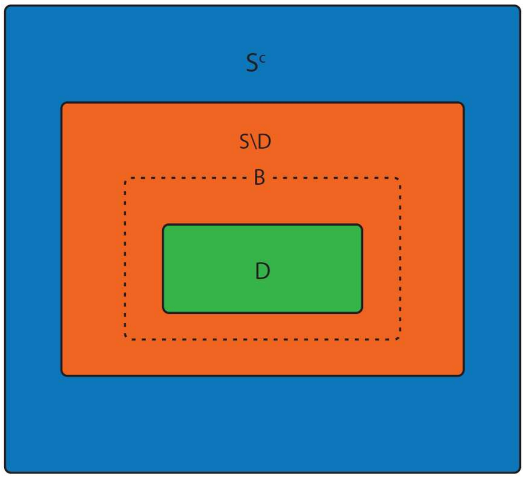


Figure 1: An illustration of credible subgroups. Region B (enclosed by dashed line) contains the true types of patients who benefit. Region D (green) includes only types of patients who benefit while region $S \backslash D$ (orange) is an uncertainty region. Region S^c (blue) represents types of patients who have no benefit.

determine whether patients are benefiting from treatment or not. Lastly, region B enclosed by the dashed circle represents the true benefiting group which we would want to estimate.

The credible subgroups D and S can be constructed from the results of the two–stage procedure. Particularly, we first fit a ZIP regression model in a Bayesian framework (described in Section 2) to get the posterior distribution of coefficients corresponding to covariate points $\mathcal{X}_i$. We then compute the marginal posterior of the CATE, and then use them to obtain a pair of credible subgroups (D, S) in the second stage (described in Section 4.4).

4.2 A Bayesian framework for estimating CATE

Under the ZIP regression model in Equation 3, the log likelihood of regression coefficients $\{\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ based on all n subjects in the data is given by

$$\begin{aligned} \ell(\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{\mathcal{D}}) = & \sum_{y_i=0} \log \left[\exp\left(\boldsymbol{Z}_i^\intercal \boldsymbol{\tau} + \psi_i \boldsymbol{S}_i^\intercal \boldsymbol{\phi} \right) + \exp\left(-\exp\left(\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + \psi_i \boldsymbol{W}_i^\intercal \boldsymbol{\gamma} \right) \right) \right] \\ & + \sum_{y_i>0} \left(y_i \left(\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + \psi_i \boldsymbol{W}_i^\intercal \boldsymbol{\gamma} \right) - \exp\left(\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + \psi_i \boldsymbol{W}_i^\intercal \boldsymbol{\gamma} \right) \right) \\ & - \sum_{y_i>0} \log\left(y_i ! \right) - \sum_{i=1}^{n} \log \left[1 + \exp\left(\boldsymbol{Z}_i^\intercal \boldsymbol{\tau} + \psi_i \boldsymbol{S}_i^\intercal \boldsymbol{\phi} \right) \right], \end{aligned} \tag{9}$$

which cannot be analytically maximized. The maximum likelihood estimation of the regression coefficients can be performed through convenient methods such as Newton–Raphson algorithm or the Expectation Maximization algorithm (Cohen 1963).

A Bayesian framework for the ZIP regression was introduced by (Ghosh et al. 2006). They assumed that the parameters $\{\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ are a priori independent, such as $\boldsymbol{\tau} \sim \mathcal{N}_q(\boldsymbol{\tau}_0, \sigma_\tau^2 \boldsymbol{I}_q)$, $\boldsymbol{\phi} \sim \mathcal{N}_m(\boldsymbol{\phi}_0, \sigma_\phi^2 \boldsymbol{I}_m)$, $\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \sigma_\beta^2 \boldsymbol{I}_p)$ and $\boldsymbol{\gamma} \sim \mathcal{N}_k(\boldsymbol{\gamma}_0, \sigma_\gamma^2 \boldsymbol{I}_k)$ are independent, and $\boldsymbol{I}_o$ is an identity matrix of size o. For each normal prior distribution, they used large variances to express flat but proper priors, and the posterior distributions of the parameters were obtained by using MCMC with data augmentation (Tanner & Wong 1987, Rodrigues 2003, Ghosh et al. 2006). In the presence of a large set of covariates, a large parameter space can severely affect the generalizability of the model due to overfitting, and the interpretation of BCSs can be difficult. To facilitate variable selection in the ZIP regression model, we adopt spike-and-slab priors, for the regression parameters in both parts of the model. These priors have been commonly used in the context of Bayesian stochastic search variable selection (George & McCulloch 1993, 1997) to select the relevant variables (i.e. those with non-zero effect).

4.3 Spike and slab prior

The spike–and–slab priors are commonly used for high dimensional variable selection in the Bayesian framework. George and McCulloch (George & McCulloch 1993) introduced a mixture of two normal distributions with a Bernoulli latent variable ϑ. For example in the ZIP model, the priors of the jth coefficient regression β_j in $\boldsymbol{\beta}$ as follow

$$\beta_j | \vartheta_j \overset{i.i.d}{\sim} \vartheta_j \mathrm{N}(0, \sigma_1^2) + (1 - \vartheta_j) \mathrm{N}(0, \sigma_2^2), \quad j = 1, \ldots, p, \tag{10}$$

where $\vartheta_j \sim \text{Bernoulli}(\omega_j)$ with probability of success ω_j, and the two normal distributions have the same zero mean but different variances, such as σ_1^2 is a large value while σ_2^2 is suitably small. We can interpret this model as if $\vartheta_j = 1$, β_j follows the *slab* distribution represented by a normal distribution with a large variance σ_1^2, and the jth term in the model is assumed to have a large effect size. When $\vartheta_j = 0$, β_j follows the *spike* distribution represented by a narrow normal distribution with a small variance σ_2^2, and the jth term in the model is assumed to have a small or zero effect size. Therefore, the spike–and–slab approach imposes two–group mixture priors on the effects and assumes the presence of small effects, which may reflect that the treatment is of low effectiveness for the subject compared to placebo. Such assumptions are relevant from a clinical trial point of view. Note that the spike and slab prior would also be used for covariates other than treatment.

The prior hierarchy for β_j is completed by choosing a prior for a hyperparameter ω_j, and a common choice is the beta distribution, i.e. $\omega_j \sim \text{Beta}(a_0, b_0)$. In practice, it can be difficult to determine the values for $\sigma_1^2, \sigma_2^2, a_0, b_0$. An alternative approach is to use a continuous bimodal distribution (Ishwaran & Rao 2005) in place of the mixture of two normal distributions, and the full prior specification for β_j can be written as

$$\begin{aligned} \beta_j|\vartheta_{\beta_j} &\overset{i.i.d}{\sim} \text{N}(0, \vartheta_{\beta_j}\sigma^2_{\beta_j}), \quad j = 1, \ldots, p. \qquad (11) \\ \sigma^2_{\beta_j}|a_{\beta_j,1}, b_{\beta_j,1} &\overset{i.i.d}{\sim} \text{Inverse-Gamma}(a_{\beta_j,1}, b_{\beta_j,1}), \\ \vartheta_{\beta_j}|v_{\beta_j,0}, \omega_{\beta_j} &\overset{i.i.d}{\sim} \omega_{\beta_j} + (1-\omega_{\beta_j})v_{\beta_j,0}, \\ \omega_{\beta_j} &\sim \text{Uniform}[0,1], \end{aligned}$$

where $v_{\beta_j,0}$ is a small positive value near zero, say $v_{\beta_j,0} = 0.001$, to create a spike. We set $a_{\beta_j,1} = b_{\beta_j,1} = 0.5$ resulting in a vague prior on $\sigma^2_{\beta_j}$ so that the variance of the slab is estimated based on the data. The marginal distribution of the slab component of the mixture becomes a Cauchy distribution with heavy tails under vague prior assumption. The mixing parameter ω_{β_j} is uniformly distributed on $[0, 1]$, so it allows prior information for each coefficient to consist of a mixture of the spike and slab, with each component weighted by the uniform probabilities ω_{β_j}. Therefore, it can handle the model complexity compared to manually assigned priors in the mixture of two normal distributions. Furthermore, the posterior mean of ω_{β_j} is used to estimate the posterior inclusion probability (PIP), which provides nonnull evidence for coefficients in the model.

The main advantage of the spike–and–slab priors is that they provide flexibility in controlling the degree of sparsity in our ZIP model by adjusting the weight ω_{β_j} of the spike in the mixture. We can carry out inference for the latent binary variables ω_{β_j} to identify which corresponding model coefficients are actually different from zero. Moreover, spike–and–slab priors have a closed–form convolution with the Gaussian distribution compared to other continuous shrinkage priors, such as Horseshoe (Carvalho et al. 2009). This advantage allows us to use approximate inference methods for BCSs based on Gaussian approximations in Section 4.4. For the other coefficients $\{\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\gamma}\}$, we adapt a continuous bimodal prior distribution in Equation 11 which has similar prior specifications to $\boldsymbol{\beta}$. Posterior distributions of all the unknown parameters can be obtained via Gibbs sampling along with data augmentation (more details in the Supplementary material). Note that the *brms* R package (Bürkner 2017) can be used to fit the ZIP model (even with horseshoe priors).

4.4 Credible subgroup estimation

Given the posterior mean of $\Delta(\mathcal{X}_i)$ (denoted as $\hat{\Delta}(\mathcal{X}_i)$), we construct credible subgroups (D, S), which bound the true benefiting subgroup B and handle multiplicity in testing $\Delta(\mathcal{X}_i)$ at each covariate point. Particularly, we determine the simultaneous credible bands for $\Delta(\mathcal{X}_i)$ over the covariate space $\boldsymbol{\mathcal{C}}$ by

$$\Delta(\mathcal{X}_i) \in \hat{\Delta}(\mathcal{X}_i) \pm \sqrt{W_\alpha Var(\Delta(\mathcal{X}_i))}, \tag{12}$$

where W_α is the $1-\alpha$ quantile of the empirical distribution $W = \sup_{\mathcal{X}_i, \mathcal{Z}_i \in \boldsymbol{\mathcal{C}}} \frac{(\Delta(\mathcal{X}_i) - \hat{\Delta}(\mathcal{X}_i))^2}{Var(\Delta(\mathcal{X}_i))}$. The exclusive subgroup D is defined by the upper bound of the simultaneous credible band and is expressed by

$$D = \{\mathcal{X}_i \in \boldsymbol{\mathcal{C}} : \hat{\Delta}(\mathcal{X}_i) - \sqrt{W_\alpha Var(\Delta(\mathcal{X}_i))} > \delta\}, \tag{13}$$

which shows that D contains *only* covariate points $\mathcal{X}_i$ for which the characteristics of subjects benefit from the treatment.

Moreover, the inclusive subgroup S is defined by the lower bound of the simultaneous credible band, i.e.

$$S = \{(\mathcal{X}_i) \in \boldsymbol{\mathcal{C}} : \hat{\Delta}(\mathcal{X}_i) + \sqrt{W_\alpha Var(\Delta(\mathcal{X}_i))} \geq \delta\}, \tag{14}$$

which includes *all* types of subjects who benefit. The credible subgroups S and D in Equation 13 and Equation 14 are then obtained via Gaussian approximation because the posterior distributions are approximately normally distributed in our application. In the non–Gaussian case, one can use a quantile–based simultaneous credible band method (Schnell et al. 2018).

5 Simulation Study

In this section, we conduct simulation studies to investigate the performance of our methodology for zero–inflated count data and to compare it with alternative approaches in different simulation settings. Our simulations examined the effects of sample size, strengths of associations between outcome and covariates, and a large number of covariates.

5.1 Simulation setups and evaluation criteria

We generate 1,000 samples $\boldsymbol{\mathcal{D}} = \{y_i, \psi_i, \mathcal{X}_i = (\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i, \boldsymbol{S}_i)\,; i = 1, \ldots, n\}$. For each sample, the zero-inflated count response y_i is sampled from the model in Equation 2 with

$$y_i \sim \begin{cases} 0 & \text{with probability } \theta_i, \\ \text{Poisson}(\mu_i) & \text{with probability } 1 - \theta_i, \end{cases} \tag{15}$$

$$\log(\mu_i) = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \gamma_1 \psi_i + \gamma_2 \psi_i w_{1i} + \gamma_3 \psi_i w_{2i}, \tag{16}$$

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \tau_1 + \tau_2 z_{1i} + \tau_3 z_{2i} + \phi_1 \psi_i + \phi_2 \psi_i s_{1i} + \phi_3 \psi_i s_{2i}. \tag{17}$$

For the zero component, the covariate z_{1i} is drawn from the Bernoulli distribution with probability of success 0.5. We let $z_{2i} = x_{2i}$, i.e., the covariate is allowed to be the same in the two components of a ZIP model. For simplicity, we assume that $\boldsymbol{W}_i = \boldsymbol{X}_i$ and $\boldsymbol{S}_i = \boldsymbol{Z}_i$. For the Poisson component, the covariate x_{1i} is drawn from the Bernoulli distribution with probability of success 0.5, x_{2i} is generated from a uniform distribution on the interval $(-4, 4)$, and ψ_i is a binary treatment and generated as 0 or 1 with equal probability at random. Moreover, the true regression coefficients are set to be $\boldsymbol{\tau} = (-1, 0, 0.5)^\intercal$, $\boldsymbol{\phi} = (-0.5, 0, -0.1)^\intercal$ to avoid a separation problem resulting in an infinite estimate in logistic regression (Albert & Anderson 1984), and $\boldsymbol{\beta} = (1, 0, -0.3, 0.1, 0, 0.5)^\intercal$. Therefor, the covariates x_1, z_1 and s_1 are not relevant in our simulated data.

We examine the performance of our proposed method under various settings of the model parameter $\boldsymbol{\gamma}$ as follows: (S1) null case by setting $\boldsymbol{\gamma} = \boldsymbol{\phi} = (0, 0, 0)^\intercal$, i.e., there is no benefit to the treatment; (S2) small effect size $\boldsymbol{\gamma} = (0.1, 0, 0.5)^\intercal$; (S3) moderate effect size $\boldsymbol{\gamma} = (0.5, 0, 0.5)^\intercal$; and (S4) larger effect size $\boldsymbol{\gamma} = (0.7, 0, 0.5)^\intercal$. Figure 2 illustrates the true CATE as a function over the covariate $x_2 \in (-4, 4)$. The horizontal line represents no treatment effectiveness in scenario S1, and the CATE values above (below) this line indicate that there is (not) a benefit to a treatment. For example, any subjects with $x_2 \in (0, 4)$ will have benefit from treatment in scenarios S2, S3 and S4. Moreover, we further examine the behavior of BCSs in a full case (S5) in which all subjects are beneficial to the treatment. In scenario S5, we assume a moderate effect size as in S3 and x_{2i} followed a uniform distribution from 0 to 4. We vary the sample size $n = 50, 100$ and 500, which we refer to as small–to–moderate sample size. Note that in our simulation studies, the choice of sample sizes is relevant to the bladder tumor dataset with $n = 87$, and similarly for the parameter values. With these simulation settings, the means of Poisson component range from 0.86 to 11.33, and on average, 70% of the responses y_i are zero and 10% of the zeros are Poisson.

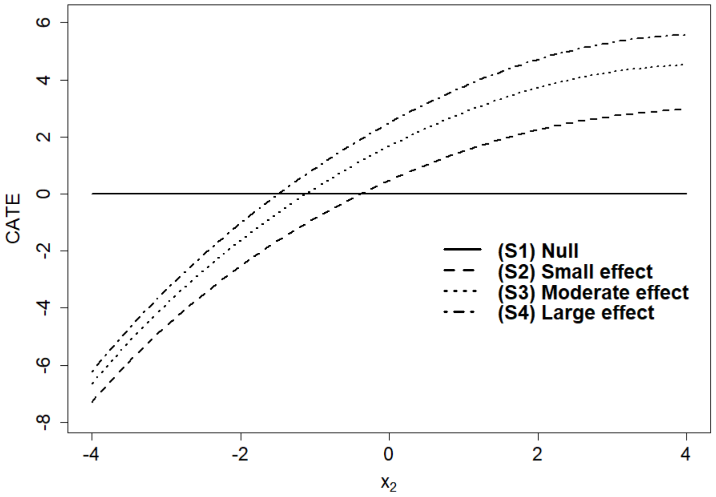


Figure 2: Illustration of CATE for various values of x_2 in simulation study. The horizontal line indicates that there is no treatment effect (S1). The CATE values above (below) the horizontal line show that there is (not) a benefit to treatment.

As competitors, we compare the proposed method using spike–and–slab prior (denoted as BCS–SS) to the BCS without using spike–and–slab prior (denoted as BCS), a horseshoe prior (denoted as BCS–HS) and pointwise method (denoted as PW). For BCS–HS approach, we use the same ZIP model as BCS–SS for constructing credible subgroups, but the spike–and–slab priors for regression coefficients are replaced by a horseshoe prior. The full prior specification is

$$\beta_j \mid \tau, Z_j \sim \mathrm{N}\left(0, Z_j^2\tau^2\right) \tag{18}$$

$$\tau, Z_j \sim \text{Half–Cauchy}(1),$$

where the scale parameter of 1 for the Half–Cauchy distribution is the default choice given in Carvalho et al. (Carvalho et al. 2009). In contrast to BCS–SS, the BCS-HS approach cannot provide explicit estimates for the inclusion probabilities. Moreover, the PW method also uses the same posterior samples from BCS–SS for constructing credible subgroups, but without correcting for multiplicity. Specifically, we formulate the credible subgroup D by identifying covariate points $\mathcal{X}_i$ from a covariate space such that $P\left(\Delta(\mathcal{X}_i) > \delta \mid \boldsymbol{\mathcal{D}}\right) > 1 - \alpha$, and the credible subgroup S is constructed by finding the covariate points having $P\left(\Delta(\mathcal{X}_i) \leq \delta \mid \boldsymbol{\mathcal{D}}\right) \leq \alpha$. For each simulated data, we run each MCMC chain for 10,000 iterations with the first half taken as burn–in. In addition, we set a credible level $\alpha = 0.8$ and a threshold $\delta = 0$.

We assessed the performance of the BCSs approach by using the same criteria as in Schnell et al. (2016) by calculating four quantities: (1) the total coverage which provides frequency that the true benefiting subgroup B is in (D, S); (2) the credible pair size which measures the proportion of population in the uncertainty region $S \backslash D$; (3) Specificity and sensitivity of D providing diagnostic accuracy of subgroup D with respect to B; and (4) Mean square error (MSE) of CATEs comparing the estimated treatment effect with the true values. In our simulation study, we address the multiplicity issues by using the total coverage. Since this coverage metric is the frequency of the true benefiting subgroup B belonging to both exclusive and inclusive credible subgroups, an approach without encountering multiplicity would have a total coverage below the nominal size, i.e., a coverage failure corresponds to a family–wise error.

5.2 Simulation results

Table 1 summarizes the average summary statistics for scenarios S2, S3 and S4 with sample size $n = 50, 100$ and 500 at 80% credible level and δ=0, and we report the results of scenario S1 and S5 in Table S1 (provided in Supplemental Document). Note that when the benefiting subgroup is empty in scenario S1, the sensitivity of D, which is the proportion of the exclusive subgroup D that is also contained in the benefiting subgroup B, is not calculable (denoted as 'NaN' in the Table S1). Similarly, the specificity of D is not provided in scenario S5 in which all subjects benefit from the treatment. For the total coverage, since the benefiting subgroup B would be empty in scenario S1, it would only be covered if the exclusive subgroup D is also empty. However, the benefiting subgroup B would correspond to the whole population in

Table 1: Average summary statistics for sample size $n = 50, 100$ and 500 at 80% credible level and δ=0 for scenarios: S2 (small effect size), S3 (moderate effect size) and S4 (large effect size).

Sample size	Scenario	Method	Total coverage	Credible pair size	Specificity of D	Sensitivity of D	MSE
50	S2	BCS	0.81	0.53	0.84	0.38	4.45
		BCS-SS	0.84	0.48	0.87	0.47	3.63
		BCS-HS	0.84	0.53	0.86	0.41	3.40
		PW	0.41	0.33	0.79	0.63	3.63
	S3	BCS	0.83	0.51	0.85	0.39	4.41
		BCS-SS	0.88	0.52	0.89	0.47	3.32
		BCS-HS	0.88	0.52	0.89	0.47	3.31
		PW	0.43	0.31	0.82	0.66	3.32
	S4	BCS	0.83	0.49	0.87	0.55	3.69
		BCS-SS	0.90	0.44	0.93	0.60	2.32
		BCS-HS	0.90	0.46	0.91	0.61	2.13
		PW	0.46	0.23	0.85	0.75	2.32
100	S2	BCS	0.84	0.53	0.86	0.41	1.66
		BCS-SS	0.86	0.44	0.89	0.67	1.56
		BCS-HS	0.86	0.45	0.88	0.61	1.53
		PW	0.46	0.31	0.82	0.73	1.56
	S3	BCS	0.85	0.39	0.89	0.71	1.44
		BCS-SS	0.89	0.38	0.92	0.80	1.25
		BCS-HS	0.89	0.38	0.90	0.79	1.21
		PW	0.55	0.17	0.85	0.88	1.25
	S4	BCS	0.85	0.27	0.93	0.77	1.42
		BCS-SS	0.89	0.22	0.95	0.88	1.29
		BCS-HS	0.90	0.21	0.95	0.84	1.28
		PW	0.69	0.12	0.87	0.94	1.29
500	S2	BCS	0.87	0.18	0.91	0.82	0.62
		BCS-SS	0.90	0.14	0.92	0.86	0.59
		BCS-HS	0.90	0.17	0.92	0.85	0.48
		PW	0.69	0.03	0.88	0.92	0.59
	S3	BCS	0.89	0.14	0.98	0.89	0.61
		BCS-SS	0.91	0.13	0.98	0.92	0.57
		BCS-HS	0.90	0.12	0.98	0.90	0.56
		PW	0.71	0.05	0.97	0.94	0.57
	S4	BCS	0.89	0.13	0.98	0.91	0.59
		BCS-SS	0.93	0.07	0.99	0.96	0.44
		BCS-HS	0.93	0.09	0.99	0.95	0.43
		PW	0.72	0.04	0.97	0.98	0.44

scenario S5, and it would only be covered if the inclusive subgroup S would need to correspond to the whole population.

In general, we observe that the total coverage, specificity and sensitivity of D increase as the sample size and effect size increase, and the decrease in credible pair size and MSE with the increase of sample size and and effect size. It appears that the BCS–HS and BCS–SS approaches have similar trends in all scenarios for all criteria, and these approaches outperform the BCS approach without shrinkage priors across all scenarios as expected. Moreover, the BCS–HS and BCS–SS approaches show more efficient performance than the PW approach in terms of total coverage, specificity of D, especially in the small sample size $n = 50$. For example, the total coverage is near or higher than the nominal level for all scenarios, whereas the total coverage of PW ranges from 0.41 to 0.46.

Moreover, the BCS–HS and BCS–SS approaches generally yield high specificity of D compared to the PW approach because the PW approach has smaller credible pair size, i.e. tighter uncertainty region $S \backslash D$, resulting in smaller total coverage and specificity of D. In contrast, the PW approach obtains the specificity of D larger than that of the BCS–HS and BCS–SS approaches. This phenomenon reflects the trade–off between specificity (one minus the type I error rate) and sensitivity (statistical power) under the multiple hypothesis testing setting, i.e., increased specificity reduces the sensitivity and vice-versa. Our proposed method can achieve high specificity of D in exchange for a reduced sensitivity of D, especially for small samples. In a regulatory setting, clinicians and researchers, who prefer a greater specificity but a slightly lower sensitivity, might consider this trade–off. The simulation results show that, overall, the proposed approach BCS with shrinkage priors has the advantage of controlling multiplicity issues by providing a total coverage above nominal level 0.8, high specificity $(87\% - 99\%)$ of D in different scenarios and relatively high sensitivity (60%) of D for small sample size with large size effect.

In Supplemental material, we use different link functions for the zero component (Equation 17) under scenario S3 to study to what extent our proposed models are sensitive to the link function misspecification. We found that the performance of the BCS is not sensitive to the choice of the link function for the zero component. Furthermore, we expand the scenario S3 to investigate the scalability of our proposed models in the presence of high–dimensional covariates (detailed in Supplemental material). Briefly, we set $n = 200$, and the total dimension $\eta = p + k + q + m = 300$, 500 and $1,000$, where the number of non-zero coefficients is

Table 2: Average summary statistics for high–dimensional setting at sample size $n = 200$, 80% credible level and δ=0.

Total dimension η	Method	Total coverage	Credible pair size	Specificity of D	Sensitivity of D	MSE
300	BCS-SS	0.81	0.52	0.78	0.69	1.33
	BCS-HS	0.81	0.52	0.78	0.69	1.31
	PW	0.54	0.36	0.55	0.79	1.33
500	BCS-SS	0.77	0.49	0.66	0.62	2.55
	BCS-HS	0.76	0.49	0.66	0.62	2.57
	PW	0.51	0.31	0.34	0.73	2.55
1,000	BCS-SS	0.74	0.44	0.61	0.54	3.24
	BCS-HS	0.74	0.44	0.60	0.54	3.25
	PW	0.48	0.26	0.22	0.62	3.24

constant. Under these settings, the number of parameters being estimated is greater than, or equals to, the sample size. Table 2 shows that we yield similar results to those described in the low–dimensional setting above. Moreover, increasing total dimension η, as expected, leads to increase the MSE. For the BCS-SS and BCS-HS approaches, the total coverage rates lay between 0.74 and 0.81, and these approaches achieve both moderate specificity $(60\% - 78\%)$ and sensitivity $(54\% - 69\%)$ of D. Hence, the simulation study confirms the advantage of our proposed methodology in high–dimensional settings.

6 Analysis of the bladder tumor dataset

The proposed BCSs method was originally motivated by the clinical bladder tumor study conducted by the Veterans Administration Cooperative Urological Research Group (VACURG) (Byar et al. 1977). The data collection is described in Byar (1980). All patients at the beginning of the trial had experienced superficial bladder tumors, which were removed through a transurethral resection. Following surgery, participants were randomly assigned to receive either pyridoxine pills along with periodic instillation of a chemotherapeutic agent thiotepa into the bladder or a placebo. Several statistical methodologies were proposed in various literature (Wellner & Zhang 2000, Baetschmann & Winkelmann 2013, Sun & Wei 2000). They found that periodic instillation of thiotepa significantly reduced the recurrence of bladder tumors compared to placebo.

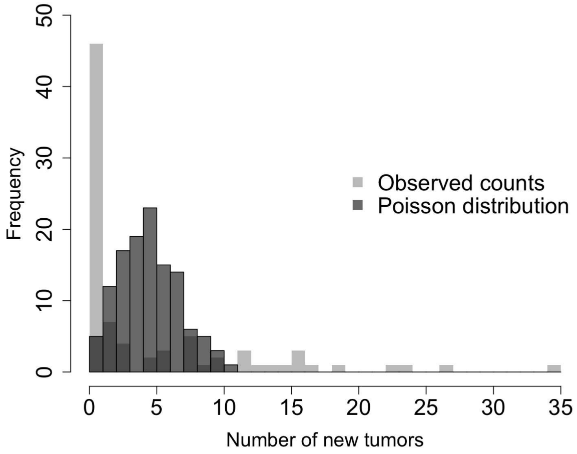


Figure 3: A light gray histogram represents the observed number of new tumors in the bladder tumor study, and a dark gray histogram refers to a Poisson distribution with mean 4.73, based on the the empirical mean of the data.

As an illustration of our proposed method, our goal is to identify characteristics of patients who benefit from thiotepa treatment. The endpoints of interest were the number of new tumors recorded over the entire observed patient record time. We have a total of 38 patients receiving the thiotepa treatment and 47 patients receiving the placebo. Following Baetschuman et al (Baetschmann & Winkelmann 2013), we include the following covariates in our analysis: the number of initial tumors (INITNR) which ranges from 1 to 8, the treatment indicator (trt) and the tumor size (Size). The natural log of duration of exposure measured in months, log(time), is considered as an offset variable.

Figure 3 shows the distribution of the total number of tumors, and it is highly peaked at zero. Therefore, the excessive zeros provide evidence of zero inflation, which support our use of the ZIP regression model for count data with excess zeros. We then consider two different link functions for the zero component in the ZIP model, including the common logit link, i.e.

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \tau_0 + \tau_1 \mathrm{INITNR}_i + \tau_2 \mathrm{Size}_i + \phi_1 \mathrm{trt}_i + \phi_2 \mathrm{INITNR}_i \times \mathrm{trt}_i + \phi_3 \mathrm{Size}_i \times \mathrm{trt}_i, \quad (19)$$

and the complementary log–log link, i.e.

$$\theta_i = \exp\left(-\exp(\tau_0 + \tau_1 \text{INITNR}_i + \tau_2 \text{Size}_i + \phi_1 \text{trt}_i + \phi_2 \text{INITNR}_i \times \text{trt}_i + \phi_3 \text{Size}_i \times \text{trt}_i)\right). \tag{20}$$

Moreover , the model for the Poisson component is given by

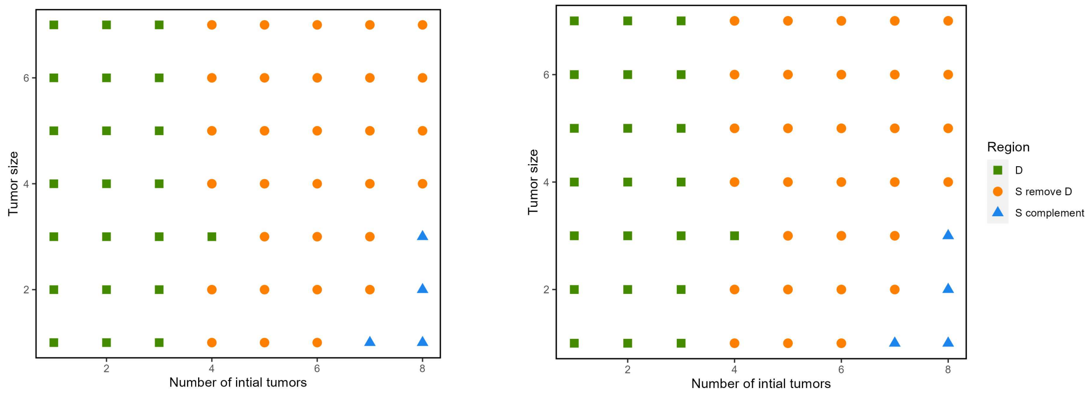
$$\log(\mu_i) = \log(\text{time}_i) + \beta_0 + \beta_1 \text{INITNR}_i + \beta_2 \text{Size}_i + \gamma_1 \text{trt}_i + \gamma_2 \text{INITNR}_i \times \text{trt}_i + \gamma_3 \text{Size}_i \times \text{trt}_i. \tag{21}$$

To identify characteristics of patients who benefit from the thiotepa treatment, we only assign spike and slab priors (described in Section 4.3) on regression coefficients corresponding to INITNR and Size for both components. We then assign $\mathrm{N}(0, 1000)$ prior on other coefficients in the model, e.g., τ_0, β_0, ϕ_1, and γ_1. We run the MCMC algorithm for $10,000$ draws discarding a burn–in of $5,000$, and the convergence of the MCMC sampler was satisfactory based on examination of trace plots (provided in Supplemental material). We then report the posterior summary statistics of the regression coefficients for both models in Table 3. It appears that both approaches provide similar estimates of coefficient in Poisson component, and the significant treatment effect indicates that the thiotepa treatment suppresses the number of bladder tumors compared to placebo at a 90% credible level. The PIP were roughly equal for all covariates under logit and cloglog links. For the zero component, INITNR has higher inclusion probability than that of SIZE, and similar results are found for interaction terms. For the Poisson component, INITNR, identified as a prognostic covariate, has substantially smaller effect size and PIP than those of a predictive covariate (identified as interacting with the treatment). On the other hand, SIZE has larger effect size and PIP than those of a predictive covariate. The results suggest that the prognostic effect of covariate INITNR and the predictive effect of covariate SIZE may be irrelevant for identifying the characteristics of patients for whom the thiotepa treatment is beneficial, and thus the sparsity in covariates would ease the interpretation of our proposed BCSs.

Using the above MCMC sampling results, we construct BCSs by setting $\delta = 0$ and credible level of 80%. Figures 4A and 4B show the credible subgroups for the ZIP model using logit and cloglog links, respectively. Overall, we obtain similar subgroups for both approaches. Each point in each panel represents a particular type of patient with their tumor size and number of initial tumors. The interpretation of BCSs in Figure 4 is that the rectangles represent characteristics of patients for whom the thiotepa treatment is beneficial, whereas the triangles represent

Table 3: Posterior summaries of covariates in Bladder Tumor dataset for ZIP model. Posterior standard (SD) deviation are in parentheses, and (*) denotes significance at 90% credible level.

Component	Parameter	Posterior Mean (Posterior SD)		PIP	
		logit	cloglog	logit	cloglog
Zero component	Intercept	-0.026 (0.701)	-0.27 (0.382)	1	1
	trt	0.81 (0.823)	-0.542 (0.564)	1	1
	INITNR	-0.23 (0.251)	0.11 (0.127)	0.567	0.367
	Size	-0.089 (0.18)	0.034(0.084)	0.297	0.152
	INITNR × trt	-0.156 (0.282)	0.118 (0.165)	0.417	0.462
	Size × trt	0.006 (0.141)	-0.038 (0.128)	0.249	0.231
Poisson component	Intercept	-1.105 (0.141)*	1.098 (0.122)*	1	1
	trt	-1.361 (0.284)*	-1.418 (0.226)*	1	1
	INITNR	0.010 (0.027)	0.007 (0.022)	0.064	0.066
	Size	-0.089 (0.045)	-0.094 (0.041)	0.551	0.616
	INITNR × trt	0.302 (0.048)*	0.298 (0.041)*	0.932	0.973
	Size × trt	-0.102 (0.124)	-0.086 (0.109)	0.307	0.261



(A) ZIP model with logit link

(B) ZIP model with cloglog link

Figure 4: The Bayesian credible subgroups for bladder tumor dataset.

types of patients who have no benefit. Therefore, we have evidence to conclude that patients with the number of initial bladder tumors lower than 5 and tumor size between 1–7 are benefiting from the thiotepa treatment. However, patients with 7 initial bladder and tumor sizes between 1–2 are not benefiting from the thiotepa treatment, and the results are similar to those with 8 initial bladder tumors and tumor sizes between 1–3. We further investigate our model choice of ZIP regression by comparing to zero–inflated negative binomial (ZINB) regression, which can accommodate excess zeros and overdispersion. As in the ZIP model, we yield similar BCSs results for ZINB model (see Supplementary material).

7 Discussion

In this article, we have introduced the BCSs for count data with excess zeros which are common in medical and public health related studies. Our approach provides insight into the apparent heterogeneity of treatment by identifying characteristics of patients who benefit from the treatment, while handling multiplicity issues. The method studied here is widely applicable as post–hoc analysis for confirmatory clinical trials, which focus on assessment of benefits and risks of new drugs compared to standard treatments.

As shown in the numerical studies, our method achieves desirable frequentist properties such as the total coverage, sensitivity and specificity of exclusive group D in low and high dimensional covariates. The advantage of BCSs is that it provides simultaneous inferences, as opposed to non-simultaneous inferences available from tree-based methods, from a pair of credible subgroups (D, S) where D is contained by the benefiting subgroup and S contains the benefiting subgroup. In addition, our strategy includes shrinkage priors, which screen the covariates to find a lower dimensional covariate space, resulting in improving the estimation and interpretation of BCSs. This renders the credible subgroups interpretable and useful in practice.

We note that in the construction of Bayesian credible subgroups, we have assumed that the threshold δ is zero for controlling the benefiting subgroups. If the alternative treatments have unequal cost, we need to find the covariates of the patients in which their CATEs exceed δ. Ideally, the choice of δ should be a clinically meaningful value determined by the subject matter experts. When such a threshold is not available, based on the posterior distribution of CATE, one can compare the posterior mean or median with a range of threshold values and decide the treatment strategy with the smallest threshold. Another possible approach is to conduct sensitivity analysis to evaluate the changes in subgroups at different thresholds.

There are several possible extensions of our approach. The first extension is model choice for zero-inflated data. Our method consists of a two–step approach which includes building the statistical model and constructing the Bayesian credible subgroups. Thus it is straightforward to implement our method with different models for zero-inflated data, such as compound Poisson random effect (Ma et al. 2009) or marginal zero inflated regression models (Martin & Hall 2017). The second extension is comparing different metrics for the CATE. We note that our metric for the CATE is the difference in expected counts, and the CATE can be measured by the relative ratio. It is worth noting that the treatment effect heterogeneity is scale–dependent, i.e., the treatment effects may be heterogenous on one scale (difference) but not on another (ratio). We believe that the metric for CATE has an important impact on the estimation and interpretation of the BCSs. Another future direction to consider are multiple endpoints for both efficacy and risk. For example, one can extend our method for benefiting subgroup estimation for multiple endpoints by using the concept of admissibility (Schnell 2017).

SUPPLEMENTARY MATERIAL

The supplementary materials for Bayesian shrinkage estimation of credible subgroups for count data with excess zeros includes: (1) sampling algorithm for posterior distribution using spike and slab priors; (2) simulation results for the null case S1 and the full case S5; (3) simulation settings for different link functions; (4) simulation settings for high–dimensional covariates; and (5) the MCMC diagnostics and ZINB for the bladder tumor data analysis.

8 Acknowledgment

The authors are very grateful to the associate editor and two reviewers for their useful comments and suggestions, which we believe have improved our work. In addition, the authors would like to thank Mr. Greg Ball (Biostatistics and Research Decision Sciences, Merck & Co., Inc.) for the constructive suggestions and comments.

9 Disclosure statement

The authors report there are no competing interests to declare.

References

Albert, A. & Anderson, J. A. (1984), 'On the existence of maximum likelihood estimates in logistic regression models', *Biometrika* **71**(1), 1–10.

Albert, J. M., Wang, W. & Nelson, S. (2014), 'Estimating overall exposure effects for zero-inflated regression models with application to dental caries', *Statistical methods in medical research* **23**(3), 257–278.

Baetschmann, G. & Winkelmann, R. (2013), 'Modeling zero-inflated count data when exposure varies: With an application to tumor counts', *Biometrical Journal* **55**(5), 679–686.

Ballarini, N. M., Rosenkranz, G. K., Jaki, T., König, F. & Posch, M. (2018), 'Subgroup identification in clinical trials via the predicted individual treatment effect', *PloS one* **13**(10), e0205971.

Beckman, R. A., Clark, J. & Chen, C. (2011), 'Integrating predictive biomarkers and classifiers into oncology clinical development programmes', *Nature reviews Drug discovery* **10**(10), 735–748.

Berger, J. O., Wang, X. & Shen, L. (2014), 'A bayesian approach to subgroup identification', *Journal of biopharmaceutical statistics* **24**(1), 110–129.

Berry, D. A. (1990), 'Subgroup analyses'.

Breiman, L. (1996), 'Heuristics of instability and stabilization in model selection', *The annals of statistics* **24**(6), 2350–2383.

Byar, D. (1980), The veterans administration study of chemoprophylaxis for recurrent stage i bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa, *in* 'Bladder tumors and other topics in urological oncology', Springer, pp. 363–370.

Byar, D., Blackard, C., Group, V. A. C. U. R. et al. (1977), 'Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage i bladder cancer', *Urology* **10**(6), 556–561.

Bürkner, P.-C. (2017), 'brms: An R package for Bayesian multilevel models using Stan', *Journal of Statistical Software* **80**(1), 1–28.

Cai, T., Tian, L., Wong, P. H. & Wei, L. (2011), 'Analysis of randomized comparative clinical trial data for personalized treatment selections', *Biostatistics* **12**(2), 270–282.

Cameron, A. C. & Trivedi, P. K. (2013), *Regression analysis of count data*, Vol. 53, Cambridge university press.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2009), Handling sparsity via the horseshoe, *in* 'Artificial intelligence and statistics', PMLR, pp. 73–80.

Cohen, A. C. (1963), *Estimation in mixtures of discrete distributions*, Statistical Pub. Society.

Cui, L., James Hung, H., Wang, S. J. & Tsong, Y. (2002), 'Issues related to subgroup analysis in clinical trials', *Journal of biopharmaceutical statistics* **12**(3), 347–358.

Ding, P., Li, X. & Miratrix, L. W. (2017), 'Bridging finite and super population causal inference', *Journal of Causal Inference* **5**(2), 20160027.

Foster, J. C., Taylor, J. M. & Ruberg, S. J. (2011), 'Subgroup identification from randomized clinical trial data', *Statistics in medicine* **30**(24), 2867–2880.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for bayesian variable selection', *Statistica sinica* pp. 339–373.

Geweke, J. (1996), 'Variable selection and model comparison in regression', *In Bayesian Statistics 5* .

Ghosh, S. K., Mukhopadhyay, P. & Lu, J.-C. J. (2006), 'Bayesian analysis of zero-inflated regression models', *Journal of Statistical planning and Inference* **136**(4), 1360–1375.

Havlir, D. V., Kendall, M. A., Ive, P., Kumwenda, J., Swindells, S., Qasba, S. S., Luetkemeyer, A. F., Hogg, E., Rooney, J. F., Wu, X. et al. (2011), 'Timing of antiretroviral therapy for hiv-1 infection and tuberculosis', *New England Journal of Medicine* **365**(16), 1482–1491.

Hernán, M. A. (2016), 'Does water kill? a call for less casual causal inferences', *Annals of epidemiology* **26**(10), 674–680.

Hilbe, J. M. (2011), *Negative binomial regression*, Cambridge University Press.

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Ishwaran, H. & Rao, J. S. (2005), 'Spike and slab variable selection: frequentist and bayesian strategies', *The Annals of Statistics* **33**(2), 730–773.

Lagakos, S. W. et al. (2006), 'The challenge of subgroup analyses-reporting without distorting', *New England Journal of Medicine* **354**(16), 1667.

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

Lamont, A., Lyons, M. D., Jaki, T., Stuart, E., Feaster, D. J., Tharmaratnam, K., Oberski, D., Ishwaran, H., Wilson, D. K. & Van Horn, M. L. (2018), 'Identification of predicted individual treatment effects in randomized clinical trials', *Statistical methods in medical research* **27**(1), 142–157.

Ma, R., Hasan, M. T. & Sneddon, G. (2009), 'Modelling heterogeneity in clustered count data with extra zeros using compound poisson random effect', *Statistics in medicine* **28**(18), 2356–2369.

Martin, J. & Hall, D. B. (2017), 'Marginal zero-inflated regression models for count data', *Journal of Applied Statistics* **44**(10), 1807–1826.

Mitchell, T. J. & Beauchamp, J. J. (1988), 'Bayesian variable selection in linear regression', *Journal of the american statistical association* **83**(404), 1023–1032.

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *Journal of econometrics* **33**(3), 341–365.

Neelon, B., O'Malley, A. J. & Smith, V. A. (2016), 'Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview', *Statistics in Medicine* **35**(27), 5070–5093.

Ngo, D., Baumgartner, R., Mt-Isa, S., Feng, D., Chen, J. & Schnell, P. (2020), 'Bayesian credible subgroup identification for treatment effectiveness in time-to-event data', *PloS one* **15**(2), e0229336.

Qian, M. & Murphy, S. A. (2011), 'Performance guarantees for individualized treatment rules', *Annals of statistics* **39**(2), 1180.

Rodrigues, J. (2003), 'Bayesian analysis of zero-inflated distributions', *Communications in Statistics-Theory and Methods* **32**(2), 281–289.

Rose, C. E., Martin, S. W., Wannemuehler, K. A. & Plikaytis, B. D. (2006), 'On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data', *Journal of biopharmaceutical statistics* **16**(4), 463–481.

Rubin, D. B. (2005), 'Causal inference using potential outcomes: Design, modeling, decisions', *Journal of the American Statistical Association* **100**(469), 322–331.

Schnell, P. M. (2017), Credible subgroups: identifying the population that benefits from treatment, PhD thesis, University of Minnesota.

Schnell, P. M., Müller, P., Tang, Q. & Carlin, B. P. (2018), 'Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials', *Clinical Trials* **15**(1), 75–86.

Schnell, P. M., Tang, Q., Offen, W. W. & Carlin, B. P. (2016), 'A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects', *Biometrics* **72**(4), 1026–1036.

Shiba, K. & Kawahara, T. (2021), 'Using propensity scores for causal inference: pitfalls and tips', *Journal of epidemiology* p. JE20210145.

Su, X., Meneses, K., McNees, P. & Johnson, W. O. (2011), 'Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors', *Journal of the Royal Statistical Society Series C: Applied Statistics* **60**(3), 457–474.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M. & Li, B. (2009), 'Subgroup analysis via recursive partitioning.', *Journal of Machine Learning Research* **10**(2).

Sun, J. & Wei, L. (2000), 'Regression analysis of panel count data with covariate-dependent observation and censoring times', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2), 293–302.

Tanner, M. A. & Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American statistical Association* **82**(398), 528–540.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Todem, D., Kim, K. & Hsu, W.-W. (2016), 'Marginal mean models for zero-inflated count data', *Biometrics* **72**(3), 986–994.

VanderWeele, T. J. (2009), 'Concerning the consistency assumption in causal inference', *Epidemiology* **20**(6), 880–883.

Wagner, G. G., Frick, J. R. & Schupp, J. (2007), The german socio-economic panel study (soep): Scope, evolution and enhancements, Technical report, SOEPpapers on Multidisciplinary Panel Data Research.

Weisberg, H. I. & Pontes, V. P. (2015), 'Post hoc subgroups in clinical trials: Anathema or analytics?', *Clinical trials* **12**(4), 357–364.

Wellner, J. A. & Zhang, Y. (2000), 'Two estimators of the mean of a counting process with panel count data', *The Annals of statistics* **28**(3), 779–814.

Westreich, D. & Cole, S. R. (2010), 'Invited commentary: positivity in practice', *American journal of epidemiology* **171**(6), 674–677.

Winkelmann, R. (2008), *Econometric analysis of count data*, Springer Science & Business Media.

Zhang, B. & Zhang, M. (2022), 'Subgroup identification and variable selection for treatment decision making', *The Annals of Applied Statistics* **16**(1), 40–59.

Zhao, W., Ma, W., Wang, F. & Hu, F. (2022), 'Incorporating covariates information in adaptive clinical trials for precision medicine', *Pharmaceutical Statistics* **21**(1), 176–195.

Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M. R. (2012), 'Estimating individualized treatment rules using outcome weighted learning', *Journal of the American Statistical Association* **107**(499), 1106–1118.

Ziegler, A., Koch, A., Krockenberger, K. & Großhennig, A. (2012), 'Personalized medicine using dna biomarkers: a review', *Human genetics* **131**, 1627–1638.