A Bayesian Approach to Account for Misclassification in Prevalence and Trend Estimation*

Martijn van Hasselt †[1], Christopher R. Bollinger[2], and Jeremy Bray[1]

[1]The University of North Carolina Greensboro, Department of Economics

[2]The University of Kentucky, Department of Economics

*We thank John Pepper and Gary Koop for helpful comments on an earlier version of this paper.

†*Corresponding author.* Address: Bryan School of Business and Economics, Department of Economics, P.O. Box 26170, Greensboro, NC 27412. Tel. (336) 256-1010. Email: mnvanhas@uncg.edu

Summary

In this paper we present a Bayesian approach to estimate the mean of a binary variable and changes in the mean over time, when the variable is subject to misclassification error. These parameters are partially identified and we derive identified sets under various assumptions about the misclassification rates. We apply our method to estimating the prevalence and trend of prescription opioid misuse, using data from the 2002-2014 National Survey on Drug Use and Health. Using a range of priors, the posterior distribution provides evidence that among middle-aged white men, the prevalence of opioid misuse increased multiple times between 2002 and 2012.

Keywords. Misclassification, partial identification, Bayesian estimation

1. INTRODUCTION

In this paper we present a Bayesian approach to estimate the population mean of a binary variable as well as changes in the mean over time, when the variable in question is subject to misclassification error.[1] Our methods can be applied when data is available from either a single cross section or multiple independent cross sections, possibly supplemented with population-level weights. To illustrate our approach, we conduct an empirical analysis of self-reported past-year misuse of opioid pain relievers, using data from the 2002-2014 waves of the National Survey on Drug Use and Health (NSDUH). We focus on opioids because they constitute a pressing public health concern (e.g. Kolodny et al., 2015) and the misreporting of substance use in surveys is a well-documented problem (Fendrich et al., 1999; Biemer and Wiesen, 2002; Ledgerwood et al., 2008; Kroutil et al., 2010; Murphy et al., 2015).[2]

Although the empirical focus in this paper is on substance use, the problem of misclassification arises in many different contexts. For example, Kreider and Pepper (2007) and Gosling and Saloniki (2014) address misclassification in self-reported disability status. Gundersen et al. (2012) and Meyer et al. (2015) document misreporting of participation in the food stamp program (SNAP). Beyond surveys, data from clinical settings or public health surveillance systems are also subject to misclassification error. The recent COVID-19 pandemic has highlighted the difficulties in estimating coronavirus infection rates. While this largely results from sample selection issues, problems with testing accurary, and in particular the potential for false negatives, can give rise to a substantial amount of misclassification (e.g., Li et al., 2020). Simple estimates of the prevalence of health conditions such as HIV, opioid misuse, or COVID-19 can affect policy decisions with widespread health and economic implications. Furthermore, accurate measures of the trend in prevalence of a condition are equally important as they motivate changes in, or even cessation of, these same policies. A failure to account for misclassification errors leads to biased estimates of critical parameters and undermines the development of effective, evidence-based policies.

Models with misclassification error have a long history in statistics and econometrics (e.g., Bross,

[1] Throughout this paper we will refer to the mean as *prevalence* and to changes in the mean over time as *trends*.

[2] Even apparently objective data such as those obtained from death certificates are subject to possible reporting errors as new data systems have been implemented and medical examiners and other officials exercise personal judgment on when to test for or report opioid use as a cause of death (Mertz et al., 2014; Ruhm, 2016; Rudd et al., 2014, 2016).

1954; Tenenbein, 1970). Unless the misclassification probabilities are known or a validation sample (i.e., a set of observations that is known to be correctly classified) is available, the prevalence is completely unidentified but the misclassification probabilities are partially identified (Bollinger and Van Hasselt, 2017a). In a Bayesian model, the lack of identification does not require a different approach to inference. While the likelihood function does not identify every model parameter, the information contained in the prior can still lead to informative posterior distributions (Kadane, 1974). For example, Gaba and Winkler (1992), Joseph et al. (1995), Evans et al. (1996) and Rahme et al. (2000) use beta priors for the misclassification rates to estimate the prevalence, resulting in posterior density intervals that are strictly contained within the unit interval.

Since the influential contributions of Kadane (1974) and Poirier (1998), Bayesian inference in models that are partially identified has been an active area of research (e.g., Poirier and Tobias, 2003; Gustafson et al., 2005; Moon and Schorfheide, 2012; Hahn et al., 2016; Bollinger and Van Hasselt, 2017b). Using the nomenclature of Moon and Schorfheide (2012), the main feature of such models is that the data are fully informative about a reduced-form parameter vector ϕ, in the sense that its posterior becomes more concentrated as more data become available. Conditional on ϕ, however, the data contains no further information about the structural parameter vector θ. The prior of the non-identified and partially identified elements of θ is then updated by the data only to the extent that ϕ and θ are a priori dependent. When θ is partially identified, prior dependence is necessary for any prior that is consistent with the model, because the bounds of the identified set are functions of ϕ. Put differently, the support of the conditional prior of θ is a function of ϕ. The prior dependence between θ and ϕ, combined with Bayesian learning about ϕ, ensures that at least some learning about θ occurs (Poirier, 1998; Poirier and Tobias, 2003).

This paper makes two main contributions. First, we extend the work of Pepper (2001) and derive identified sets for prevalence and trend under a range of assumptions about changes in misclassification over time. Our results show that under sufficiently strong assumptions, the direction (upward or downward) of a trend is identified. Second, we develop a Bayesian approach to inference, where the identified sets are used to specify a range of priors that researchers might entertain in practice. The posterior distribution takes a simple form and random samples from it are easily generated. While our development focuses on estimating prevalence and trends from repeated cross sections, a case that is relevant for many nationally representative survey samples, our approach can

be valuable in other settings as well. In the supplemental appendix, we discuss how our methods might be adapted for use in regression models.

In partially identified models, bounds are often only informative under strong assumptions. For example, to identify the direction of a trend, Pepper (2001) assumes that the probability of misclassification error has a known upper bound. In practice, there can be considerable uncertainty about the appropriate value of such a bound. An advantage of a Bayesian approach is that this uncertainty can be incorporated into the prior. Additionally, information contained in the prior can lead to more precise inference relative to a classical bounds analysis. This is apparent in our empirical analysis. The classical bounding approach identifies the direction of the trend only under very strict assumptions. Without these assumptions, the estimated bounds move farther apart and become practically useless. In contrast, for a range of prior distributions and assumptions, the Bayesian posterior provides strong evidence that among middle-aged white men, the prevalence of opioid misuse increased several times between 2002 and 2012.

The remainder of this paper is organized as follows. In Section 2 we discuss the misclassification model and the identified sets for the prevalence and trend under different assumptions about the misclassification rates. Section 3 discusses a range of prior distributions and shows how to draw a sample from the posterior. Section 4 presents our empirical analysis and Section 5 concludes. A supplemental appendix contains more details about the identified sets, presents additional empirical results for select subgroups, and suggests ways for extending our methods to a regression context.

2. THE MODEL

2.1. The Misclassification Problem

The model we present here is based on Bollinger and Van Hasselt (2017a), extended to the case of a repeated cross section. Let $Y_{it}^* = 0, 1$ be the true value of a binary indicator for individual $i = 1, \ldots, n_t$ in time period $t = 1, \ldots, T$, and let $\pi_t = E(Y_{it}^*)$ be its mean (the true prevalence). Instead of Y_{it}^*, we observe a possibly misclassified variable $Y_{it} = 0, 1$, where $p_t = \Pr(Y_{it} = 1 | Y_{it}^* = 0)$ is the probability of a false positive and $q_t = \Pr(Y_{it} = 0 | Y_{it}^* = 1)$ the probability of a false negative.

The observed prevalence $\mu_t = E(Y_{it})$ is related to (π_t, p_t, q_t) through the equation

$$\mu_t = \pi_t(1 - q_t) + (1 - \pi_t)p_t. \tag{1}$$

We aim to learn about the prevalence π_t and $\Delta\pi_{t,j} = \pi_{t+j} - \pi_t$, the trend between periods t and $t+j$. It is clear from equation (1) that without additional information the parameters (π_t, p_t, q_t) are completely unidentified (e.g., Gaba and Winkler (1992)). It is common to assume that $p_t + q_t < 1$, which ensures that the covariance between Y_{it}^* and Y_{it} is positive (Bollinger, 1996; Lewbel, 2007; Chen et al., 2008a,b). This assumption and equation (1) imply that $p_t \leq \mu_t$ and $q_t \leq 1 - \mu_t$, and the misclassification probabilities are now partially identified. The true prevalence π_t, however, remains completely unidentified. As a result, $-1 \leq \Delta\pi_{t,j} \leq 1$ and nothing can be learned about the direction of the trend.

Additional information in the form of restrictions on the misclassification rates can yield non-trivial bounds on the prevalence and the trend. In what follows, we consider a number of cases that lead to partial identification. Throughout the discussion we maintain the assumption that $p_t + q_t < 1$. Also, in the context of reporting prescription opioid misuse, it is highly unlikely that an individual who does not misuse actually reports doing so (Bollinger and David, 1997). Thus, in all but one of the cases we discuss below, we set p_t equal to zero in all time periods.

2.2. Assumptions and Identified Sets

In this section we consider five different assumptions about the misclassification rates. The first and most restrictive assumption is that the rate of false negatives (under-reporting) is constant over time. We subsequently allow this rate to vary over time in different ways and show the impact that this has on the identified sets for the prevalence and the trend. Our final assumption is an extension that allows for the possibility of false positives. Details about the derivation of the parameter bounds can be found in the supplemental appendix.

Case I. The first and most restrictive case we consider is the assumption that the probability of false negatives is constant over time.

Assumption C-I. *(i)* $q_t = q^*$ *for* $t = 1, \ldots, T$, *and (ii)* $p_t = 0$.

Letting $M = \max_t \mu_t$, it follows from Assumption C-I and equation (1) that

$$\mu_t \leq \pi_t \leq \frac{\mu_t}{M}, \quad t = 1, \ldots, T. \tag{2}$$

The trend in prevalence between periods t and $t + j$ is bounded as follows.

$$\begin{aligned} \Delta\mu_{t,j} \leq \Delta\pi_{t,j} \leq \frac{\Delta\mu_{t,j}}{M}, \qquad & \text{if} \quad \Delta\mu_{t,j} \geq 0, \\ \frac{\Delta\mu_{t,j}}{M} \leq \Delta\pi_{t,j} \leq \Delta\mu_{t,j}, \qquad & \text{if} \quad \Delta\mu_{t,j} < 0. \end{aligned} \tag{3}$$

Equations (2) and (3) show that the restrictions on (p_t, q_t) carry substantial identifying information. Since there are only false negatives, the true prevalence in each time period is at least as large as the observed prevalence, and may have an upper bound well below 1. Also, (3) shows that $\Delta\pi_{t,j}$ has the same sign as $\Delta\mu_{t,j}$: if the observed prevalence increases (decreases) between time periods t and $t + j$, then so does the unobserved true prevalence.

Case II. We now assume that the rate of false negatives is non-decreasing over time. This occurs, for example, when Y_{it}^* is an indicator for stigmatized behavior and stigma is increasing over time (Pepper, 2001).

Assumption C-II. *(i) $q_t \geq q_s$ when $t > s$, and (ii) $p_t = 0$.*

Defining $M_t^+ = \max_{s \geq t} \mu_s$, it follows that

$$\mu_t \leq \pi_t \leq \frac{\mu_t}{M_t^+}, \quad t = 1, \ldots, T. \tag{4}$$

While the prevalence is still partially identified, a comparison of (2) and (4) shows that the upper bound on π_t is now larger. Under Assumption C-II the trend in prevalence between periods t and $t + j$ is bounded as follows.

$$\begin{aligned} \Delta\mu_{t,j} \leq \Delta\pi_{t,j} \leq \frac{\mu_{t+j}}{M_{t+j}^+} - \mu_t, \qquad & \text{if} \quad \Delta\mu_{t,j} \geq 0, \\ \frac{\Delta\mu_{t,j}}{M_t^+} \leq \Delta\pi_{t,j} \leq \frac{\mu_{t+j}}{M_{t+j}^+} - \mu_t, \qquad & \text{if} \quad \Delta\mu_{t,j} < 0. \end{aligned} \tag{5}$$

Thus, if the observed prevalence increases between periods t and $t + j$, then so does the true prevalence. This is intuitive: if the observed trend is positive while the rate of false negatives

increases (or at least, does not decrease), then the unobserved true prevalence must be increasing as well. Pepper (2001), using an assumption comparable to C-II, derives a similar result. On the other hand, when $\Delta\mu_{t,j} < 0$, equation (5) shows that the sign of $\Delta\pi_{t,j}$ is not necessarily identified. While the lower bound is negative, the upper bound could be positive. In this case, a decrease in the observed trend results from either a decrease in the true prevalence, or from an increase in false negative reporting that more than offsets a stable or even increasing true prevalence.

Case III. The third case we examine is the mirror image of Case II and assumes that the probability of false negatives is non-increasing over time.

Assumption C-III. *(i)* $q_t \leq q_s$ *when* $t > s$*, and (ii)* $p_t = 0$*.*

Defining $M_t^- = \max_{s \leq t} \mu_s$, Assumption C-III implies that

$$\mu_t \leq \pi_t \leq \frac{\mu_t}{M_t^-}, \quad t = 1, \ldots, T. \tag{6}$$

The true prevalence is again partially identified, but the bounds are farther apart compared to Case I, where q_t is constant over time. The bounds on the trend under Assumption C-III are given below.

$$\begin{aligned} \mu_{t+j} - \frac{\mu_t}{M_t^-} \leq \Delta\pi_{t,j} \leq \frac{\Delta\mu_{t,j}}{M_{t+j}^-}, \qquad & \text{if} \quad \Delta\mu_{t,j} \geq 0, \\ \mu_{t+j} - \frac{\mu_t}{M_t^-} \leq \Delta\pi_{t,j} \leq \Delta\mu_{t,j}, \qquad & \text{if} \quad \Delta\mu_{t,j} < 0. \end{aligned} \tag{7}$$

Equation (7) shows that when the observed prevalence decreases, so does the true prevalence. This occurs because the rate of false negative reporting cannot increase. Hence, a decrease in observed prevalence has to be a result from a decrease in the actual prevalence. On the other hand, the direction of the trend in the true unobserved prevalence is not necessarily identified when $\Delta\mu_{t,j} \geq 0$. An observed increase could result from an increase in the true prevalence but also from a decrease in false negative reporting that more than offsets a stable or even decreasing true prevalence.

Case IV. The prior two cases are restrictive in terms of the structure they impose on q_t. In the fourth case we therefore assume that q_t varies over time but remains within some distance of an unknown "base rate" $\bar{q}$. We will refer to this as the assumption of bounded variation.

Assumption C-IV. *(i) For some* $x \in (0, 1]$ *and* $\bar{q} \in [0, (1 - M)/(1 + x)]$*,* q_t *satisfies* $(1 - x)\bar{q} \leq q_t \leq (1 + x)\bar{q}$*; and (ii)* $p_t = 0$*.*

For the identified set of each q_t to be non-empty, the base rate $\bar{q}$ has to satisfy $(1-x)\bar{q} \leq 1-M$. Assumption C-IV, however, imposes the slightly stronger restriction that $(1+x)\bar{q} \leq 1-M$. This ensures that a maximum (positive or negative) deviation of $100(x)\%$ from the base rate is possible in each time period. We also note that under Assumption C-IV, the case $x = 1$ leads to $0 \leq q_t \leq 2\bar{q}$ for all t. Thus, the assumption that q_t is time-varying but remains below some unknown, fixed upper bound in each time period is subsumed under Assumption C-IV.

From equation (1) and the bounded variation in q_t it follows that

$$\frac{\mu_t}{1-(1-x)\bar{q}} \leq \pi_t \leq \frac{\mu_t}{1-(1+x)\bar{q}}.$$

Minimizing the lower bound and maximizing the upper bound over $\bar{q} \in [0, (1-M)/(1+x)]$ yields the following prevalence bounds:

$$\mu_t \leq \pi_t \leq \frac{\mu_t}{M}. \tag{8}$$

Perhaps surprisingly, these bounds are the same as under Assumption C-I. While allowing q_t to vary over time leads to a larger identified set, limiting the percentage deviation in each period narrows the bounds to the point where these opposing effects exactly offset each other. For the trend, define $a \coloneqq 1-x$ and $b \coloneqq 1+x$ and let $\Delta\pi^L_{t,j}$ and $\Delta\pi^U_{t,j}$ denote the lower and upper bounds, respectively. It is shown in the supplemental appendix that these bounds are given by

$$\Delta\pi^L_{t,j} = \begin{cases} \frac{\mu_{t+j}}{1-(a/b)(1-M)} - \frac{\mu_t}{M} & \text{if } a\mu_{t+j} < b\mu_t, \\ \Delta\mu_{t,j} & \text{if } a\mu_{t+j} \geq b\mu_t,\ M > \frac{(b-a)\sqrt{b\mu_t}}{b\sqrt{a\mu_{t+j}}-a\sqrt{b\mu_t}}, \\ \min\left\{\Delta\mu_{t,j}, \frac{\mu_{t+j}}{1-(a/b)(1-M)} - \frac{\mu_t}{M}\right\} & \text{if } a\mu_{t+j} \geq b\mu_t,\ M \leq \frac{(b-a)\sqrt{b\mu_t}}{b\sqrt{a\mu_{t+j}}-a\sqrt{b\mu_t}}, \end{cases} \tag{9}$$

$$\Delta\pi^U_{t,j} = \begin{cases} \frac{\mu_{t+j}}{M} - \frac{\mu_t}{1-(a/b)(1-M)} & \text{if } b\mu_{t+j} \geq a\mu_t, \\ \Delta\mu_{t,j} & \text{if } b\mu_{t+j} < a\mu_t,\ M > \frac{(b-a)\sqrt{b\mu_{t+j}}}{b\sqrt{a\mu_t}-a\sqrt{b\mu_{t+j}}}, \\ \max\left\{\Delta\mu_{t,j}, \frac{\mu_{t+j}}{M} - \frac{\mu_t}{1-(a/b)(1-M)}\right\} & \text{if } b\mu_{t+j} < a\mu_t,\ M \leq \frac{(b-a)\sqrt{b\mu_{t+j}}}{b\sqrt{a\mu_t}-a\sqrt{b\mu_{t+j}}}. \end{cases} \tag{10}$$

As an example, suppose that the observed prevalence increases between periods t and $t+j$ (so that $\Delta\mu_{t,j} > 0$ and $b\mu_{t+j} > a\mu_t$) but the increase is modest: $a\mu_{t+j} < b\mu_t$. From (9) and (10), it follows

that

$$\frac{\mu_{t+j}}{1-(a/b)(1-M)} - \frac{\mu_t}{M} \leq \Delta\pi_{t,j} \leq \frac{\mu_{t+j}}{M} - \frac{\mu_t}{1-(a/b)(1-M)}.$$

Comparing this with the trend bounds when q_t is constant (cf. (3)), it is easy to show that the lower bound is less than $\Delta\mu_{t,j}$, whereas the upper bound exceeds $\Delta\mu_{t,j}/M$. Thus, the identified set is again larger than under Assumption C-I. We also note that a constant rate of false negatives can be obtained as a limit of the bounded variation assumption when $x \downarrow 0$. In this case a and b both converge to 1 and the lower and upper bounds in (9) and (10) converge to the bounds in (3).

Case V. The final assumption we discuss is an extension of Case IV and allows for a non-zero but constant rate of false positives p.

Assumption C-V. *(i) For some $x \in (0,1)$ and $\bar{q} \in [0,(1-M)/(1+x)]$, q_t satisfies $(1-x)\bar{q} \leq q_t \leq (1+x)\bar{q}$; and (ii) $p_t = p$.*

Since $p \leq \mu_t$ for all t, the upper bound for p is $m \coloneqq \min_s \mu_s$. For a given value of p, we have the following prevalence bounds:

$$\frac{\mu_t - p}{1-p-a\bar{q}} \leq \pi_t \leq \frac{\mu_t - p}{1-p-b\bar{q}}.$$

The lower bound is minimal when $\bar{q} = 0$ and the upper bound is maximal when $\bar{q} = (1-M)/b$, so that

$$\frac{\mu_t - p}{1-p} \leq \pi_t \leq \frac{\mu_t - p}{M-p}.$$

The bounds on p shown above are decreasing in p, so that

$$\frac{\mu_t - m}{1-m} \leq \pi_t \leq \frac{\mu_t}{M} \tag{11}$$

Comparing (8) and (11), we see that allowing false positives reduces the lower bound and results in a larger identified set. The true prevalence may be below the observed prevalence due to the possibility of false positives. Regarding the trend, we use (1) and observe that for any given value of p, the difference $\Delta\pi_{t,j}$ is maximized when $(q_t, q_{t+j}) = (a\bar{q}, b\bar{q})$ and minimized for $(q_t, q_{t+j}) = (b\bar{q}, a\bar{q})$. Therefore,

$$\frac{\mu_{t+j} - p}{1-p-a\bar{q}} - \frac{\mu_t - p}{1-p-b\bar{q}} \leq \Delta\pi_{t,j} \leq \frac{\mu_{t+j} - p}{1-p-b\bar{q}} - \frac{\mu_t - p}{1-p-a\bar{q}} \tag{12}$$

The lower bound on the trend is obtained by minimizing the left-hand side of (12) subject to $0 \leq p \leq m$ and $0 \leq \bar{q} \leq (1-M)/b$. It is shown in the supplemental appendix that if $\Delta\mu_{t,j} < 0$ and Assumption C-V holds, the lower bound is attained at $\bar{q} = (1-M)/b$ and given by

$$\Delta\pi^L_{t,j} = \begin{cases} \frac{\mu_{t+j}-m}{c-m} - \frac{\mu_t - m}{M-m} & \text{if } \mu_{t+j} \leq c - \left(\frac{c-m}{M-m}\right)^2 M + \left(\frac{c-m}{M-m}\right)^2 \mu_t, \\ \frac{\mu_{t+j}-p^*_L}{c-p^*_L} - \frac{\mu_t - p^*_L}{M-p^*_L} & \text{if } c - \left(\frac{c-m}{M-m}\right)^2 M + \left(\frac{c-m}{M-m}\right)^2 \mu_t < \mu_{t+j} < c - M\left(\frac{c}{M}\right)^2 + \left(\frac{c}{M}\right)^2 \mu_t, \\ \frac{\mu_{t+j}}{c} - \frac{\mu_t}{M} & \text{if } \mu_{t+j} \geq c - M\left(\frac{c}{M}\right)^2 + \left(\frac{c}{M}\right)^2 \mu_t, \end{cases} \tag{13}$$

where

$$p^*_L = \frac{c\sqrt{M-\mu_t} - M\sqrt{c-\mu_{t+j}}}{\sqrt{M-\mu_t} - \sqrt{c-\mu_{t+j}}}.$$

When $\Delta\mu_{t,j} \geq 0$, there is no convenient way to characterize $\Delta\pi^L_{t,j}$, because it depends on the relative magnitudes of $(a, b, \mu_t, \mu_{t+j}, m, M)$. Solutions to minimizing the left-hand side of (12), subject to the boundary restrictions, can be found by inspecting solutions to the Kuhn-Tucker first-order conditions.

The upper bound on the trend is found by maximizing the right-hand side of (12) subject to $0 \leq p \leq m$ and $0 \leq \bar{q} \leq (1-M)/b$. If Assumption C-V holds and $\Delta\mu_{t,j} > 0$, the upper bound is attained at $\bar{q} = (1-M)/b$ and given by

$$\Delta\pi^U_{t,j} = \begin{cases} \frac{\mu_{t+j}}{c} - \frac{\mu_t}{M} & \text{if } \mu_{t+j} \leq M - c\left(\frac{M}{c}\right)^2 + \left(\frac{M}{c}\right)^2 \mu_t, \\ \frac{\mu_{t+j}-p^*_L}{c-p^*_L} - \frac{\mu_t - p^*_L}{M-p^*_L} & \text{if } M - c\left(\frac{M}{c}\right)^2 + \left(\frac{M}{c}\right)^2 \mu_t < \mu_{t+j} < M - c\left(\frac{M-m}{c-m}\right)^2 + \left(\frac{M-m}{c-m}\right)^2 \mu_t, \\ \frac{\mu_{t+j}-m}{c-m} - \frac{\mu_t - m}{M-m} & \text{if } \mu_{t+j} \geq M - c\left(\frac{M-m}{c-m}\right)^2 + \left(\frac{M-m}{c-m}\right)^2 \mu_t, \end{cases} \tag{14}$$

When $\Delta\mu_{t,j} \leq 0$ instead, there is again no convenient expression for $\Delta\pi^U_{t,j}$. The upper bound can be found by inspecting solutions to the Kuhn-Tucker first-order conditions for maximizing the right-hand side of (12), subject to the boundary restrictions.

In summary, we have presented the implications of different assumptions about p_t and q_t for the identified sets of π_t and $\Delta\pi_{t,j}$. The focus on conditional error probabilities is common in much of the misclassification literature. In contrast, Pepper (2001) imposes restrictions on the joint distribution of (Y^*_{it}, Y_{it}). In our notation $P(Y^*_{it} = 1, Y_{it} = 0) = \pi_t q_t$ and $P(Y^*_{it} = 0, Y_{it} = 1) = (1-\pi_t)q_t$. Pepper

(2001) assumes that false negatives are at least as likely as false positives, so that $P(Y_{it}^* = 1, Y_{it} = 0) \geq P(Y_{it}^* = 0, Y_{it} = 1)$. In addition, the *total* fraction of misclassified observations is assumed to lie below some known upper bound:

$$P(Y_{it}^* = 1, Y_{it} = 0) + P(Y_{it}^* = 0, Y_{it} = 1) \leq P.$$

In this case the true prevalence satisfies the bounds $\mu_t \leq \pi_t \leq \min\{\mu_t + P, 1\}$. Our results for the prevalence provide a useful extension of Pepper's (2001) bounds for two reasons. First, restrictions on the joint distribution of (Y_{it}^*, Y_{it}) are restrictions on the triple (π_t, p_t, q_t), whereas we only restrict (p_t, q_t) and investigate the implications for π_t. Second, as noted by Pepper (2001), the upper bound P on total false reports must either be known or a value must be assumed by the researcher. Setting a reasonable value for P may be difficult in practice. The prevalence bounds we present here do not depend on any unknown constants.

3. BAYESIAN INFERENCE

3.1. Nonidentification and the Posterior

We now consider Bayesian inference about the prevalence under Assumptions C-I through C-V. A Bayesian model that is consistent with these assumptions incorporates the parameter bounds from the previous section into the prior distribution. We initially assume that a simple random sample is available in each time period. We postpone a discussion of more complex survey designs and the use of sampling weights until section 3.3.

Let μ, π and q be T-dimensional parameter vectors with t-th elements μ_t, π_t and q_t, respectively, and let p be a scalar (recall that under Assumptions C-I through C-V), p is constant over time). We use $\mathbf{Y} = \{Y_{it}; i = 1, \ldots, n_t, t = 1, \ldots, T\}$ to denote the full set of observations across all individuals and time periods. Let $n_{t1} = \sum_i Y_{it}$ and $n_{t0} = \sum_i (1 - Y_{it})$ be the observed number of ones and zeroes in time period t, respectively, and define $n_t = n_{t1} + n_{t0}$. If the samples from different periods are independent and each individual only appears in a single period, the likelihood for the full sample can be written as

$$f(\mathbf{Y}|\mu, \pi, q, p) = \prod_{t=1}^{T} \mu_t^{n_{t1}} (1 - \mu_t)^{n_{t0}}. \tag{15}$$

The likelihood is a function of μ alone and therefore does not separately identify π, q and p.

Let $f(\mu, \pi)$ be a prior distribution. Since π is not identified, the joint posterior of (μ, π) can be decomposed as (Kadane, 1974; Poirier, 1998)

$$\begin{aligned} f(\mu, \pi|\mathbf{Y}) &\propto f(\mathbf{Y}|\mu) \cdot f(\mu) \cdot f(\pi|\mu) \\ &\propto f(\mu|\mathbf{Y}) \cdot f(\pi|\mu). \end{aligned}$$

A similar expression holds for the joint posterior of μ, p and q. The marginal posterior of π is obtained by integrating out μ:

$$f(\pi|\mathbf{Y}) \propto \int f(\mu|\mathbf{Y}) f(\pi|\mu) d\mu.$$

Learning about π occurs indirectly through the conditional prior. As more data become available, the posterior $f(\mu|\mathbf{Y})$ becomes more concentrated around some value, say $\tilde{\mu}$. The marginal posterior of π will then get close to the conditional prior $f(\pi|\tilde{\mu})$ and uncertainty about π remains, even in large samples. When information about the rates of misreporting is available, a researcher could use it to specify a prior for (μ, p, q) instead of for (μ, π).[3] A similar argument can be used to show that in large samples the posterior of π again gets close to the conditional prior $f(\pi|\tilde{\mu})$.

3.2. Prior Distributions

The prior distributions we propose are based on Assumptions C-I through C-V and specified in terms of the misclassification rates. Since μ is identified, the prior $f(\mu)$ will have a negligible influence on the posterior in large samples. However, as noted earlier, the priors $f(p, q|\mu)$ or $f(\pi|\mu)$ remain influential in large samples and their specification needs to be considered carefully.

Case I. Under Assumption C-I we know that $q^* \leq 1 - M$. Without specific knowledge about misreporting, a researcher might use a uniform prior on the interval $[0, 1 - M]$, conditional on μ. The conditional prior of π_t is then $f(\pi_t|\mu) = \mu_t/[(1 - M)\pi_t^2]$ for $\mu_t \leq \pi_t \leq \mu_t/M$. This density is decreasing in π_t, so that it is relatively more likely that the true prevalence is close to the observed prevalence (i.e., the lower bound of the identified set). If instead small values of q^* are believed to

[3] For example, Meyer et al. (2015) and Meyer et al. (2018) provide estimates of the amount of misreporting in SNAP participation.

be more likely than large values, we can use a power-type prior $f(q^*|\mu) = C(q^*)^{-\alpha}$, where $0 < \alpha < 1$ and C is a normalizing constant. The induced prior for π_t is then $f(\pi_t|\mu) = C\mu_t^{1+\alpha}/(\pi_t^2(\pi_t - \mu_t)^\alpha)$, which also places a relatively high probability on values of π_t near μ_t. Finally, suppose a researcher wishes to use a (conditional) prior on π_t directly. One possibility is the uniform distribution on the interval $[\mu_t, \mu_t/M]$. The induced prior for q^* is then $f(q^*|\mu) = M/[(1-M)(1-q^*)^2]$, which puts a relatively high probability on values of q^* near $1 - M$. Thus, a uniform prior for the true prevalence can be justified if we believe that the rate of false negative reporting is likely to be high.

Case II. When q_t is assumed to be non-decreasing (Assumption C-II), we can construct a prior of the form

$$\begin{aligned} f(q|\mu) &= f(q_1|\mu) \prod_{t=2}^{T} f(q_t|q_{t-1}, \ldots, q_1, \mu) \\ &= f(q_1|\mu) \prod_{t=2}^{T} f(q_t|q_{t-1}, \mu). \end{aligned}$$

The conditional priors $f(q_t|q_{t-1}, \ldots, q_1, \mu)$ for $t \geq 3$ are chosen to be independent of $q_{t-2}, \ldots, q_1$ because q_t satisfies the restriction $q_t \geq q_{t-1}$. The probability of a false negative in the first period satisfies $q_1 \leq 1 - M$, and we can specify a prior with support on this range, as in Case I. Similarly, for $t \geq 2$, we have the inequalities $q_{t-1} \leq q_t \leq 1 - M_t^+$, and we choose conditional priors $f(q_t|q_{t-1}, \mu)$ with support on this interval. If lower misreporting rates are considered more likely, we choose a power-type distribution for each q_t that puts more probability near the lower bound of the support (as in Case I). An alternative choice for $f(q|\mu)$ is to use a series of uniform distributions on the intervals discussed above. The choice of continuous distributions for $f(q_t|q_{t-1}, \mu)$, however, implies that q_t is strictly increasing with prior probability 1. This can result in a large probability of unreasonably high values of q_t in later time periods. To avoid this in the empirical application, we therefore use a discrete-continuous mixture prior that assigns a positive probability to the false negative rate staying the same between $t - 1$ and t. Specifically, if $\lambda \in (0, 1)$ is the mixture proportion, we use (for $t = 2, \ldots, T$)

$$q_t \begin{cases} = q_{t-1} & \text{with probability } \lambda, \\ \sim f(q_t|q_{t-1}, \mu) & \text{with probability } 1 - \lambda, \end{cases}$$

where, as discussed above, $f(q_t|q_{t-1},\mu)$ is a uniform or power-type distribution supported on the interval $[q_{t-1}, 1-M_t^+]$.

Case III. If the rate of false-negative reporting is assumed to be non-increasing, we can construct a prior in a way that resembles Case II:

$$\begin{aligned} f(q|\mu) &= f(q_T|\mu)\prod_{t=1}^{T-1} f(q_{T-t}|q_{T-t+1},\ldots,q_T,\mu) \\ &= f(q_T|\mu)\prod_{t=1}^{T-1} f(q_{T-t}|q_{T-t+1},\mu). \end{aligned}$$

For q_T, we choose a prior (conditional on μ) that is supported on the interval $[0, 1-M]$. For $t<T$, the misreporting rate satisfies $q_{t+1} \leq q_t \leq 1-M_t^-$, and we choose a distribution $f(q_t|q_{t+1},\mu)$ supported on that interval. If we want to ensure that there is a non-zero probability that q_t stays the same between successive periods, we can again use a mixture distribution:

$$q_t \begin{cases} = q_{t+1} & \text{with probability } \lambda, \\ \sim f(q_t|q_{t+1},\mu) & \text{with probability } 1-\lambda, \end{cases}$$

where $f(q_t|q_{t+1},\mu)$ is a continuous distribution supported on the interval $[q_{t+1}, 1-M_t^-]$.

Cases IV-V. Under the assumption of bounded variation, the probability of a false negative in period t can be written as $q_t = v_t\bar{q}$, where $1-x \leq v_t \leq 1+x$. We assume that x is chosen by the researcher (e.g., $x=0.10$ or $x=0.25$). A prior for q can be obtained by combining a distribution for $\bar{q}$ with a distribution for $(v_1,\ldots,v_T)$. Since $\bar{q} \leq (1-M)/(1+x)$, possible (conditional) priors for $\bar{q}$ are the uniform or power-type distribution on $[0,(1-M)/(1+x)]$. Candidate priors for v_t include the uniform distribution on the interval $[1-x,1+x]$ and the normal distribution with mean 1, truncated to the interval $[1-x,1+x]$. Finally, under Assumption C-IV we simply set $p=0$, whereas under Assumption C-V we can use a uniform or power-type prior for p supported on the interval $[0,m]$.

3.3. Survey Design and Sampling from the Posterior

Since we are interested in inference about population prevalence and trends, it is necessary to consider the sampling design. In our empirical analysis, we use data $\mathbf{Y}$ from the NSDUH, which does not constitute a random sample from the population and invalidates the likelihood function in (15). Suppose, however, that a set of individual-level sampling weights w_{it} is available, where $N_t = \sum_{i=1}^{n_t} w_{it}$ is the size of the population at time t. Thus, observation Y_{it} is thought to represent w_{it} individuals in the population. We assume that the size of the population and the weights are known (as is typically done; incorporating uncertainty about the weights is beyond the scope of this paper) and follow an approach proposed by Gunawan et al. (2017) to conduct Bayesian inference about (μ, π, q). Their approach is based on data augmentation (Tanner and Wong, 1987) and consists of two steps. First, use the sampling weights to generate pseudo-random samples from the population. Second, use these samples to conduct inference about the parameters in the usual Bayesian way.

To describe the steps involved in more detail, let $Y_t = (Y_{1t}, \ldots, Y_{n_t,t})$ be the observed sample at time t, so that $\mathbf{Y} = (Y_1, \ldots, Y_T)$. Similarly, let $\tilde{Y}_t = (\tilde{Y}_{1t}, \ldots, \tilde{Y}_{n_t,t})$ be a random sample from the population at time t, and let $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_T)$. The vector of sampling weights at time t is $w_t = (w_{1t}, \ldots, w_{n_t,t})$ and we define $\mathbf{w} = (w_1, \ldots, w_T)$. Conditional on $(\mathbf{Y}, \mathbf{w})$, the variable $\tilde{Y}_{it}$ has a Bernoulli distribution with parameter $\tilde{p}_t$, where

$$\tilde{p}_t = P(\tilde{Y}_{it} = 1 | Y_t, w_t) = \frac{\sum_{j=1}^{n_t} w_{jt} Y_{jt}}{\sum_{j=1}^{n_t} w_{jt}}. \tag{16}$$

The samples $\tilde{\mathbf{Y}}$ are not observed. With data augmentation they are treated as an additional set of unknown parameters. The posterior distribution of μ and $\tilde{\mathbf{Y}}$ can be decomposed as

$$f(\mu, \tilde{\mathbf{Y}} | \mathbf{Y}, \mathbf{w}) = f(\mu | \tilde{\mathbf{Y}}, \mathbf{Y}, \mathbf{w}) \cdot f(\tilde{\mathbf{Y}} | \mathbf{Y}, \mathbf{w}).$$

The second term on the right-hand side is the product of the Bernoulli distributions in (16). Also, the conditional posterior of μ depends only on the random samples from the population, so that

$$f(\mu | \tilde{\mathbf{Y}}, \mathbf{Y}, \mathbf{w}) = f(\mu | \tilde{\mathbf{Y}})$$

$$\propto f(\mu) f(\tilde{\mathbf{Y}}|\mu)$$
$$\propto f(\mu) \prod_{t=1}^{T} \mu_t^{\tilde{n}_{t1}} (1-\mu_t)^{\tilde{n}_{t0}}, \tag{17}$$

where $\tilde{n}_{t1} = \sum_{i=1}^{n_t} \tilde{Y}_{it}$ and $\tilde{n}_{t0} = \sum_{i=1}^{n_t} (1 - \tilde{Y}_{it})$ for $t = 1, \ldots, T$. A random draw from the joint posterior $f(\mu, \tilde{\mathbf{Y}}|\mathbf{Y}, \mathbf{w})$ can now be generated by first drawing $\tilde{\mathbf{Y}}$ from (16) and then drawing μ from (17). Note that because the conditional posterior in (17) only depends on $\tilde{n}_{t1}$ and $\tilde{n}_{t0}$, it is not necessary to sample each $\tilde{Y}_{it}$ individually. Instead, we can sample $\tilde{n}_{t1}$ from the binomial distribution with parameters n_t and $\tilde{p}_t$. Assuming that a conditional prior $f(q|\mu)$ or $f(p, q|\mu)$ has been specified, the steps to generate a sample from the posterior can now be summarized as follows.

SAMPLING FROM THE POSTERIOR:

(1) For $t = 1, \ldots, T$, sample values $\tilde{n}_{t1}$ from a binomial distribution with parameters n_t and $\tilde{p}_t$, and calculate $\tilde{n}_{t0} = n_t - \tilde{n}_{t1}$;

(2) Given the sampled value $(\tilde{n}_{t1}, \tilde{n}_{t0})$, sample μ from the posterior distribution in (17);

(3) Given the sampled value μ:

   i. (Cases I-IV) if $p = 0$, sample q from the conditional prior $f(q|\mu)$ and calculate $\pi_t = \mu_t/(1 - q_t)$ for $t = 1, \ldots, T$;

   ii. (Case V) if $p \neq 0$, sample p and q from $f(p, q|\mu)$ and calculate $\pi_t = (\mu_t - p)/(1 - p - q_t)$ for $t = 1, \ldots, T$;

(4) Return to (1) and repeat.

In Section 4 we use a uniform prior for μ, so that step (2) involves generating a random draw from a beta distribution with parameters $(\tilde{n}_{t1} + 1, \tilde{n}_{t0} + 1)$. Finally, as referred to earlier, if the Bayesian model specifies a conditional prior for π instead of q, step (3) is modified by sampling a value of $f(\pi|\mu)$ and a value from $f(p, q|\pi, \mu)$.

4. ESTIMATING PREVALENCE AND TREND OF OPIOID MISUSE

4.1. Observed Prevalence

For our empirical analysis we use publicly available data from the 2002-2014 waves of the National Survey on Drug Use and Health (NSDUH).[4] The NSDUH provides a nationally representative sample of the non-institutionalized U.S. population aged 12 years old or older, and collects detailed information about the use and misuse of various substances, including alcohol, tobacco, marijuana, prescription drugs and illegal drugs. Data from the NSDUH is therefore a primary source of information for looking at trends in the use and misuse of prescription opioids.

Our variable of interest is an indicator for misuse of prescription pain relievers during the past year. The NSDUH imputed this indicator based on an individual's response to the question "How long has it been since you last used any prescription pain reliever that was not prescribed for you or that you took only for the experience or feeling it caused?" We use the indicator and individual-level sampling weights to estimate the population prevalence of past-year misuse, as well as the one-year changes in prevalence.

The observed prevalence μ_t, based on the self-reported misuse indicator, is shown in Figure 1. The left panel represents the population of individuals 18 and older. Between 2002 and 2007, the estimated prevalence rose from 4.4% to 4.9%, an increase of more than 10%. Between 2007 and 2012, the prevalence fluctuated before starting a seemingly downward trend in 2013. The year 2011 seems to be an anomaly, with the prevalence temporarily dropping down to 4.1%. The reason for this is unclear but we suspect it may result from some extreme values in the sampling weights. The 95% confidence intervals for the observed prevalence in each year largely overlap, making it difficult to draw any definite conclusions about a trend.

The right panel of Figure 1 shows the observed prevalence for white men, ages 26 to 49 years old. This population is of interest because recent evidence suggests that middle-aged white men are at a relatively high risk of prescription drug abuse (Case and Deaton, 2015). As is apparent from the figure, the observed prevalence among middle-aged white men was substantially higher than in the overall population 18 and older. It rose from about 5.5% in 2002 to 7.3% in 2010, an

[4]The data is available for download from https://www.datafiles.samhsa.gov. Public use files for 2015-2018 have also been released, but we are not using these for our analysis due to a major survey redesign in 2015 that impacted the prescription drug module of the questionnaire.

increase of more than 30%. The prevalence may have started to decline in 2013 but, given the width of the confidence intervals, it is hard to discern any clear trend. In Section 4.3, we apply our Bayesian approach to the subsample of middle-aged white men. Results for several other subgroups are given in the supplemental appendix.

4.2. Prior Specification

We use the prior distributions proposed and discussed in more detail in Section 3.2. For all power-type distributions, we set $\alpha = 2$. When q_t is assumed to be constant over time, we use a uniform and power distributions supported on the identified set. When q_t is thought to be non-decreasing (Assumption C-II) or non-increasing (Assumption C-III), we use a mixture prior: with probability $\lambda = 0.9$, q_t remains the same between adjacent periods; with probability 0.10, q_t follows a uniform or power distribution on the identified set, conditional on either q_{t-1} or q_{t+1}. Under Assumption C-IV, q_t deviates at most $100(x)\%$ from a base rate $\bar{q}$, and we set $x = 0.25$. After sampling $\bar{q}$ from a power distribution, we generate $q_t = v_t\bar{q}$ by drawing v_t from a distribution truncated to the interval $[0.75, 1.25]$. Specifically, we use a uniform distribution and two normal distributions with mean 1 and standard deviations 0.25 and 0.0625. These reflect increasingly strong beliefs that v_t is close to 1 or, equivalently, that q_t is close to $\bar{q}$. Finally, under Assumption C-V we augment the prior with a power distribution for p, supported on the interval $[0, m]$.

The following section shows estimates of the classical bounds and the Bayesian 95% highest posterior density (HPD) intervals. Within these intervals, circles indicate the posterior mean. While other summary statistics can be calculated from the posterior, we focus on HPD intervals for the one-year trend, the posterior probability that this trend is positive, and a comparison of the average prevalence between two sub-periods. This narrower focus allows us to compare results across different prior distributions more easily. Additional posterior graphs and summary statistics are collected in the supplemental appendix.

4.3. Posterior Summaries for the True Prevalence

The posterior results presented here are based on $100,000$ simulated draws from the posterior distribution.[5] As noted in Section 3.3, generating a random draw from the posterior involves generating a draw from the posterior $f(\mu|\mathbf{Y})$ followed by generating a random draw from the conditional prior $f(q|\mu)$ or $f(p,q|\mu)$. When q_t is assumed to be constant, Figure 2 shows that the HPD intervals for the prevalence are much narrower than estimates of the identified set, especially when a power prior is used for q^*. This is no longer the case for the one-year change in prevalence. The classical bounds often lie within the limits of the HPD interval, though this may partially occur because the estimates of the identified set do not account for uncertainty in the bounds. Table 1 shows the posterior probabilities that a given one-year change in prevalence is positive. For example, the probability that the true prevalence of misuse increased between 2002 and 2003 is about 78% under both priors. The posterior probabilities and means seem robust to the choice of the two priors for q^* (uniform and power) that we consider here. There is strong evidence that the prevalence increased—relative to the previous year—in 2006 and 2007, with posterior probabilities of 95% and 82%, respectively. Until 2012 the prevalence shows no clear trend, but in 2013 and 2014 the probability of a positive one-year change drop below 14%. This suggests that the misuse prevalence started a downward trend in those years.

Estimates of the identified sets and 95% HPD intervals for the trend are shown in Figure 3, when q_t is assumed to be either non-decreasing (Case II) or non-increasing (Case III). As expected, these sets and intervals are much wider compared to the case where q_t is constant. The striking feature of this figure is that the Bayesian HPD intervals are now much narrower than the (estimated) classical bounds. Comparing the left- and right-hand sides of Figure 3 we also see that the location of the bounds strongly depends on whether false negatives are assumed to be (weakly) increasing or decreasing.

As shown in Table 2, the posteriors under Case II and Case III again provide evidence that the true prevalence increased in the periods 2005-2006 and 2006-2007. For example, assuming that q_t

[5]The computational time is modest. For example, generating $100,000$ draws from the posterior under the assumption that q_t is constant takes around 180-190 seconds. The largest fraction of this time is used to simulate from the (beta) posterior of μ_t (see Section 3.3). It is important to note that the marginal posterior is the same across all joint posteriors of (μ_t, π_t), so that the additional computing time to simulate samples from other posteriors is much smaller. For example, generating samples of size $100,000$ each from 6 different posteriors takes around 220 seconds. More than 85% of that time is used to generate the posterior draws of μ_t.

year	uniform prior	power prior
2003	0.7757	0.7757
2004	0.7127	0.7127
2005	0.2185	0.2185
2006	0.9527	0.9527
2007	0.8151	0.8151
2008	0.1968	0.1968
2009	0.4197	0.4197
2010	0.6464	0.6464
2011	0.4791	0.4791
2012	0.5621	0.5621
2013	0.1361	0.1361
2014	0.1344	0.1344

Table 1: posterior probability of an increase in prevalence relative to the previous year (q_t constant)

was non-decreasing, the posterior probability of an increase during the period 2005-2006 was 95.7% with either a uniform or a power mixture component in the prior (recall that in Case II and III we use priors for q_t that are discrete-continuous mixtures). We also note that if q_t is assumed to be non-decreasing, the posterior probabilities of a positive one-year change are uniformly larger, compared to when q_t is non-increasing. This was to be expected. For example, from Figure 1 we see a large increase in observed prevalence from 2005 to 2006. If false negatives stayed the same or increased in this period, as assumed in Case II, then the increase in true prevalence was even higher. On the other hand, if false negatives (weakly) decreased, then part of the observed increase in prevalence may be due to less misreporting, and the evidence for an increase in the true prevalence is weaker (i.e., the probability of a positive trend is smaller).

Next, we consider the case where q_t is assumed to deviate no more than 25% from an unknown base rate $\bar{q}$, with a constant rate of false positives $p = 0$ (Case IV) or $p > 0$ (Case V). Figure 4 shows the identified sets and HPD intervals for the trend in true prevalence. The identified sets for the trend cover a wide range of positive and negative values and are uninformative about the direction of the change in any given year. The 95% HPD intervals are again much narrower.

The results in Table 3 show strong evidence for an increase in misuse from 2005 to 2006. Moreover, the posterior probability of an increase in the prevalence during that period becomes larger as the prior distribution of v_t, the factor measuring the deviation from the base rate, becomes more concentrated around 1. For example, assuming that $p > 0$ as in Case V, the posterior

year	q_t non-decreasing (Case II)		q_t non-increasing (Case III)	
	uniform	power	uniform	power
2003	0.7961	0.7954	0.7139	0.7150
2004	0.7385	0.7378	0.6557	0.6565
2005	0.2810	0.2791	0.1990	0.1992
2006	0.9571	0.9570	0.8871	0.8881
2007	0.8313	0.8311	0.7526	0.7537
2008	0.2604	0.2593	0.1793	0.1793
2009	0.4666	0.4665	0.3839	0.3844
2010	0.6760	0.6760	0.5932	0.5942
2011	0.5216	0.5215	0.4368	0.4377
2012	0.5988	0.5985	0.5138	0.5154
2013	0.2002	0.1999	0.1235	0.1240
2014	0.2020	0.2009	0.1222	0.1225

Table 2: posterior probability of an increase in prevalence relative to the previous year; priors are mixtures with a uniform or power component

probability of an increase in true prevalence between 2005 and 2006 is 90.4% under a uniform prior for v_t on the interval $[0.75, 1.25]$. If that prior changes to a truncated $N(0, (0.0625)^2)$ distribution, the probability increases to 93.7%. Table 3 also shows evidence that the prevalence decreased after 2012. For example, allowing for false positives and depending on the prior, the probability of a decrease in prevalence between 2012 and 2013 ranged from 82 to 85 percent.

So far, we have focused on the quantity $\Delta\pi_{t,1} = \pi_{t+1} - \pi_t$ and the posterior probability that it is positive. There are, of course, many other parameters that could be of interest. For example, inspection of Figure 1 suggests that for middle-aged white men, the prevalence of misuse may have been higher in the period 2006-2009 compared to 2002-2005, but was more or less stable in the period 2010-2012. To assess the posterior evidence for this, we use the sample from the posterior of π_t and first calculate the difference between the average true prevalence during 2002-2005 ($\bar{\pi}_0$) and the average during 2006-2009 ($\bar{\pi}_1$). Figure 5 shows kernel density estimates of the posterior of $\bar{\pi}_1 - \bar{\pi}_0$, assuming bounded variation in q_t and $p = 0$ (Case IV) or $p > 0$ (Case V). For Case IV in the left graph, the prior of the base false negative rate $\bar{q}$ is a power distribution. For Case V in the right graph, the priors on both the base false negative rate and the false positive rate are power distributions. Summary statistics of the posteriors are given in Table 4.

In both cases there is strong evidence that the average prevalence of prescription opioid misuse

year	(Case IV: $p = 0$)			(Case V: $p > 0$)		
	uniform	TN_1	TN_2	uniform	TN_1	TN_2
2003	0.7256	0.7282	0.7560	0.7394	0.7414	0.7633
2004	0.6681	0.6713	0.6954	0.6793	0.6826	0.7011
2005	0.2742	0.2691	0.2426	0.2611	0.2571	0.2352
2006	0.8887	0.8933	0.9306	0.9043	0.9091	0.9365
2007	0.7519	0.7557	0.7883	0.7646	0.7669	0.7943
2008	0.2578	0.2537	0.2224	0.2455	0.2412	0.2146
2009	0.4356	0.4352	0.4273	0.4288	0.4272	0.4205
2010	0.6169	0.6192	0.6345	0.6204	0.6232	0.6373
2011	0.4851	0.4838	0.4813	0.4861	0.4849	0.4839
2012	0.5464	0.5481	0.5544	0.5494	0.5489	0.5542
2013	0.2002	0.1963	0.1613	0.1814	0.1808	0.1507
2014	0.2019	0.1972	0.1599	0.1864	0.1825	0.1553

Table 3: posterior probability of an increase in prevalence relative to the previous year. q_t is assumed not to deviate more than 25% from the base rate. The priors of the relative deviation v_t are uniform, $N(1, (0.25)^2)$ and $N(1, (0.0625)^2)$, all truncated to the interval $[0.75, 1.25]$. The latter two distributions are labeled TN_1 and TN_2.

was higher in 2006-2009 than it was in 2002-2005. The posterior distributions are centered on a mean difference of about 2.1 to 2.2 percentage points (this corresponds increases in the average prevalence of roughly 23% under Assumption C-IV and 44% under Assumption C-V). The 95% HPD intervals cover mostly positive values and the posterior probability of an increase in the average prevalence exceeds 97% for both cases and all priors considered here. Results presented in the supplemental appendix show that there is no evidence of a substantial change in prevalence between 2006-2009 and 2010-2012. Moreover, the results reported here appear to be unique to white men: individuals in the same age range who are *not* white men displayed a pattern of misuse that was more constant over time.

Case	Prior	mean	std. dev.	2.5%	50%	97.5%	95% HPD	$P(+)$
	uniform	0.0211	0.0235	-0.0018	0.0166	0.0805	[-0.0062,0.0732]	0.9721
IV	TN_1	0.0209	0.0219	-0.0003	0.0167	0.0772	[-0.0050,0.0696]	0.9746
	TN_2	0.0205	0.0130	0.0066	0.0173	0.0573	[0.0038,0.0486]	0.9962
	uniform	0.0217	0.0215	0.0033	0.0172	0.0774	[-0.0027,0.0661]	0.9817
V	TN_1	0.0218	0.0208	0.0042	0.0173	0.0752	[-0.0004,0.0651]	0.9839
	TN_2	0.0211	0.0131	0.0070	0.0177	0.0580	[0.0038,0.0490]	0.9973

Table 4: posterior summary of the difference $\bar{\pi}_1 - \bar{\pi}_0$ in average prevalence between the periods 2006-2009 and 2002-2005. $P(+)$ is the probability of an increase in average prevalence.

5. CONCLUSION

Misclassification error is a frequent concern in self-reported survey data. Examples include reports of participation in social programs and reports of certain types of behavior (e.g., substance misuse). In this paper we analyze the implications of misclassification in the context of a repeated cross section. We derive the identified sets for the means of a binary variable (the prevalence) as well as changes in the mean over time (the trend). These sets are sensitive to what is assumed about the probability of a misclassification error. We consider 5 different cases. In the first four cases, motivated by the context of prescription opioid misuse, we assume that the probability of a false positive (i.e, individuals incorrectly reporting misuse) is zero. In the fifth and final case, we allow for the possibility of false positives.

A second contribution of this paper is that we show how to conduct Bayesian inference about the true prevalence and trends when these parameters are only partially identified. We apply this approach to an analysis of prescription opioid misuse, based on data from the NSDUH. The observed prevalence for white, middle-aged men is relatively high, which is why we mostly restrict our analysis to this population. We find that the estimated identified sets (intervals) are very wide and have limited usefulness. The Bayesian HPD intervals, on the other hand, are typically much narrower and provide information about the plausible values of the prevalence and the trend. Under a variety of assumptions and prior distributions, we find evidence that the prevalence of prescription opioid misuse increased several times between 2002 and 2014. An analysis of select subgroups shows that these patterns are mostly unique to middle-aged white men.

Our paper highlights the strong impact of prior assumptions on identified sets and HPD intervals. This is necessarily the case in models with unidentified or partially identified parameters. We have experimented with a range of prior assumptions and prior distributions while remaining silent about which of these are most appropriate or reasonable in the context of the opioid epidemic. A related issue is that there may be substantial heterogeneity in misreporting behavior across the population, and researchers may wish to use different priors for distinct subgroups. Even with very little knowledge about the extent of misreporting, other outcomes such as emergency room visits or drug-related fatalities could provide direct prior information about the true prevalence of misuse. Since this information may have a significant impact on posterior inference, it is crucial to motivate

and calibrate priors distributions carefully.

Finally, we have illustrated our approach by analyzing the prevalence of past-year misuse of prescription opioids. For policy makers seeking to address this critical public health problem, an analysis of additional, potentially misreported outcomes such as the incidence of misuse, is likely to be of interest. Also, a natural next step is to determine the impact of misuse, at either the individual, county or state level, on a range of health and socioeconomic outcomes. In the supplemental appendix to this paper, we briefly discuss how our approach might be adapted to inform estimation in these contexts. We aim to pursue these issues in more detail in future work.

Although bounding estimators and identified sets have long been the subject of academic pursuit, they have not been of much policy relevance, largely because the estimated bounds are often so far apart. Our approach provides an avenue by which these bounds can be narrowed and thus may become more informative for policy. We have illustrated their potential importance in the context of the prevalence of opioid misuse, a critical policy issue for the last several years. The more recent COVID-19 pandemic, however, further highlights the need to better estimate the true prevalence of health conditions within a population. And while the advent of big data has allowed us to greatly narrow traditional, frequentist confidence intervals, the fundamental issues of misclassification error, missing data and their implications for identifying population prevalence and trends remain front and center. Thus, it is essential that further work apply and refine our methods to give policy makers more accurate information on critical policy parameters and emerging health issues.

REFERENCES

Altekruse, S. F., Cosgrove, C. M., Altekruse, W. C., Jenkins, R. A., and Blanco, C. (2020). Socioeconomic risk factors for fatal opioid overdoses in the united states: Findings from the mortality disparities in american communities study (MDAC). *PloS ONE*, 15(1):e0227966.

Biemer, P. P. and Wiesen, C. (2002). Measurement error evaluation of self-reported drug use: a latent class analysis of the us national household survey on drug abuse. *Journal of the Royal Statistical Society: Series A*, 165:97–119.

Bollinger, C. R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics*, 73:387–399.

Bollinger, C. R. and David, M. H. (1997). Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association*, 92:827–835.

Bollinger, C. R. and Van Hasselt, M. (2017a). A Bayesian analysis of binary misclassification. *Economics Letters*, 156:68–73.

Bollinger, C. R. and Van Hasselt, M. (2017b). Bayesian moment-based inference in a regression model with misclassification error. *Journal of Econometrics*, 200:282–294.

Bross, I. (1954). Misclassification in 2x2 tables. *Biometrics*, 10:478–486.

Case, A. and Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century. *Proceedings of the National Academy of Sciences*, 112(49):15078–15083.

Chamberlain, G. and Imbens, G. (2003). Nonparametric applications of bayesian inference. *Journal of Business and Economic Statistics*, 21(1):12–18.

Chen, X., Hu, Y., and Lewbel, A. (2008a). Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments. *Economics Letters*, 100(3):381–384.

Chen, X., Hu, Y., and Lewbel, A. (2008b). A note on the closed-form identification of regression models with a mismeasured binary regressor. *Statistics and Probability Letters*, 78(12):1473–1479.

Evans, M., Guttman, I., Haitovsky, Y., and Swartz, T. (1996). Bayesian analysis of binary data subject to misclassification. In Berry, D. A., Chaloner, K. M., and Geweke, J. K., editors, *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, pages 67–78. Wiley.

Fendrich, M., Johnson, T. P., Sudman, S., Wislar, J. S., and Spiehler, V. (1999). Validity of drug use reporting in a high-risk community sample: a comparison of cocaine and heroin survey reports with hair tests. *American Journal of Epidemiology*, 149:955–962.

Gaba, A. and Winkler, R. L. (1992). Implications of errors in survey data: A Bayesian model. *Management Science*, 38:913–925.

Gosling, A. and Saloniki, E.-C. (2014). Correction of misclassification error in disability rates. *Health economics*, 23(9):1084–1097.

Gunawan, D., Panagiotelis, A., Griffiths, W., and Chotikapanich, D. (2017). Bayesian weighted inference from surveys. Unpublished working paper.

Gundersen, C., Kreider, B., Pepper, J., and Jolliffe, D. (2012). Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported. *Journal of the American Statistical Association*, 107:958–975.

Gustafson, P., Gelfand, A. E., Sahu, S. K., Johnson, W. O., Hanson, T. E., Joseph, L., and Lee, J. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables [with comments and rejoinder]. *Statistical Science*, pages 111–140.

Hahn, P. R., Murray, J. S., and Manolopoulou, I. (2016). A Bayesian partial identification approach to inferring the prevalence of accounting misconduct. *Journal of the American Statistical Association*, 111:14–26.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141:263–272.

Kadane, J. B. (1974). The role of identification in Bayesian theory. In Fienberg, S. and Zellner, A., editors, *Studies in Bayesian Econometrics and Statistics*. North-Holland.

Keyes, K. M., Cerdá, M., Brady, J. E., Havens, J. R., and Galea, S. (2014). Understanding the rural–urban differences in nonmedical prescription opioid use and abuse in the united states. *American journal of public health*, 104(2):e52–e59.

Kolodny, A., Courtwright, D. T., Hwang, C. S., Kreiner, P., Eadie, J. L., Clark, T. W., and Alexander, G. C. (2015). The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction. *Annual review of public health*, 36:559–574.

Kreider, B. and Pepper, J. V. (2007). Disability and employment: Reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association*, 102:432–441.

Kroutil, L. A., Vorburger, M., Aldworth, J., and Colliver, J. D. (2010). Estimated drug use based on direct questioning and open-ended questions: responses in the 2006 national survey on drug use and health. *International journal of methods in psychiatric research*, 19:74–87.

Ledgerwood, D. M., Goldberger, B. A., Risk, N. K., Lewis, C. E., and Price, R. K. (2008). Comparison between self-report and hair analysis of illicit drug use in a community sample of middle-aged men. *Addictive behaviors*, 33:1131–1139.

Lewbel, A. (2007). Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551.

Li, Y., Yao, L., Li, J., Chen, L., Song, Y., Cai, Z., and Yang, C. (2020). Stability issues of rt-pcr testing of sars-cov-2 for hospitalized patients clinically diagnosed with covid-19. *Journal of medical virology*.

Mertz, K. J., Janssen, J. K., and Williams, K. E. (2014). Underrepresentation of heroin involvement in unintentional drug overdose deaths in allegheny county, pa. *Journal of forensic sciences*, 59:1583–1585.

Meyer, B. D., Mittag, N., and Goerge, R. M. (2018). Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. Working Paper 25143, National Bureau of Economic Research.

Meyer, B. D., Mok, W. K. C., and Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226.

Moon, H. R. and Schorfheide, F. (2012). Bayesian and frequentist inference in partially identified models. *Econometrica*, 80:755–782.

Murphy, S. M., Friesner, D. L., and Rosenman, R. (2015). Opioid misuse among adolescents: new evidence from a misclassification analysis. *Applied health economics and health policy*, 13:181–192.

Nguimkeu, P., Denteh, A., and Tchernis, R. (2019). On the estimation of treatment effects with endogenous misreporting. *Journal of Econometrics*, 208:487–506.

Pepper, J. V. (2001). How do response problems affect survey measurement of trends in drug use? In Manski, C. F., Pepper, J. V., and Petrie, C. V., editors, *Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us*, pages 321–348. National Academy Press.

Poirier, D. J. (1980). Partial observability in bivariate probit models. *Journal of Econometrics*, 12:209–217.

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory*, 14:483–509.

Poirier, D. J. and Tobias, J. L. (2003). On the predictive distributions of outcome gains in the presence of an unidentified parameter. *Journal of Business & Economics Statistics*, 21(2):258–268.

Rahme, E., Joseph, L., and Gyorkos, T. W. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics*, 49:119–128.

Rubin, D. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9:130–134.

Rudd, R. A., Aleshire, N., Zibbell, J. E., and Gladden, R. M. (2016). Increases in drug and opioid overdose deaths – united states, 2000–2014. *American Journal of Transplantation*, 16:1323–1327.

Rudd, R. A., Paulozzi, L. J., Bauer, M. J., Burleson, R. W., Carlson, R. E., Dao, D., Davis, J. W., Dudek, J., Eichler, B. A., Fernandes, J. C., et al. (2014). Increases in heroin overdose deaths – 28 states, 2010 to 2012. *Morbidity and Mortality Weekly Report*, 63:849–854.

Ruhm, C. J. (2016). Taking the measure of a fatal drug epidemic. Technical report.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, 65:1350–1361.

Figures

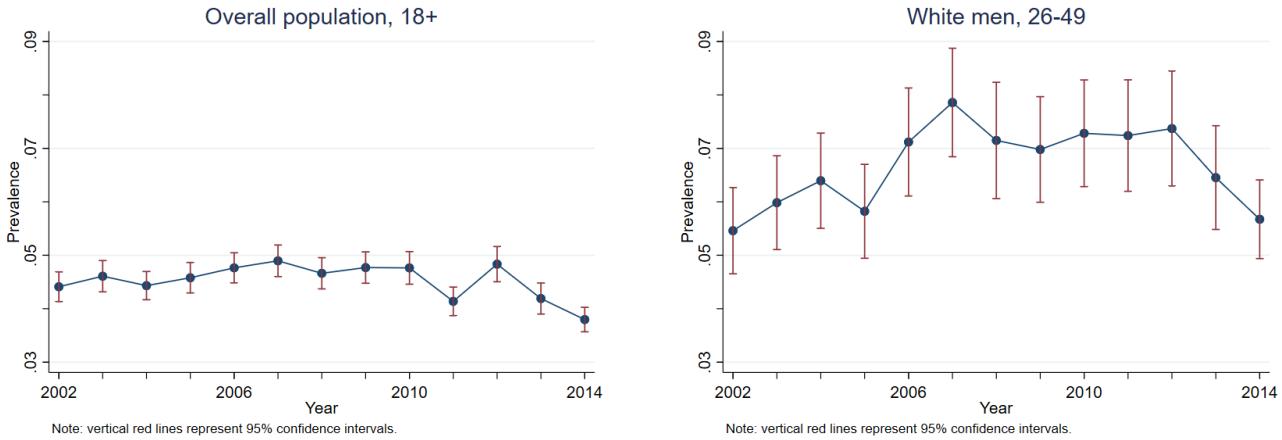

Figure 1: Past-year misuse of prescription pain relievers (2002-2014 NSDUH)

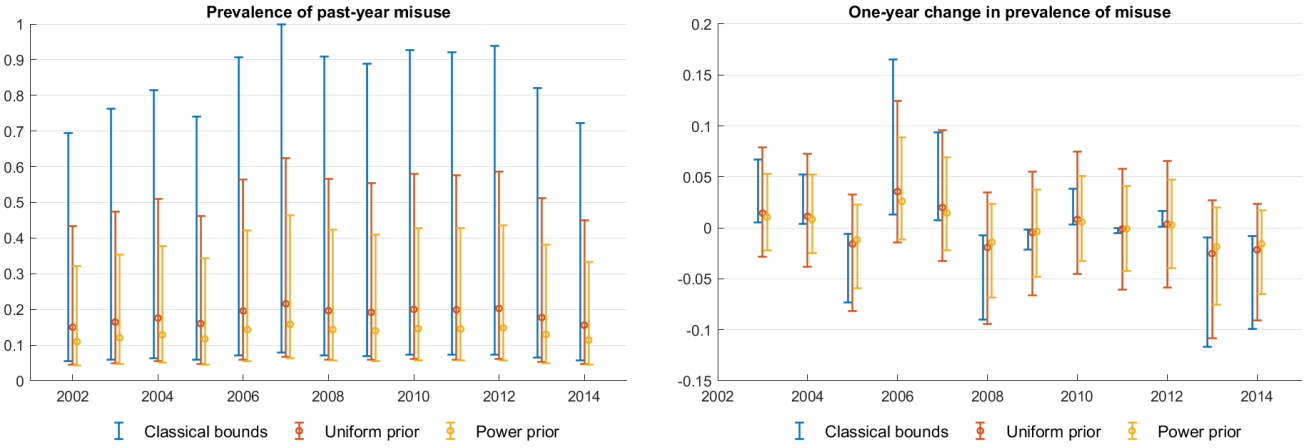

Figure 2: classical bounds and 95% HPD intervals for the prevalence and the one-year change in prevalence (q_t constant)

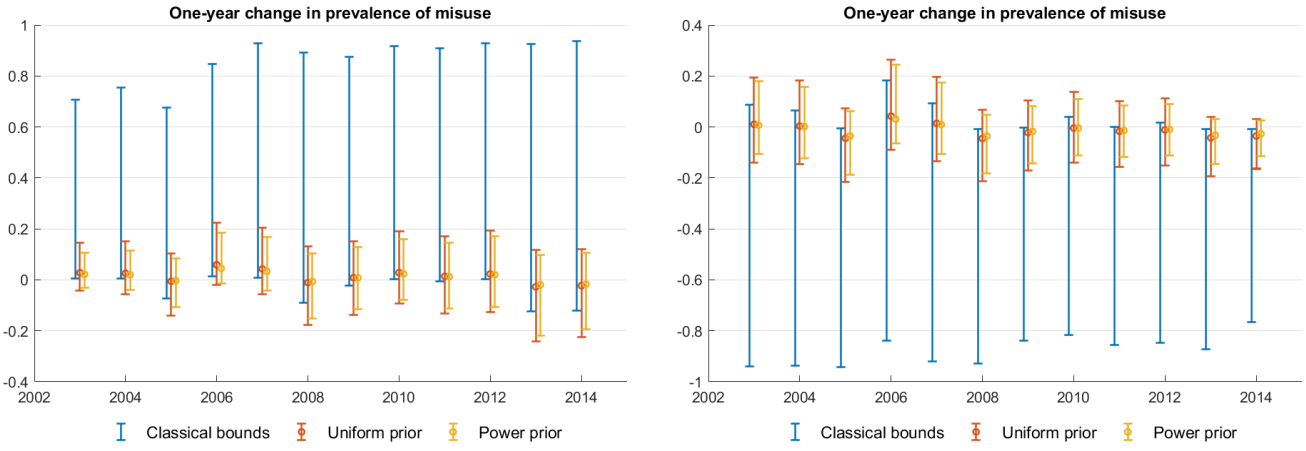

Figure 3: classical bounds and 95% HPD intervals for the one-year change in prevalence. Left: q_t non-decreasing (Case II); right: q_t non-increasing (Case III).

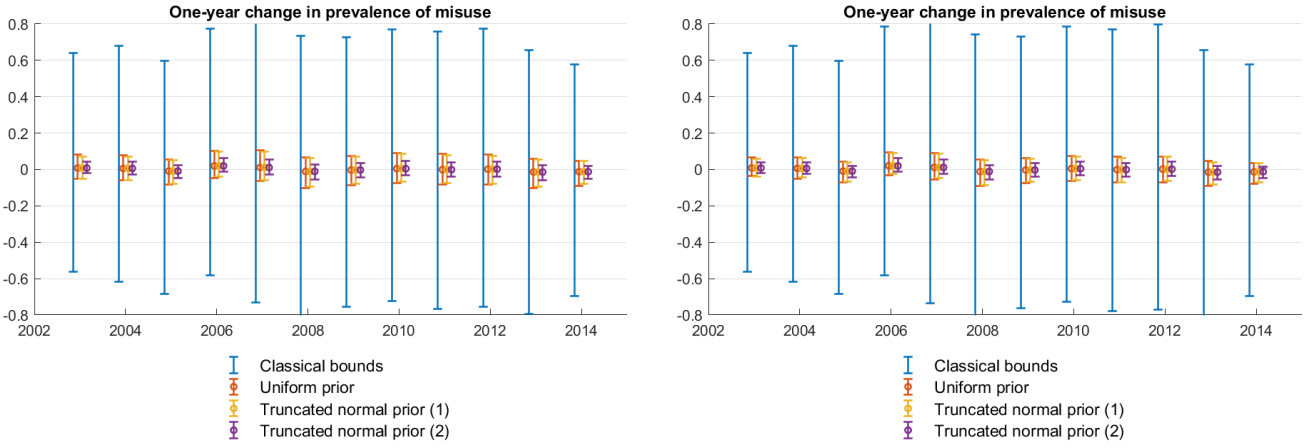

Figure 4: classical bounds and 95% HPD intervals for the one-year change in prevalence; q_t is assumed not to deviate more than 25% from the base rate. Left: $p = 0$; right: $p > 0$

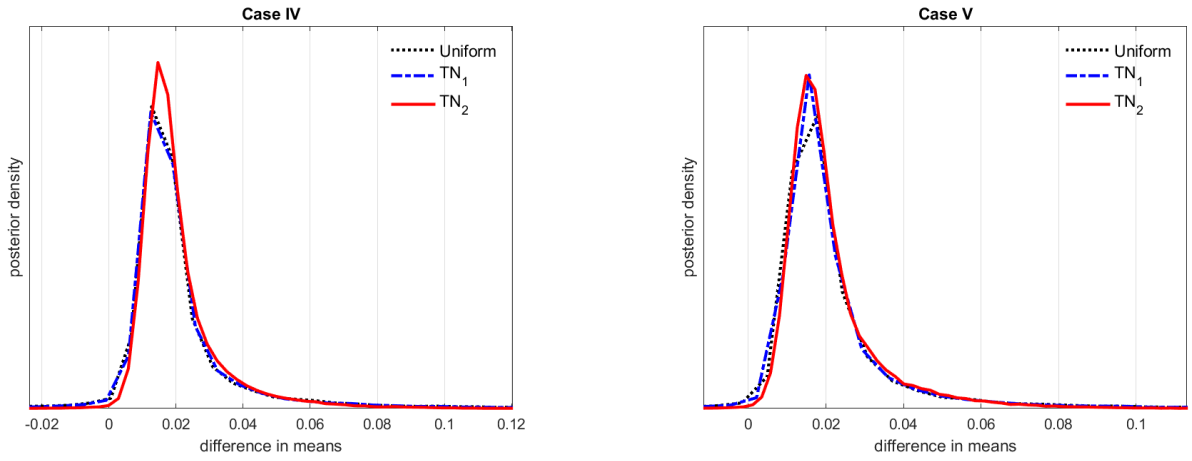

Figure 5: posterior of difference in average prevalence between the periods 2002-2005 and 2006-2009, Case IV (left; $p_t = 0$) and Case V (right; $p_t = p > 0$).