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

Martijn van Hasselt †1, Christopher R. Bollinger2, and Jeremy Bray1

1The University of North Carolina Greensboro, Department of Economics
2The 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) 334-5463. 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 the prevalence of opioid misuse increases 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. 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 multiple waves of the National Survey on Drug Use and Health (NSDUH).

Imperfect recall and intentional misreporting lead to misclassification errors in surveys. Imperfect recall is often related to the way survey questions are formulated (e.g., “Did you do X during the past 12 months?”), whereas intentional misreporting is a concern when the survey asks about behaviors that are stigmatized or even illegal (Pepper, 2001). While our empirical application is concerned with substance abuse, 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, and Gundersen et al. (2012) and Meyer et al. (2015) document misreporting of participation in the food stamp program (SNAP). A failure to account for these errors leads to biased estimates of the prevalence and trends.

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

---

1Throughout this paper we will refer to the mean as prevalence and to changes in the mean over time as trends.
models with unidentified and partially identified parameters has been an active area of research (e.g., Gustafson et al., 2005; Moon and Schorfheide, 2012; Hahn et al., 2016; Bollinger and Van Hasselt, 2017b). Following the notation and nomenclature of Moon and Schorfheide (2012), the main feature of such models is that the data are informative about a reduced-form parameter vector \( \phi \), but not about the structural parameter vector \( \theta \). The prior of the non-identified and partially identified elements of \( \theta \) is then updated by the data only to the extent that \( \phi \) and \( \theta \) are a priori dependent. When \( \theta \) 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 \( \phi \). Put differently, the support of the conditional prior of \( \theta \) is a function of \( \phi \).

This paper makes two main contributions. First, earlier Bayesian work on binary misclassification uses beta prior distributions for the prevalence and misclassification probabilities (Gaba and Winkler, 1992; Joseph et al., 1995; Evans et al., 1996; Rahme et al., 2000). In contrast, we follow the same general strategy that Bollinger and Van Hasselt (2017b) propose in the context of a regression model with a binary, misclassified regressor. We start by formulating the identified sets for the parameters of interest. These sets are subsequently used as inputs for specifying a range of priors that researchers might entertain in practice. The advantage of this approach is that it highlights what can and cannot be learned from the data and how the prior distribution remains influential even in large samples. The resulting posterior distribution takes a simple form, and a sample from this distribution can be easily generated. Also, unlike the aforementioned papers, we extend this framework to the analysis of repeated cross sections, which is relevant for many nationally representative survey samples, and analyze both the prevalence and the trend. Second, we extend the work of Pepper (2001). We consider a range of different assumptions about how misclassification probabilities evolve over time and the impact of these assumptions on identified sets for the prevalence and the trend. Our results show that under sufficiently strong assumptions, the direction (upward or downward) of a trend is identified.

In partially identified models, strong assumptions are often needed for the bounds to be informative. 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 may 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 distribution. Additionally, information embod-
ied by the prior can lead to more precise inference relative to a classical bounds analysis. This is apparent in our empirical application, where we estimate the prevalence and trend of prescription opioid misuse from the 2002-2014 waves of the NSDUH. We focus on opioid misuse because it is a current and 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). The classical bounding approach identifies the direction of the trend only under very strict assumptions. If these assumptions are not imposed, 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 the prevalence of 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 sets of assumptions about the misclassification rates. Section 3 discusses a range of prior distributions and shows how to draw a sample from the posterior. The empirical analysis of prescription opioid misuse is presented in section 4 and section 5 concludes. Details about the derivation of the identified sets can be found in the supplemental appendix.

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 $(\pi_t, p_t, q_t)$ through the equation

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

\footnote{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).}
We aim to learn about the prevalence $\pi_t$ 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 $(\pi_t, p_t, q_t)$ 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 $\pi_t$, 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

$$\mu_t \leq \pi_t \leq \frac{\mu_t}{M}, \quad t = 1, \ldots, T.$$  \hfill (2)
The trend in prevalence between periods $t$ and $t + j$ is bounded as follows.

$$
\Delta \mu_{t,j} \leq \Delta \pi_{t,j} \leq \frac{\Delta \mu_{t,j}}{M}, \quad \text{if} \quad \Delta \mu_{t,j} \geq 0,
$$

$$
\frac{\Delta \mu_{t,j}}{M} \leq \Delta \pi_{t,j} \leq \Delta \mu_{t,j}, \quad \text{if} \quad \Delta \mu_{t,j} < 0.
$$

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

$$
\mu_t \leq \pi_t \leq \frac{\mu_t}{M^+_t}, \quad t = 1, \ldots, T.
$$

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

$$
\Delta \mu_{t,j} \leq \Delta \pi_{t,j} \leq \frac{\mu_{t+j}}{M^+_{t+j}} - \mu_t, \quad \text{if} \quad \Delta \mu_{t,j} \geq 0,
$$

$$
\frac{\Delta \mu_{t,j}}{M^+_t} \leq \Delta \pi_{t,j} \leq \frac{\mu_{t+j}}{M^+_{t+j}} - \mu_t, \quad \text{if} \quad \Delta \mu_{t,j} < 0.
$$

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 or negative. In this case, a
decline in the observed trend may result from either a decrease in the true prevalence, or from an
increase in the true prevalence that is more than offset by an increase in false negative reporting.

**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

$$\mu_t \leq \pi_t \leq \frac{\mu_t}{M_t}, \quad t = 1, \ldots, T. \quad (6)$$

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.

$$\mu_{t+j} - \frac{\mu_t}{M_t} \leq \Delta \pi_{t,j} \leq \frac{\Delta \mu_{t,j}}{M_{t+j}}, \quad \text{if } \Delta \mu_{t,j} \geq 0,$$

$$\mu_{t+j} - \frac{\mu_t}{M_t} \leq \Delta \pi_{t,j} \leq \Delta \mu_{t,j}, \quad \text{if } \Delta \mu_{t,j} < 0. \quad (7)$$

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 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 that
is more than offset by a decrease in false negative reporting.

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

\[
\mu_t \leq \pi_t \leq \frac{\mu_t}{M}. \quad (8)
\]

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 \(a := 1 - x\) and \(b := 1 + x\) and let \(\Delta\pi_{t,j}^L\) and \(\Delta\pi_{t,j}^U\) denote the lower and upper bounds, respectively. It is shown in the supplemental appendix that these bounds are given by

\[
\Delta\pi_{t,j}^L = \begin{cases} 
\frac{\mu_t}{M} - \mu_t & \text{if } a\mu_{t+j} < b\mu_t, \\
\Delta\mu_{t,j} & \text{if } a\mu_{t+j} \geq b\mu_t, \quad M > \frac{(b-a)\sqrt{\mu_t}}{b\sqrt{a\mu_{t+j}} - a\sqrt{b\mu_t}}, \\
\min \left\{ \Delta\mu_{t,j}, \frac{\mu_t}{M} - \frac{\mu_t}{1 - (a/b)(1-M)} \right\} & \text{if } a\mu_{t+j} \geq b\mu_t, \quad M \leq \frac{(b-a)\sqrt{\mu_t}}{b\sqrt{a\mu_{t+j}} - a\sqrt{b\mu_t}}.
\end{cases} \quad (9)
\]

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

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

\[
\frac{\mu_t}{1 - (a/b)(1-M)} - \frac{\mu_t}{M} \leq \Delta\pi_{t,j} \leq \frac{\mu_t}{1 - (a/b)(1-M)} - \frac{\mu_t}{M}.
\]
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
\[
(1-x)\bar{q} \leq q_t \leq (1+x)\bar{q}; \text{ and (ii) } p_t = p.
\]

Since \( p \leq \mu_t \) for all \( t \), the upper bound for \( p \) is \( m := \min_s \mu_s \). 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
\[
\frac{\mu_t - m}{1 - m} \leq \pi_t \leq \frac{\mu_t}{M} \quad (11)
\]
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,
\[
\frac{\mu_{t+j} - p}{1 - p - a\bar{q}} - \frac{\mu_t - p}{1 - p - b\bar{q}} \leq \Delta \pi_{t,j} \leq \frac{\mu_{t+j} - p}{1 - p - b\bar{q}} - \frac{\mu_t - p}{1 - p - a\bar{q}} \quad (12)
\]
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

\[
\Delta \pi_{t,j}^L = \begin{cases} 
\frac{\mu_{t+1} - m}{c-m} - \frac{\mu_{t}}{M-m} & \text{if } \mu_{t+1} \leq c - \left( \frac{c-m}{M-m} \right)^2 M + \left( \frac{c-m}{M-m} \right)^2 \mu_t, \\
\frac{\mu_{t+1} - \mu_{t}}{c-p_t} - \frac{\mu_{t}}{M-p_t} & \text{if } c - \left( \frac{c-m}{M-m} \right)^2 M + \left( \frac{c-m}{M-m} \right)^2 \mu_t < \mu_{t+1} < c - M \left( \frac{c}{M} \right)^2 + \left( \frac{c}{M} \right)^2 \mu_t, \\
\frac{\mu_{t+1} - m}{c-m} & \text{if } \mu_{t+1} \geq c - M \left( \frac{c}{M} \right)^2 + \left( \frac{c}{M} \right)^2 \mu_t,
\end{cases}
\]

(13)

where

\[
p_t^* = \frac{c\sqrt{M - \mu_t} - M\sqrt{c - \mu_{t+1}}}{\sqrt{M - \mu_t} - \sqrt{c - \mu_{t+1}}}
\]

When \( \Delta \mu_{t,j} \geq 0 \), there is no convenient way to characterize \( \Delta \pi_{t,j}^L \), because it depends on the relative magnitudes of \((a, b, \mu_t, \mu_{t+1}, m, M)\). 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

\[
\Delta \pi_{t,j}^U = \begin{cases} 
\frac{\mu_{t+1}}{c} - \frac{\mu_{t}}{M} & \text{if } \mu_{t+1} \leq M - c \left( \frac{M}{c} \right)^2 + \left( \frac{M}{c} \right)^2 \mu_t, \\
\frac{\mu_{t+1} - \mu_{t}}{c-p_t} - \frac{\mu_{t}}{M-p_t} & \text{if } M - c \left( \frac{M}{c} \right)^2 + \left( \frac{M}{c} \right)^2 \mu_t < \mu_{t+1} < M - c \left( \frac{M-m}{c-m} \right)^2 + \left( \frac{M-m}{c-m} \right)^2 \mu_t, \\
\frac{\mu_{t+1} - m}{c-m} - \frac{\mu_{t}}{M-m} & \text{if } \mu_{t+1} \geq M - c \left( \frac{M-m}{c-m} \right)^2 + \left( \frac{M-m}{c-m} \right)^2 \mu_t,
\end{cases}
\]

(14)

When \( \Delta \mu_{t,j} \leq 0 \) instead, there is again no convenient expression for \( \Delta \pi_{t,j}^U \). 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 \( \pi_t \) 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 \((\pi_t, p_t, q_t)\), whereas we only restrict \((p_t, q_t)\) and investigate the implications for \(\pi_t\). 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 \(\mu, \pi\) and \(q\) be \(T\)-dimensional parameter vectors with \(t\)-th elements \(\mu_t, \pi_t\) 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 \(Y = \{Y_{it}; i = 1, \ldots, n_t, t = 1, \ldots, T\}\) 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

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

The likelihood is a function of \(\mu\) alone and therefore does not separately identify \(\pi, q\) and \(p\).

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

\[ f(\mu, \pi | Y) \propto f(Y | \mu) \cdot f(\mu) \cdot f(\pi | \mu) \]

\[ \propto f(\mu | Y) \cdot f(\pi | \mu). \]

A similar expression holds for the joint posterior of \( \mu, p \) and \( q \). The marginal posterior of \( \pi \) is obtained by integrating out \( \mu \):

\[ f(\pi | Y) \propto \int f(\mu | Y) f(\pi | \mu) d\mu. \]

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

### 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 \( \mu \) 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 \( \mu \). The conditional prior of \( \pi_t \) 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 \( \pi_t \), 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 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 \( \pi_t \) is then \( f(\pi_t | \mu) = C \mu_t^{1+\alpha} / (\pi_t^2 (\pi_t - \mu_t)^\alpha) \),

\(^3\)For example, Meyer et al. (2015) and Meyer and Mittag (2018) provide estimates of the amount of misreporting in SNAP participation.
which also places a relatively high probability on values of $\pi_t$ near $\mu_t$. Finally, suppose a researcher wishes to use a (conditional) prior on $\pi_t$ 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

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

$$= f(q_1|\mu) \prod_{t=2}^{T} f(q_t|q_{t-1}, \mu).$$

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 $q_{t-1} \leq q_t \leq 1-M_i^+$, 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 $t = 2, \ldots, T$)

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

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

12
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:

\[
f(q|\mu) = f(q_T|\mu) \prod_{t=1}^{T-1} f(q_{T-t}|q_{T-t+1}, \ldots, q_T, \mu)
\]

\[
= f(q_T|\mu) \prod_{t=1}^{T-1} f(q_{T-t}|q_{T-t+1}, \mu).
\]

For \(q_T\), we choose a prior (conditional on \(\mu\)) 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 \(Y\) 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 \((\mu, \pi, q)\). 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 \( Y_t = (Y_{1t}, \ldots, Y_{n_t}) \) be the observed sample at time \( t \), so that \( Y = (Y_1, \ldots, Y_T) \). Similarly, let \( \tilde{Y}_t = (\tilde{Y}_{1t}, \ldots, \tilde{Y}_{n_t}) \) be a random sample from the population at time \( t \), and let \( \tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_T) \). The vector of sampling weights at time \( t \) is \( w_t = (w_{1t}, \ldots, w_{n_t}) \) 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 \( \tilde{p}_t \)

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

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

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

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

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

\[
\propto f(\mu) f(\tilde{Y}|\mu)
\]

\[
\propto f(\mu) \prod_{t=1}^{T} \mu^{\tilde{n}_{t1}} (1 - \mu)^{\tilde{n}_{t0}},
\]

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 $f(\mu, \tilde{Y}|Y, w)$ can now be generated by first drawing $\tilde{Y}$ from (16) and then drawing $\mu$ 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 $\tilde{Y}_i$ 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 $\mu$ from the posterior distribution in (17);
3. Given the sampled value $\mu$:
   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 $\mu$, 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 $\pi$ 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. The Sample

For our empirical analysis we use publicly available data from the 2002-2014 waves of the National Survey on Drug Use and Health (NSDUH).\footnote{Public use files for 2015-2017 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.} 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.

We restrict the sample to individuals who were at least 18 years old at the time of the NSDUH interview. Our analysis variable is an indicator for the 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.

4.2. Observed Prevalence

The observed prevalence of past-year misuse of prescription pain relievers ($\mu_t$) is shown in Figure 1. 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 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.

Figure 2 shows the prevalence of self-reported misuse 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 comparing Figures 1 and 2, the observed prevalence among middle-aged white men is substantially higher than in the overall population. We therefore focus the remainder of our analysis on this group. Self-reported past-year misuse of prescription pain relievers rose from about 5.4% in 2002 to 6.8% in 2010, an increase of more than 25%. Figure 2 also suggests that the prevalence may have started to decline in 2013. Throughout the sample period, however, the 95% confidence intervals mostly overlap, again making it difficult to discern any sustained upward or downward trend.
4.3. Prior Specification

The posterior results presented in the next section are based on 100,000 simulated draws from the posterior distribution. As noted in section 3.3, generating a random draw from the posterior involves generating a draw from the posterior $f(\mu|Y)$ followed by generating a random draw from the conditional prior $f(q|\mu)$ or $f(p,q|\mu)$. We use the conditional priors discussed in more detail in section 3.2. When $q_t$ is thought to be non-decreasing (Assumption C-II) or non-increasing (Assumption C-III), we use the mixture probability $\lambda = 0.9$ that $q_t$ remains the same between adjacent periods.

Under Assumption C-IV that $q_t$ deviates at most $100(x)\%$ from a base rate $\bar{q}$, 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 $[0, m]$.

In the following section we present estimates of the classical bounds and the Bayesian 95% highest posterior density (HPD) intervals. Within these intervals, circles indicate the posterior mean. While other posterior summaries can be calculated, we mainly 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.

4.4. Posterior Summaries for the True Prevalence

When $q_t$ is assumed to be constant, Figure 3 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 $q^*$. The HPD intervals for the trend are mostly narrower but not always (e.g., 2003 and 2005). This may 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 0.5958 under both priors. While these probabilities do not reveal a clear trend in prevalence, there is strong evidence that the prevalence increased—relative to the previous year—in 2006, 2010 and 2012, with posterior probabilities exceeding 90%. For the remaining years, the posterior is less informative. The posterior probabilities and the locations of the posterior means seem robust to the choice of the two priors for $q^*$ (uniform and power) that we consider here.

<table>
<thead>
<tr>
<th>year</th>
<th>uniform prior</th>
<th>power prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>0.5958</td>
<td>0.5958</td>
</tr>
<tr>
<td>2004</td>
<td>0.2466</td>
<td>0.2466</td>
</tr>
<tr>
<td>2005</td>
<td>0.6108</td>
<td>0.6108</td>
</tr>
<tr>
<td>2006</td>
<td>0.9716</td>
<td>0.9716</td>
</tr>
<tr>
<td>2007</td>
<td>0.4760</td>
<td>0.4760</td>
</tr>
<tr>
<td>2008</td>
<td>0.3254</td>
<td>0.3254</td>
</tr>
<tr>
<td>2009</td>
<td>0.3000</td>
<td>0.3000</td>
</tr>
<tr>
<td>2010</td>
<td>0.9613</td>
<td>0.9613</td>
</tr>
<tr>
<td>2011</td>
<td>0.0510</td>
<td>0.0510</td>
</tr>
<tr>
<td>2012</td>
<td>0.9052</td>
<td>0.9052</td>
</tr>
<tr>
<td>2013</td>
<td>0.0610</td>
<td>0.0610</td>
</tr>
<tr>
<td>2014</td>
<td>0.1499</td>
<td>0.1499</td>
</tr>
</tbody>
</table>

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

Estimates of the identified sets and 95% HPD intervals for the trend are shown in Figure 4, 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) identified sets. Comparing the left- and right-hand sides of Figure 4 we also see that the location of the identified sets 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, 2009-2010 and 2011-2012. For example, assuming that $q_t$ was non-decreasing, the posterior probability of an increase during the period 2005-2006 was 97.4% 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 result was to be expected. For
example, from Figure 2 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).

<table>
<thead>
<tr>
<th>year</th>
<th>$q_t$ non-decreasing (Case II)</th>
<th>$q_t$ non-increasing (Case III)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uniform</td>
<td>power</td>
</tr>
<tr>
<td>2003</td>
<td>0.6343</td>
<td>0.6326</td>
</tr>
<tr>
<td>2004</td>
<td>0.3107</td>
<td>0.3083</td>
</tr>
<tr>
<td>2005</td>
<td>0.6446</td>
<td>0.6437</td>
</tr>
<tr>
<td>2006</td>
<td>0.9741</td>
<td>0.9742</td>
</tr>
<tr>
<td>2007</td>
<td>0.5208</td>
<td>0.5203</td>
</tr>
<tr>
<td>2008</td>
<td>0.3809</td>
<td>0.3802</td>
</tr>
<tr>
<td>2009</td>
<td>0.3564</td>
<td>0.3563</td>
</tr>
<tr>
<td>2010</td>
<td>0.9650</td>
<td>0.9650</td>
</tr>
<tr>
<td>2011</td>
<td>0.1182</td>
<td>0.1184</td>
</tr>
<tr>
<td>2012</td>
<td>0.9134</td>
<td>0.9134</td>
</tr>
<tr>
<td>2013</td>
<td>0.1307</td>
<td>0.1298</td>
</tr>
<tr>
<td>2014</td>
<td>0.2187</td>
<td>0.2173</td>
</tr>
</tbody>
</table>

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

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 of either $p = 0$ (Case IV) or $p > 0$ (Case V). Figure 5 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 posterior probabilities of a positive trend in Table 3 show evidence for an increase in misuse in the periods 2005-2006, 2009-2010 and 2011-2012. For these years, the probabilities increase 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 probability of an increase in true prevalence between 2005 and 2006 is 91.8% 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 95.3%.
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 \( TN_2 \).

So far, we have focused on the quantity \( \Delta \pi_{t+1} = \pi_{t+1} - \pi_t \) and the posterior probability that this one-year change is positive. There are, of course, many other parameters that could be of interest. For example, inspection of Figure 2 suggests that the prevalence of misuse may have been higher in the period 2006-2009 compared to 2002-2005, followed by a further increase in the period 2010-2012. To assess the posterior evidence for this, we use the posterior sample of \( \pi_t \) values and calculate difference between the average true prevalence during 2002-2005 (\( \bar{\pi}_0 \)) and the average during 2006-2009 (\( \bar{\pi}_1 \)). Figure 6 shows kernel density estimates of the posterior of \( \bar{\pi}_1 - \bar{\pi}_0 \), assuming bounded variation in \( q_t \) and either no false positives (Case IV) or a constant rate of false positives (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 false positive rate has a prior power distribution as well. 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 was higher in 2006-2009 than in 2002-2005. The posterior distributions are centered on a mean difference of about 1.3 to 1.4 percentage points (this corresponds increases in the average prevalence of roughly 16% under assumption C-IV and 22% under assumption C-V). The 95% HPD intervals cover mostly positive values and the posterior probability of an increase in the average prevalence

<table>
<thead>
<tr>
<th>year</th>
<th>(Case IV: ( p = 0 ))</th>
<th>(Case V: ( p \geq 0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uniform ( TN_1 ) ( TN_2 )</td>
<td>uniform ( TN_1 ) ( TN_2 )</td>
</tr>
<tr>
<td>2003</td>
<td>0.5712 (0.5740) (0.5845)</td>
<td>0.5798 (0.5811) (0.5891)</td>
</tr>
<tr>
<td>2004</td>
<td>0.3004 (0.2998) (0.2714)</td>
<td>0.2844 (0.2834) (0.2593)</td>
</tr>
<tr>
<td>2005</td>
<td>0.5873 (0.5877) (0.6012)</td>
<td>0.5967 (0.5952) (0.6073)</td>
</tr>
<tr>
<td>2006</td>
<td>0.8970 (0.9028) (0.9454)</td>
<td>0.9177 (0.9213) (0.9534)</td>
</tr>
<tr>
<td>2007</td>
<td>0.4805 (0.4821) (0.4793)</td>
<td>0.4809 (0.4814) (0.4768)</td>
</tr>
<tr>
<td>2008</td>
<td>0.3725 (0.3661) (0.3478)</td>
<td>0.3619 (0.3577) (0.3423)</td>
</tr>
<tr>
<td>2009</td>
<td>0.3428 (0.3417) (0.3196)</td>
<td>0.3335 (0.3313) (0.3137)</td>
</tr>
<tr>
<td>2010</td>
<td>0.8829 (0.8875) (0.9316)</td>
<td>0.9003 (0.9053) (0.9409)</td>
</tr>
<tr>
<td>2011</td>
<td>0.1302 (0.1252) (0.0805)</td>
<td>0.1129 (0.1076) (0.0714)</td>
</tr>
<tr>
<td>2012</td>
<td>0.8264 (0.8308) (0.8722)</td>
<td>0.8436 (0.8458) (0.8827)</td>
</tr>
<tr>
<td>2013</td>
<td>0.1391 (0.1339) (0.0922)</td>
<td>0.1216 (0.1161) (0.0832)</td>
</tr>
<tr>
<td>2014</td>
<td>0.2219 (0.2180) (0.1808)</td>
<td>0.2057 (0.2031) (0.1727)</td>
</tr>
</tbody>
</table>
exceeds 95% for both cases and all priors considered here. Results presented in the supplemental appendix show that the average prevalence of misuse likely increased further in the period 2010-2012.

<table>
<thead>
<tr>
<th>Case</th>
<th>Prior</th>
<th>mean</th>
<th>std. dev.</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
<th>95% HPD</th>
<th>(P(+))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uniform</td>
<td>0.0132</td>
<td>0.0205</td>
<td>-0.0110</td>
<td>0.0103</td>
<td>0.0624</td>
<td>[-0.0151,0.0569]</td>
<td>0.9501</td>
</tr>
<tr>
<td>IV</td>
<td>(TN_1)</td>
<td>0.0133</td>
<td>0.0194</td>
<td>-0.0083</td>
<td>0.0103</td>
<td>0.0592</td>
<td>[-0.0116,0.0550]</td>
<td>0.9539</td>
</tr>
<tr>
<td></td>
<td>(TN_2)</td>
<td>0.0129</td>
<td>0.0101</td>
<td>0.0021</td>
<td>0.0107</td>
<td>0.0405</td>
<td>[0.0000,0.0352]</td>
<td>0.9867</td>
</tr>
<tr>
<td></td>
<td>uniform</td>
<td>0.0137</td>
<td>0.0186</td>
<td>-0.0044</td>
<td>0.0106</td>
<td>0.0572</td>
<td>[-0.0082,0.0516]</td>
<td>0.9638</td>
</tr>
<tr>
<td>V</td>
<td>(TN_1)</td>
<td>0.0137</td>
<td>0.0175</td>
<td>-0.0025</td>
<td>0.0107</td>
<td>0.0559</td>
<td>[-0.0070,0.0490]</td>
<td>0.9674</td>
</tr>
<tr>
<td></td>
<td>(TN_2)</td>
<td>0.0133</td>
<td>0.0098</td>
<td>0.0026</td>
<td>0.0110</td>
<td>0.0400</td>
<td>[0.0003,0.0340]</td>
<td>0.9909</td>
</tr>
</tbody>
</table>

Table 4: posterior summary of difference in average prevalence between the periods 2002-2005 and 2006-2009. \(P(+)\) 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 restrict our analysis to this population. We find that the estimated identified sets (intervals) are mostly 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.

Our analysis 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. To this end, further research should be done to calibrate the prior distribution. Secondary data sources, for example on emergency room visits or drug-related fatalities, could provide relevant prior information that is likely to have a significant impact on the resulting posterior inferences. 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. We aim to pursue these issues in future work.
REFERENCES


Figure 1: Past-year misuse of prescription pain relievers (2002-2014 NSDUH)
Figure 2: Past-year misuse of prescription pain relievers (2002-2014 NSDUH)

Figure 3: Classical bounds and 95% HPD intervals for the prevalence and the one-year change in prevalence ($q_t$ constant)
Figure 4: 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 5: 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 6: posterior of difference in average prevalence between the periods 2002-2005 and 2006-2009, Case IV (left) and Case V (right).