CENTER FOR POLICY RESEARCH
THE MAXWELL SCHOOL
WORKING PAPER SERIES

A Framework for Measurement Error in Self-Reported Health Conditions

Ling Li and Perry Singleton

Paper No. 191
August 2016

ISSN: 1525-3066

Maxwell Syracuse University | Center for Policy Research
426 Eggers Hall
Syracuse University
Syracuse, NY 13244-1020
(315) 443-3114/ email: ctrpol@syr.edu
http://www.maxwell.syr.edu/CPR_Working_Papers.aspx

CENTER FOR POLICY RESEARCH –Summer 2016

Leonard M. Lopoo, Director
Professor of Public Administration and International Affairs (PAIA)

Associate Directors

Margaret Austin
Associate Director, Budget and Administration

John Yinger
Trustee Professor of Economics and PAIA
Associate Director, Metropolitan Studies

SENIOR RESEARCH ASSOCIATES

Badi Baltagi, Economics
Robert Bifulco, PAIA
Leonard Burman, PAIA
Thomas Dennison, PAIA
Alfonso Flores-Lagunes, Economics
Sarah Hamersma, PAIA
William C. Horrace, Economics
Yilin Hou, PAIA
Hugo Jales, Economics
Duke Kao, Economics
Jeffrey Kubik, Economics
Yoonseok Lee, Economics
Amy Lutz, Sociology
Yingyi Ma, Sociology
Jerry Miner, Economics
Cynthia Morrow, PAIA
Jan Ondrich, Economics
John Palmer, PAIA
David Popp, PAIA
Stuart Rosenthal, Economics
Michah Rothbart, PAIA
Rebecca Schewe, Sociology
Amy Ellen Schwartz, PAIA/Economics
Perry Singleton, Economics
Michael Wasylenko, Economics
Peter Wilcoxen, PAIA

GRADUATE ASSOCIATES

Emily Cardon, PAIA
Carlos Diaz, Economics
Alex Falevich, Economics
Wancong Fu, Economics
Boqian Jiang, Economics
Hyunseok Jung, Economics
Yusun Kim, PAIA
Ling Li, Economics
Michelle Lofton, PAIA
Judson Murchie, PAIA
Brian Ohl, PAIA
Jindong Pang, Economics
Laura Rodriquez-Ortiz, PAIA
Fabio Rueda De Vivero, Economics
David Schwegman, PAIA
Shulin Shen, Economics
Iuliia Shybalkina, PAIA
Kelly Stevens, PAIA
Saied Toossi, PAIA
Rebecca Wang, Sociology
Xirui Zhang, Economics

STAFF

Kelly Bogart, Administrative Specialist
Kathleen Nasto, Administrative Assistant
Candi Patterson, Computer Consultant
Mary Santy, Administrative Assistant
Katrina Wingle, Administrative Assistant

Abstract

This study develops and estimates a model of measurement error in self-reported health conditions. The model allows self-reports of a health condition to differ from a contemporaneous medical examination, prior medical records, or both. The model is estimated using a two-sample strategy, which combines survey data linked medical examination results and survey data linked to prior medical records. The study finds substantial inconsistencies between self-reported health, the medical record, and prior medical records. The study proposes alternative estimators for the prevalence of diagnosed and undiagnosed conditions and estimates the bias that arises when using self-reported health conditions as explanatory variables.

JEL No. I12, J22

Keywords: Measurement Error, Disease Prevalence, Diabetes, Hypertension

Author: Ling Li, Department of Economics, Center for Policy Research, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244 (315) 443-9056; Lli37@syr.edu

Corresponding Author:
Perry Singleton; Associate Professor, Department of Economics; Senior Research Associate, Center for Policy Research; Syracuse University. Contacts: 426 Eggers Hall; Syracuse University, Syracuse, NY 13244; (315) 443-3114; psinglet@syr.edu.

The authors would like to thank John Cawley, Gary Engelhardt, Alfonso Flores-Lagunes, Bruce Meyer, Jeffrey Kubik, Yoonseok Lee, and seminar participants at Cornell University for valuable comments.

I. Introduction

Several surveys collect data on previously diagnosed health conditions, and these data are used for in a variety of applications, from estimating the prevalence of health conditions to estimating the effect of health conditions on labor market outcomes. However, several recent studies question the validity of self-reported health conditions. For example, Baker, Stabile, and Deri (2004) link survey data to prior medical records and find substantial inconsistencies between self-reported health conditions and the medical record. Additionally, Johnston, Propper, and Shields (2009) use survey data linked to results from a medical examination and find substantial inconsistencies between self-reported hypertension and a clinical test. These inconsistencies lead to measurement error, which not only bias the estimated prevalence of health conditions, but also the correlation between health conditions and other outcomes of interest.[1]

To validate data on self-reported health conditions for the US, some studies use survey data linked to medical records, while others use survey data linked to a medical examination.[2] Currently, no study uses survey data linked to both the medical record and a medical examination, as no such data linkage exists. To address this shortcoming, this study proposes a two-sample estimation strategy. The study first develops a model of measurement error in self-reported health conditions. The model is composed of three binary variables: an indicator

[1] For a reviews of measurement error in survey data, see Bound, Brown, and Mathiowetz (2001) and Meyer, Mok, and Sullivan (2015).
[2] Studies that validate self-reported health or health behaviors using administrative data include Madow (1973); Martin et al. (2000); and Suziedelyte and Johar (2013). Studies that validate self-reported health conditions using medical examination results include Butler, Burkhauser, Mitchell, and Pincus (1987), Cawley and Choi (2015), and Johnston, Propper, and Shields (2009)

of the self-report, an indicator of the medical examination result, and an indicator of the medical record. With three binary variables, the joint probability distribution consists of eight population moments. The study then estimates these moments using two separate data linkages: survey data linked to medical records and survey data linked to medical exam results. The latter come from the study Baker, Stabile, and Deri (2004), who use the Canadian National Population Health Survey linked to the Ontario Health Insurance Plan (OHIP). The former comes from the National Health and Nutrition Examination Survey. Given the available data, the analysis focuses on two conditions: hypertension and diabetes.

The study yields several results. First, the study provides an alternative estimate for the prevalence of undiagnosed health conditions. In many studies, undiagnosed conditions are defined as those that are not self-reported at the time of the survey, but are detected upon medical examination.[3] However, this definition may overstate the prevalence of undiagnosed conditions if individuals had been previously diagnosed – and thus have a medical record – but simply fail to report the condition at the time of the survey. After accounting for this possibility, the prevalence of undiagnosed hypertension decreases from 9.0 percent to 2.4 percent, and the prevalence of undiagnosed diabetes decreases from 2.7 percent to 1.3 percent.

Second, the study provides an alternative estimate for the prevalence of diagnosed health conditions. In many studies, diagnosed conditions are defined as those that are self-reported, regardless of whether they test positive for the condition at the time of the survey.

[3] For example, Cowie et al (2006) estimate that approximately 2.8 percent of the population in 2002 had undiagnosed diabetes, and Sug Yoon et al (2012) estimate that approximately 5.2 percent of the population in 2009 had undiagnosed hypertension.

While this may be plausible for individuals whose health had improved, this may also reflect individuals who were never formally diagnosed, but report the condition nonetheless, perhaps to justify non-employment or eligibility for disability benefits.[4] After accounting for this possibility, the prevalence of diagnosed hypertension decreases from 20.0 percent to 15.5 percent, and the prevalence of diagnosed diabetes decreases from 6.0 percent to 5.0 percent.

Third, the study examines the bias that may arise when estimating the causal effect of health conditions on other outcomes of interest, such as labor supply.[5] In a simplified model, the bias is proportional to $cov(S, u)/var(S)$, where S is the self-reported variable and u is the measurement error.[6] The proportional bias is estimated for various definitions of *true* health using the estimated distribution of measurement error. According to the calculations, the proportion bias ranges from 0.308 to 0.710 for hypertension from 0.187 to 0.363 for diabetes. The bias is smallest when true health is defined by the medical record only and greatest when true health is defined by the medical examination only.

The results underscore the potential biases that may arise when using self-reported health conditions. A notable limitation is that, to estimate measurement error, the study employs a two-sample strategy using data from Canada and the US. Ideally, survey data would be linked to both medical records and medical examinations, obviating the need for the two-

[4] Studies that examine the endogeneity of self-reported health include Bound (1991); Dwyer and Mitchell (1999); and Benitez-Silva, Buchinsky, Chan, Cheidvasser, and Rust (2004).
[5] Currie and Madrian (1999) raise concern for undiagnosed health conditions when estimating the effect of health on labor market outcomes. However, it remains unclear how undiagnosed health conditions affect work capacity, or how selection into medical screening affects the association between self-reported health and labor market outcomes.
[6] For a more technical discussion of measurement error, see Bound, Brown and Mathiowetz (2001).

sample strategy. And, when using two-sample strategy, the data would ideally represent the same populations. However, there is no representative survey of the US that links survey data to comprehensive medical records. Thus, this study is the first reasonable attempt to estimate measurement error in self-reported health conditions – relative to both the medical record and medical examination – given the available data.

II. Methodology

A. Model of Measurement Error

The empirical objective is to determine whether self-reports of specific health conditions are consistent with a contemporaneous medical examination or prior medical records. This is accomplished in two steps. The first step is to specify a population-level model of measurement error, which specifically allows the self-report of a health condition to differ from a medical examination, prior medical records, or both. The second step is to estimate the moments of the model using population-based survey data.

The model of measurement error consists of three binary variables. The first variable is a self-report of a previous diagnosis for the condition: the variable equals one if a survey participant reports a previous diagnosis and zero otherwise. The second variable is the result of a medical examination at the time of the survey: the variable equals one if a survey participant tests positive for the condition and zero otherwise. The third variable is an indicator of the medical record: the variable equals one if the survey participant has a medical record of the condition and zero otherwise.

With three binary variables, the joint probability distribution consists of eight moments. The joint probability distribution is given by the following table:

	Medical Examination (E)	
Self-Report (S)	No (E=0)	Yes (E=1)
No (S=0)	$\pi_{00} = \pi_{000} + \pi_{001}$	$\pi_{01} = \pi_{010} + \pi_{011}$
Yes (S=1)	$\pi_{10} = \pi_{100} + \pi_{101}$	$\pi_{11} = \pi_{110} + \pi_{111}$

The rows correspond to the self-report, and the columns correspond to the medical examination. These two variables yield four population moments, denoted π_{SE}. The first subscript corresponds to the value of the self-report, and the second subscript corresponds to the value of the medical examination. For example, π_{00} represents the percent of the population who do not self-report the condition and who do not test positive for the condition at the time of the survey. To incorporate the medical record, each π_{SE} is disaggregated into those with and without a medical record, denoted π_{SER}. Thus, π_{000} represents the percent of the population who do not self-report the condition, who do not test positive for the condition at the time of the survey, and who do not have medical record of the condition.

The model has three important empirical applications. First, the model highlights the difficulty in defining and measuring the prevalence of undiagnosed health conditions. To measure prevalence, several studies define undiagnosed conditions as those that are not self-reported at the time of the survey, but are detected upon medical examination. This case corresponds to π_{01} in the model above. However, π_{01} may include individuals who had been previously diagnosed, and thus have a medical record, but who fail to report the condition at the time of the survey. This occurs with probability π_{011}. An important consideration is

whether π_{011} should be excluded from estimates of undiagnosed health conditions. If so, prevalence of undiagnosed conditions should be measured as π_{010}, rather than π_{01}.

Second, the model highlights the difficulty in defining and measuring the prevalence of diagnosed health conditions. To measure prevalence, several studies define diagnosed conditions as those that are self-reported, regardless of whether they test positive for the condition at the time of the survey. This case corresponds to $\pi_{11} + \pi_{10}$ in the model above. However, the latter term may include individuals who were never formally diagnosed, but report the condition nonetheless, perhaps to justify non-employment or eligibility for disability benefits. This occurs with probability π_{010}. An important consideration is whether π_{010} should be excluded from estimates of diagnosed health conditions. If so, the prevalence of diagnosed conditions should be measured as $\pi_{11} + \pi_{101}$, rather than $\pi_{11} + \pi_{10}$.

Third, the model helps to characterize the biases that may arise when using self-reported health conditions as explanatory variables. For example, a structural model of an outcome Y as a function of health condition S^* is given by the following equation:

(1) $$Y = \beta_0 + \beta_1 S^* + \varepsilon.$$

For example, the model may be used to examine the causal effect of a health condition S^* on labor supply Y. The causal effect is denoted by the parameter β_1. The variable S^* is defined by the states of health that do and do not affect the outcome. For example, S^* may be defined by the result of a medical examination, regardless of whether the condition had been previously diagnosed or self-reported, as in Johnston, Propper, and Shields (2009). Alternatively, S^* may be defined solely by the medical record, as in Baker, Stabile, and Deri (2004). Another

possibility is that the S^* is measured by a combination of a medical examination and the medical record.

When true health S^* is replaced with self-reported health, denoted S, the estimate of β_1 may be biased. To characterize the bias, the self-report of the health condition is expressed as the sum of S^* and an error term u: $S = S^* + u$. When S^* is substituted in (1), the equation becomes

(2) $Y = \beta_0 + \beta_1 S + \varepsilon - \beta_1 u.$

By construction, S is correlated with u. If ε is uncorrelated with S^* and u, then the least squares estimate of β_1 converges in probability to $\beta_1[1 - cov(S,u)/var(S)]$. Thus, the bias due to measurement error is proportional to $cov(S,u)/var(S)$. This bias may be estimated given a definition of true health S^* and values for the eight population moments π_{SER}.

B. Data and Estimation Strategy

To estimate the eight population moments π_{SER}, the study would ideally use survey data matched to both medical examination results and medical records. However, no such data exist for a representative sample of the US population. As an alternative, this study uses two separate data linkages: survey data linked to medical records, and survey data linked to medical exam results. Intuitively, the joint distribution is composed of several moments. Some moments can be estimated using survey data linked to medical records; others can be estimated from survey data linked to medical exam results. These estimates, combined, yield the underlying joint distribution π_{SER} in the population.

Survey data linked to medical examinations come from the National Health and Nutrition Examination Survey (NHANES). The NHANES was designed, in part, to estimate the

prevalence of undiagnosed health conditions in the US population. This is accomplished by first asking participants if they have ever been diagnosed for certain health conditions by a medical professional, and then testing participants for these conditions by medical examination. These data are used two estimate four population moments π_{11}, π_{00}, π_{10}, and π_{01}.

Information on survey data linked to medical records comes from a study by Baker, Stabile, and Deri (2004). The study examines whether self-reported health conditions in survey data are consistent with previous medical records. The survey data come from 1996/1997 version of the Canadian National Population Health Survey (CNPHS), and the data from medical record come from the Ontario Health Insurance Plan (OHIP). The study is limited to Ontario, as the OHIP data come from Ontario only. As the authors state, OHIP records provide a comprehensive view of previous health services, as alternative services are either expensive or prohibited.

Using these data, the authors find substantial inconsistencies between self-reported health and the medical record. To characterize these inconsistencies, the authors calculate rates of false-negatives and false-positives for various health conditions. The rate of false-negatives is defined as the percent of individuals who fail to self-report a medical condition, conditional on having a medical record for the condition. Conversely, the rate of false-positives is defined as the percent of individuals who self-report a medical condition, conditional on having no medical record for the condition. They find that, for many conditions, more than 50 percent of individuals who have a medical record for a condition fail to report it in the survey. Rates of false-positive reporting are considerably lower.

The estimated rates of false-negative and false-positive reporting are used to identify population moments of the model above. To link the two, the rate of false-negative reporting is expressed as,

$R_{FN} = \frac{\pi_{001}+\pi_{011}}{\pi_{001}+\pi_{111}+\pi_{101}+\pi_{011}}$.

Similarly, the rate of false-positive reporting is expressed as,

$R_{FP} = \frac{\pi_{110}+\pi_{100}}{\pi_{000}+\pi_{110}+\pi_{100}+\pi_{010}}$.

The study by Baker, Stabile, and Deri (2004) provides estimates of R_{FN} and R_{FP}.

The model contains eight population moments, but the data thus far provide only six: π_{11}, π_{00}, π_{10}, π_{01}, R_{FN}, and R_{FN}. Thus, to identify the joint distribution, two additional assumptions are made. The first assumption is that $\pi_{00} = \pi_{000}$, so that $\pi_{001} = 0$. Intuitively, individuals who do not self-report a condition and do not test positively for the condition by medical examination are assumed to have no medical record of the condition. The second assumption is that $\pi_{11} = \pi_{111}$, so that $\pi_{110} = 0$. Intuitively, individuals who self-report a condition and test positively for the condition by medical examination are assumed to have a medical record of the condition. Both assumptions rely on the medical examination (E) to validate self-reported health (S), which implies whether a medical record (R) should or should not exist.

With six estimates and two assumptions, the eight population moments π_{SER} are identified. Details of the calculation and estimation are provided in the Appendix.

The identification strategy requires the rates R_{FN} and R_{FP} to be the same between the NHANES and the NPHS/OHIP. For this assumption to be credible, it is important that the data are comparable. The CNPHS/OHIP data come from years 1996/1997. Thus, the analysis uses

NHANES data from calendar years 1999, the first year of data, to 2003. The sample in Baker, Stabile, and Deri (2004) is restricted to individuals who are aged 16 and not attending school. The NHANES is similarly restricted. The NHANES oversamples certain groups, so all estimations use sample weights.

An obvious concern is that the NHANES is representative of the US, whereas the CNPHS and OHIP are representative of Ontario. While not ideal, the few US-based studies that validate self-reported health conditions using medical records (Harlow and Linet 1989) are limited in scope. For example, Martin et al (2000) focus on enrollees of a single insurance firm, and medical records come from claims within the firm.

Another concern pertains to the survey questions of health conditions. In the NHANES, survey participants are asked, "[Have you] ever been told by a doctor or other health professional that [you have] [this condition]?" In the CNPHS, survey participants are asked, "Do [you] have any of the following long-term conditions that have been diagnosed by a healthcare professional?".[7] While both questions ask about health conditions diagnosed by medical professionals, the question in the CNPHS may be interpreted in the present tense, whereas the question in the NHANES may be interpreted in past tense. This difference may result in lower prevalence rates in the CNPHS, which would result in a higher R_{FN} and lower R_{FP} relative to the NHANES.

Given the available data, the analysis focuses on two health conditions: hypertension and diabetes. In the NHANES, survey participants are first asked whether they have been

[7] "Healthcare professional" is defined to exclude alternative healthcare providers, such as acupuncturists, and "long-term" is defined as a condition that is expected to last six months or more.

previously diagnosed for hypertension and diabetes, and then are tested for these conditions by medical examination. For hypertension, the self-report variable S equals one if the survey participant had been diagnosed at least twice for hypertension by a medical professional, and the medical exam variable E equals one if the survey participant tests positive for hypertension based the on the average of up four blood pressure readings.[8] For diabetes, the self-report variable S equals one if the survey participant had been diagnosed for diabetes, and the medical exam variable E equals one if the survey participant tests positive for diabetes based on a test of fasting plasma glucose.[9] In regards to R_{FN} and R_{FP}, Baker, Stabile, and Deri (2003) report multiple estimates based on various specifications of the medical record. The specification used in this study requires at least two OHIP records for a specific condition during the two years prior to the survey.

III. Results

A. Estimates of Joint Distribution: π_{SER}

Table 1 presents estimates of R_{FN}, R_{FP}, and π_{SER} for hypertension and diabetes. The first two columns report estimates of R_{FN} and R_{FP}, derived from Baker, Stabile, and Deri (2004). For both conditions, the rate of false-negative reporting ranges between 20 and 30 percent. This suggests that many people who have a medical record for a condition, and thus may test positive for the condition by medical examination, may fail to report the condition

[8] A diagnosis of hypertension is based on blood pressure readings of both systolic and diastolic pressure. Hypertension is defined as systolic greater than or equal to 140 mm Hg or diastolic greater than or equal to 90 mm Hg.

[9] The test for fasting plasma glucose is administered to only half of the sample. Diabetes is defined as fasting plasma glucose greater than or equal to 126 mg/dl.

nonetheless. The rate of false-positive reporting is much lower, ranging from 1 to 6 percent. This suggests that few individuals falsely claim or self-diagnose a condition.

The next four columns report estimates for π_{11}, π_{00}, π_{10}, and π_{01}. These estimates are derived solely from the NHANES. As shown, only 40.3 percent of those who self-report hypertension actually test positive for hypertension by medical examination $(\pi_{11}/(\pi_{11}+\pi_{10}))$. Conversely, only 47.3 percent of those who test positive for hypertension by medical examination actually self-report hypertension $\left(\pi_{11}/(\pi_{11}+\pi_{01})\right)$. These figures for diabetes are 66.5 percent and 60.1 percent, respectively. Based solely on π_{01}, the prevalence of undiagnosed hypertension and diabetes is 9.0 percent and 2.7 percent, respectively.

The final four columns disaggregate π_{10} and π_{01} into those with and without a medical record. In regards to π_{01}, the empirical question is whether individuals who test positive for the condition, but fail to self-report it, have a medical record for the condition nonetheless. As shown, an estimated 72.8 percent of individuals who test positive for hypertension, but fail to report it, have a medical record for hypertension $\left(\pi_{011}/(\pi_{010}+\pi_{011})\right)$. This figure for diabetes is 50.1 percent. If π_{011} should be excluded from estimates of undiagnosed conditions, then the prevalence of hypertension is closer to 2.4 percent (π_{010}) than 9.0 percent (π_{01}), and the prevalence of diabetes is closer to 1.3 percent than 2.7 percent.

In regards to π_{10}, the empirical question is whether individuals who self-report a condition, but do not test positive for the condition by medical examination, have a medical record for the condition. As shown, an estimated 62.2 percent of individuals who self-report hypertension, but do not test positive for hypertension by medical examination, have a medical record $\left(\pi_{101}/(\pi_{100}+\pi_{101})\right)$. This figure for diabetes is 48.4 percent. If π_{100} should be

excluded from estimates of diagnosed conditions, then the prevalence of hypertension is closer to 15.5 percent ($\pi_{11} + \pi_{101}$) than 20.0 percent ($\pi_{100} + \pi_{10}$), and the prevalence of diabetes is closer to 5.0 percent than 6.0 percent.

B. Estimates of Proportional Bias: $\boldsymbol{cov(S, u)/var(S)}$

Stated above, the model of measurement error helps to characterize the biases that may arise when using self-reported health conditions as explanatory variables. Specifically, if β_1 is the causal effect of true health S^* on outcome Y, and if S^* is replaced with self-reported health S, then the estimate of β_1 converges in probability to $\beta_1[1 - cov(S, u)/var(S)]$. Thus, the bias due to measurement error is proportional to $cov(S, u)/var(S)$.

Given estimates π_{SER}, the proportional bias term $cov(S, u)/var(S)$ is estimated for various definitions of true health S^*. These estimates are presented in Table 2. The calculations for hypertension are reported in the first panel, and the calculations for diabetes are reported in the second panel. Each panel contains four rows, corresponding to different definitions of S^*.

In the first row, S^* is defined by the medical record only. In this case, measurement error u equals $S - R$. Based on the estimates of π_{SER}, the next four columns report estimates of the mean of S, variance of S, mean of u, and covariance of S and u. The final column reports the proportional bias. As shown, the bias is 0.308 for hypertension and 0.187 for diabetes. These estimates are similar to Baker, Stabile, and Deri (2003), who estimate a proportional bias of 0.355 for hypertension and 0.195 for diabetes.

In the second row, S^* is defined by the medical examination only. In this case, measurement error u equals $S - E$. As shown, the proportional bias is considerably greater,

reaching 0.710 for hypertension and 0.363 for diabetes. The estimate for hypertension is similar to the estimate by Johnston, Propper, and Shields (2009), who use survey data matched to medical examinations from the Health Survey for England.[10] They estimate a proportional bias for hypertension of 0.68.

In the third and fourth rows, S^* is defined by a combination of the medical record and the medical examination. In the third row, true health is defined by either the medical record or the examination; in the fourth row, true health requires both a medical record and a positive result by medical examination. As shown, the estimates of proportional bias fall between the estimates in the first and second rows.

Thus, the proportional bias is smallest when S^* is defined by the medical record and largest when S^* is defined by the medical examination. The results reflect that, in the former case, only two sources of measurement error exist: π_{100} and π_{011}. However, in the latter case, two additional sources of error arise: π_{101} and π_{010}. These two additional sources of error necessarily increase the proportional bias term $cov(S,u)/var(S)$.

C. Sensitivity to Assumptions $\boldsymbol{\pi_{00} = \pi_{000}}$ and $\boldsymbol{\pi_{11} = \pi_{111}}$

To identify the joint distribution π_{SER}, it was assumed that $\pi_{00} = \pi_{000}$ and $\pi_{11} = \pi_{111}$, which imply that $\pi_{001} = 0$ and $\pi_{110} = 0$, respectively. Although both assumptions are reasonable, an important question is whether the estimates of π_{SER} are sensitive to these assumptions. To relax these assumptions, it is assumed that a share γ of π_{00} instead has a medical record, so $\pi_{001} = \gamma\pi_{00}$ and $\pi_{000} = (1-\gamma)\pi_{00}$. Similarly, it is assumed a share δ of

[10] A similar analysis is conducted for arthritis by Butler, Burkhauser, Mitchell, and Pincus (1987).

π_{11} instead does not have a medical record, so $\pi_{110} = \delta\pi_{11}$ and $\pi_{111} = (1-\delta)\pi_{11}$. In the baseline results presented above, the shares γ and δ were assumed zero.

In this case, the rate of false-negative reporting is expressed as,

$R_{FN} = \frac{\gamma\pi_{00}+\pi_{011}}{\gamma\pi_{00}+(1-\delta)\pi_{11}+\pi_{101}+\pi_{011}}$.

Similarly, the rate of false-positive reporting is expressed as,

$R_{FP} = \frac{\delta\pi_{11}+\pi_{100}}{(1-\gamma)\pi_{00}+\delta\pi_{11}+\pi_{100}+\pi_{010}}$.

The empirical question is whether the estimates of π_{SER} differ for various values of γ and δ.

The calculations of π_{SER}, described in the Appendix, yield two findings. First, the estimates of π_{011} and π_{010} depend only on γ, not δ. Stated above, π_{01} has been interpreted as the prevalence of undiagnosed health conditions, and π_{011} is the prevalence of these conditions that had indeed been diagnosed, according to the medical record. The finding suggests that the disaggregation of π_{01} into those with and without a medical record does not require $\delta = 0$ ($\pi_{110} = 0$).

The sensitivity of π_{011} and π_{010} to different values of γ is given in panel A of Figures 1 and 2. The estimates for hypertension are presented in Figure 1, and the estimates for diabetes are presented in Figure 2. Each panel graphs the estimates of π_{011} and π_{010} based on the value of γ, ranging from 0 to 0.05.

As shown, the estimates of π_{011} and π_{010} are sensitive to the value of γ. In regards to hypertension, the estimate of π_{011} decreases from 6.5 percent when γ equals zero to 3.0 percent when γ equals 0.05. As a result, the share of π_{01} with a medical record decreases from 72.85 percent to 33.7 percent (π_{011}/π_{01}). The estimates of for diabetes appear more

sensitive. As shown, the estimate of π_{011} decreases quickly from 1.3 percent when γ equals zero to zero when γ reaches approximately .015.

The second finding is that the estimates of π_{101} and π_{100} depend only on δ, not γ. The term π_{10} pertains to individuals who self-report a condition, but do not test positive for the condition by medical examination. The term π_{101} includes individuals who had been diagnosed for the condition by a medical professional, accurately self-report the diagnosis during the survey, but perhaps recovered from the condition at the time of the survey. The finding suggests that the disaggregation of π_{10} into those with and without a medical record does not require $\gamma = 0$ ($\pi_{001} = 0$).

The sensitivity of π_{011} and π_{010} to different values of δ is given in panel B of Figures 1 and 2. As shown, the estimates of π_{011} and π_{010} appear less sensitive to the value of δ. In regards to hypertension, the estimate of π_{101} increases from 7.4 percent when δ equals zero to 7.8 percent when δ equals 0.05. As a result, the share of π_{10} with a medical record increases from 62.2 percent to 65.5 percent (π_{101}/π_{10}). The estimates of for diabetes also appear less sensitive. As shown, the estimate of π_{011} increases from 1.0 percent when δ equals zero to 1.2 percent when δ reaches 0.05.

IV. Discussion and Conclusion

This study develops and estimates a model of measurement error in self-reported health conditions. The model allows self-reports of a health condition to differ from a contemporaneous medical examination, prior medical records, or both. The model is estimated using a two-sample strategy, which combines survey data linked medical examination results

and survey data linked to prior medical records. The study finds substantial inconsistencies between self-reported health, the medical record, and prior medical records.

The study has three empirical applications. First, the study provides an alternative estimator of undiagnosed health conditions. Several studies define undiagnosed conditions as those that are not self-reported at the time of the survey, but are detected upon medical examination. The alternative estimator excludes those that are not self-reported, but are included in the medical record nonetheless. Using the alternative estimator, the prevalence of undiagnosed hypertension decreases from 9.0 percent to 2.4 percent, and the prevalence of undiagnosed diabetes decreases from 2.7 percent to 1.3 percent.

Second, the study provides an alternative estimator of diagnosed conditions. Several studies define diagnosed conditions as those that are self-reported, regardless of whether it is detected upon medical examination. The alternative estimator excludes those that are self-reported, are not reported upon medical examination, and not included in the medical record. Using this alternative estimator, the prevalence of diagnosed hypertension decreases from 20.0 percent to 15.5 percent, and the prevalence of diagnosed diabetes decreases from 6.0 percent to 5.0 percent.

Finally, the study examines the bias that may arise when estimating the causal effect of health conditions on other outcomes of interest. In a simplified model, the proportional bias is greatest when true health is defined by the medical examination only. In this case, the proportional bias is 0.710 for hypertension and 0.363 for diabetes.

Using a two-sample strategy, this study is the first reasonable attempt to estimate measurement error in self-reported health conditions relative to both the medical record and

medical examination. However, the estimation strategy has two notable limitations. First, this study utilizes data from Canada and the US, which raises concerns about the comparability of the two populations. Second, the estimation strategy requires assumptions about specific population parameters – specifically $\pi_{001} = 0$ and $\pi_{110} = 0$. While these assumptions are reasonable, it is imperative to test these assumptions empirically. These limitations can be addressed using survey data linked to both medical examination results and medical records, once such data become available.

Appendix

The empirical objective is to estimate the joint probability distribution of three binary variables: an indicator of a self-reported health condition, an indicator of the result from a medical examination, and an indicator of the medical record. The joint probability distribution is denoted π_{SER}, where the three subscripts correspond to the three binary variables.

The joint distribution is estimated from two separate data linkages. The first is survey data linked to results from a medical examination. These data provide estimates of π_{11}, π_{00}, π_{10}, and π_{01}. The second is survey data linked to medical records. These data provide estimates of rates of false-negative and false-positive reporting, given by

$$R_{FN} = \frac{\pi_{001}+\pi_{011}}{\pi_{001}+\pi_{111}+\pi_{101}+\pi_{011}},$$

and

$$R_{FP} = \frac{\pi_{110}+\pi_{100}}{\pi_{000}+\pi_{110}+\pi_{100}+\pi_{010}},$$

respectively.

To identify the system, two additional assumptions are made. The first assumption is $\pi_{00} = \pi_{000}$, so that $\pi_{001} = 0$. The second assumption is $\pi_{11} = \pi_{111}$, so that $\pi_{110} = 0$. Both assumptions rely on the medical examination (E) to validate self-reported health (S), which implies whether a medical record (R) should or should not exist.

With six estimates and two assumptions, the system is identified. Specifically,

$$\pi_{011} = \frac{\tilde{R}_{FN}(\pi_{11}+\pi_{10})-\tilde{R}_{FN}\tilde{R}_{FP}(\pi_{00}+\pi_{01})}{1-\tilde{R}_{FN}\tilde{R}_{FP}},$$

and

$$\pi_{100} = \frac{\tilde{R}_{FP}(\pi_{00}+\pi_{01})-\tilde{R}_{FN}\tilde{R}_{FP}(\pi_{11}+\pi_{10})}{1-\tilde{R}_{FN}\tilde{R}_{FP}}.$$

Additionally, $\pi_{010} = \pi_{01} - \pi_{011}$ and $\pi_{101} = \pi_{10} - \pi_{100}$.

To evaluate the sensitivity of the estimates to the assumptions that $\pi_{00} = \pi_{000}$ and $\pi_{11} = \pi_{111}$, it is assumed that a share γ of π_{00} has a medical record and that a share δ of π_{11} does not have a medical record. In this case, the rates of false-negative and false-positive reporting are given by

$$R_{FN} = \frac{\gamma\pi_{00}+\pi_{011}}{\gamma\pi_{00}+(1-\delta)\pi_{11}+\pi_{101}+\pi_{011}},$$

And

$$R_{FP} = \frac{\delta\pi_{11}+\pi_{100}}{(1-\gamma)\pi_{00}+\delta\pi_{11}+\pi_{100}+\pi_{010}},$$

respectively.

In this case, the estimate of π_{011} depends only on γ, not δ, and the estimate of π_{100} depends only on δ, not γ. Specifically,

$$\pi_{011} = \frac{\tilde{R}_{FN}[\gamma\pi_{00}+\pi_{11}+\pi_{10}]-\tilde{R}_{FN}\tilde{R}_{FP}[(1-\gamma)\pi_{00}+\pi_{01}]-\left(\frac{1}{1-R_{FN}}\right)\gamma\pi_{00}}{1-\tilde{R}_{FN}\tilde{R}_{FP}},$$

And,

$$\pi_{100} = \frac{\tilde{R}_{FP}[\pi_{00}+\delta\pi_{11}+\pi_{01}]-\tilde{R}_{FN}\tilde{R}_{FP}[(1-\delta)\pi_{11}+\pi_{10}]-\left(\frac{1}{1-R_{FP}}\right)\delta\pi_{11}}{1-\tilde{R}_{FN}\tilde{R}_{FP}}.$$

References

Baker, Michael, Mark Stabile, and Catherine Deri. 2004. "What Do Self-Reported, Objective Measurof Health Measure?" *Journal of Human Resources* 39(4): 1067-1093.

Benitez-Silva, Hugo, Moshe Buchinsky, Hui Man Chan, Sofia Cheidvasser, and John Rust. 2004. "How Large is the Bias in Self-Reported Disability?" *Journal of Applied Econometrics* 19: 649-670.

Bound, John. 1991. "Self-Reported versus Objective Measures of Health in Retirement Models." *Journal of Human Resources* 26(1): 106-138.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In J. Heckman and E. Leamer (eds.) *Handbook of Econometrics* volume 5.

Butler, J.S., Richard Burkhauser, Jean Mitchell, and Theodore Pincus, 1987. "Measurement Error in Self-Reported Health Variables." Review of Economics and Statistics 69(4): 644-650.

Cawley, John and Anna Choi, 2015. "Health Disparities Across Education: The Role of Differential Reporting Error." *NBER Working Paper #21317.*

Cowie, Catherine, Keith Rust, Danita Byrd-Hold, Mark Eberhardt, Katherine Flegal, Michael Engelgau, Sharon Saydah, Desmond Williams, Linda Geiss, and Edward Gregg. 2006. "Prevalence of Diabetes and Impaired Fasting Glucose in Adults in the U.S. Population: National Health and Nutrition Examination Survey 1999-2002." *Diabetes Care* 29(6): 1263-1268.

Currie, Janet and Brigitte Madrian. 1999. "Health, Health Insurance and the Labor Market." In O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics* edition 1, volume 3, number 3.

Dwyer, Debra and Olivia Mitchell. 1999. "Health Problems as Determinants of Retirement: Are Self-Rated Measures Endogenous?" *Journal of Health Economics* 18: 173-193.

Harlow, Sioban, and Martha Linet. 1989. "Agreement between Questionnaire Data and Medical Records." *American Journal of Epidemiology* 129(1): 233-48.

Johnston, David, Carol Propper, and Michael Shields. 2009. "Comparing Subjective and Objective Measures of Health: Evidence from Hypertension for the Income/Health Gradient." *Journal of Health Economics* 28(3): 540-552.

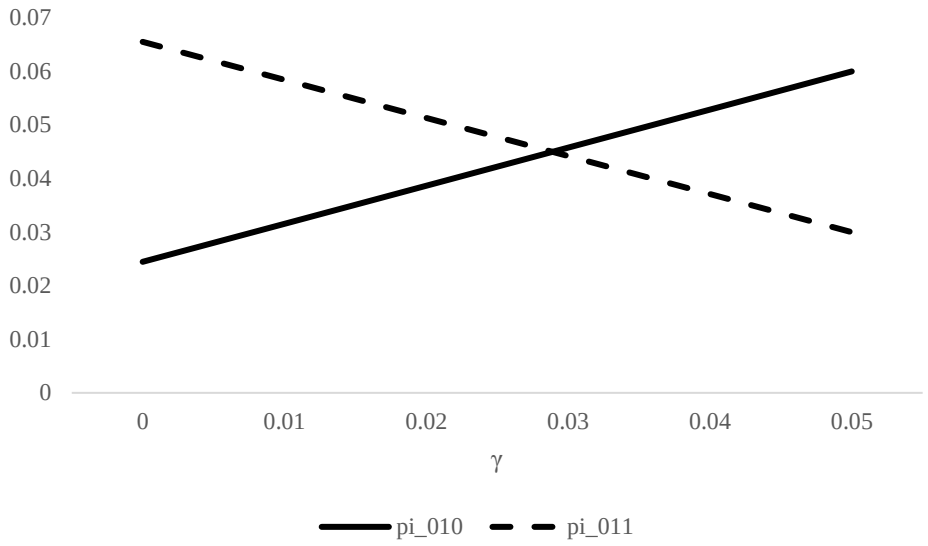
Meyer, Bruce, Wallace Mok, and James Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economics Perspectives* 29(4): 199-226.

Madow, William. 1973. "Net Differences in Interview Data on Chronic Conditions and Information Derived from Medical Records." *National Center for Health Statistics* Series 2, No. 23.

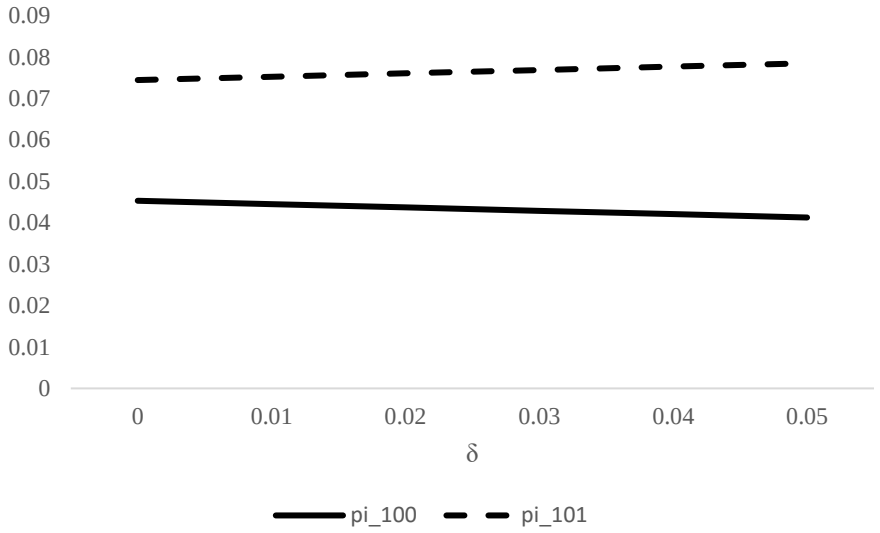
Martin, Linda, Marilyn Leff, Ned Calonge, Carol Garrett, and David Nelson. 2000. "Validation of Self-Reported Chronic Conditions and Health Services in a Managed Care Population." *American Journal of Preventive Medicine* 18(3): 215-218.

Suziedelyte, Agne and Meliyanni Johar. 2013. "Can you trust survey responses? Evidence Using Objective Health Measures." *Economics Letters* 121: 163-166.

Yoon, Sung Sug, Vicki Burt, Tatiana Louis, and Margaret Carroll. 2012. "Hypertension Among Adults in the United States, 2009 -2010." NCHS Data Brief No. 107.



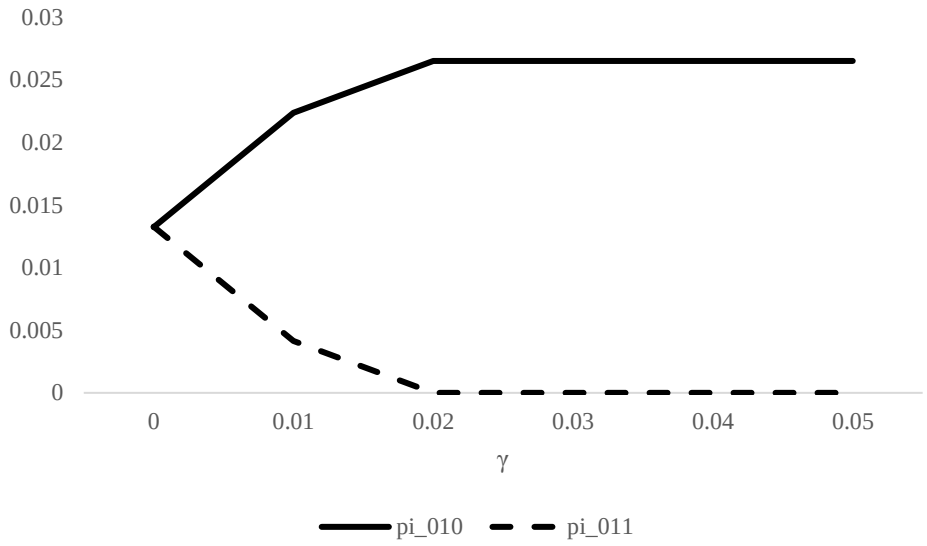
Panel A: Sensitivity of π_{010} and π_{010} to γ



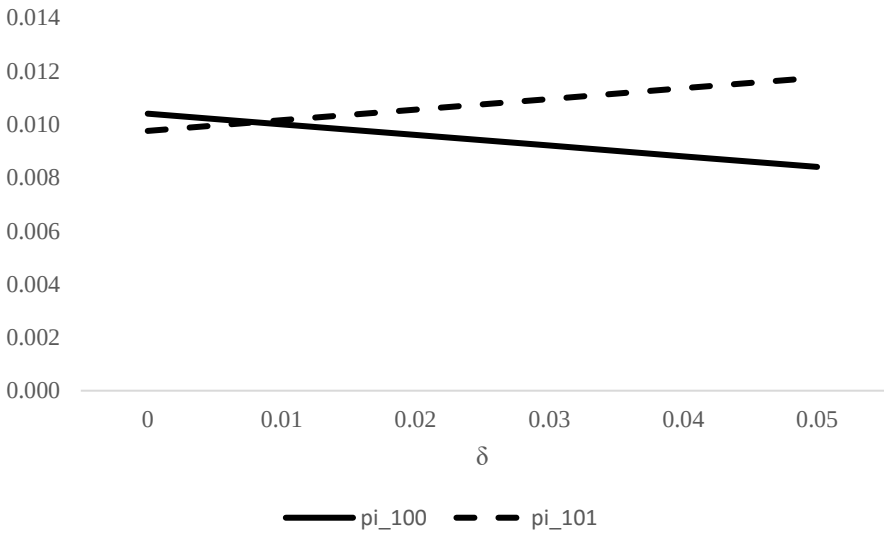
Panel B: Sensitivity of π_{100} and π_{101} to δ

Figure 1: Sensitivity of π_{SER} for Hypertension

The figure illustrates the sensitivity of π_{SER} to values of γ and δ. These terms are related according to the equations: $\pi_{001} = \gamma\pi_{00}$, $\pi_{000} = (1-\gamma)\pi_{00}$, $\pi_{110} = \delta\pi_{11}$ and $\pi_{111} = (1-\delta)\pi_{11}$. The first subscript of π indicates the value of the self-report; the second subscript indicates the value of the medical exam result; the third subscript indicates the value of the medical record.



Panel A: Sensitivity of π_{010} and π_{010} to γ



Panel B: Sensitivity of π_{100} and π_{101} to δ

Figure 2: Sensitivity of π_{SER} for Diabetes

The figure illustrates the sensitivity of π_{SER} to values of γ and δ. These terms are related according to the equations: $\pi_{001} = \gamma\pi_{00}$, $\pi_{000} = (1-\gamma)\pi_{00}$, $\pi_{110} = \delta\pi_{11}$ and $\pi_{111} = (1-\delta)\pi_{11}$. The first subscript of π indicates the value of the self-report; the second subscript indicates the value of the medical exam result; the third subscript indicates the value of the medical record.

Table 1

Joint Distribution of Self-Report, Medical Exam, and Medical Record

			Self-Report/Exam Result/Medical Record							
	R_{FN}	R_{FP}	π_{11}	π_{00}	π_{10}	π_{01}	π_{100}	π_{101}	π_{010}	π_{011}
Hypertension	0.297	0.058	0.081	0.710	0.120	0.090	0.045	0.074	0.024	0.065
			(0.002)	(0.004)	(0.003)	(0.002)	(0.000)	(0.004)	(0.006)	(0.002)
Diabetes	0.211	0.011	0.040	0.913	0.020	0.027	0.010	0.010	0.013	0.013
			(0.002)	(0.003)	(0.002)	(0.002)	(0.000)	(0.002)	(0.005)	(0.001)

The table presents estimates of the joint distribution of three binary variables: an indicator of the self-report, an indicator of the medical examination result, and an indicator of the medical record. The joint distribution is characterized by π_{SER}, where the subscripts correspond to the three binary variables. The terms R_{FN} and R_{FP} are rates of false-negative and false-positive reporting, relative to the medical record. These values are derived from Baker, Stabile, and Deri (2004). The estimates of π_{11}, π_{00}, π_{10}, and π_{01} are derived the National Health and Nutrition Examination Survey.

Table 2

Proportional Bias of by Definition of True

Hypertension	True health: S^*	Error: u	$E(S)$	$Var(S)$	$E(u)$	$cov(S,u)$	Proportional Bias
	Record	S-R	0.200	0.160	-0.020	0.049	0.308
	Exam	S-E	0.200	0.160	0.030	0.114	0.710
	Either	S-max(E,R)	0.200	0.160	-0.045	0.054	0.338
	Both	S-min(E,R)	0.200	0.160	0.120	0.096	0.597

Diabetes	True health: S^*	Error: u	$E(S)$	$Var(S)$	$E(u)$	$cov(S,u)$	Proportional Bias
	Record	S-R	0.060	0.057	-0.003	0.011	0.187
	Exam	S-E	0.060	0.057	-0.006	0.021	0.363
	Either	S-max(E,R)	0.060	0.057	-0.016	0.011	0.201
	Both	S-min(E,R)	0.060	0.057	0.020	0.019	0.335

The table presents estimates of the proportional bias of β_1 when true health S^* is replaced with self-reported health S. The term β_1 is the causal effect of true health S^* on an outcome variable Y in the following linear model: $Y = \beta_0 + \beta_1 S^* + \varepsilon$. The proportional bias is estimated for four definitions of true health S^* using estimates of the joint distribution π_{SER}.