 DOI: 10.1111/j.1541-0420.2006.00665.x

Robustness of Prevalence Estimates Derived from Misclassified Data from Administrative Databases

Martin Ladouceur,[1,2] Elham Rahme,[3] Christian A. Pineau,[4] and Lawrence Joseph[1,2,*]

[1]Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada

[2]Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Quebec, H3A 1A2, Canada

[3]Department of Medicine, and Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada

[4]Division of Rheumatology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada

[*]*email:* Lawrence.Joseph@mcgill.ca

Summary. Because primary data collection can be expensive, researchers are increasingly using information collected in medical administrative databases for scientific purposes. This information, however, is typically collected for reasons other than research, and many such databases have been shown to contain substantial proportions of misclassification errors. For example, many administrative databases contain fields for patient diagnostic codes, but these are often missing or inaccurate, in part because physician reimbursement schemes depend on medical acts performed rather than any diagnosis. Errors in ascertaining which individuals have a given disease biases not only prevalence estimates, but also estimates of associations between the disease and other variables, such as medication use. We attempt to estimate the prevalence of osteoarthritis (OA) among elderly Quebeckers using a government administrative database. We compare a naive estimate relying solely on the physician diagnoses of OA listed in the database, to estimates from several different Bayesian latent class models, which adjust for misclassified physician diagnostic codes via use of other available diagnostic clues. We find that the prevalence estimates vary widely, depending on the model used and assumptions made. We conclude that any inferences from these databases need to be interpreted with great caution, until further work estimating the reliability of database items is carried out.

Key words: Administrative databases; Bayesian latent class models; Diagnosis; Misclassification; Prevalence; Robustness; Sensitivity; Specificity.

1. Introduction

Medical administrative databases are increasingly being used for research purposes. As one example, Gabriel, Crowson, and O'Fallon (1995) estimated the prevalence of osteoarthritis (OA) using an administrative database, where OA status for each individual was ascertained from a physician diagnosis variable. Accurate disease ascertainment is important not only for prevalence estimates, useful for planning purposes or tracking changes in prevalence over time, but also for etiologic research, such as determining whether use of a given medication may be associated with the disease. Many databases claim near universal coverage of the target population, and while most are known to contain imperfect and incomplete information (Green and Wintfeld, 1993; Swerdlow, Douglas, and Vaughn, 1993; Rushton and Romaniuk, 1997; Quan, Parsons, and Ghali, 2004; Wilchesky, Tamblyn, and Huang, 2004), they are usually assumed to be sufficiently accurate for reasonable estimation. Nevertheless, it is not difficult to demonstrate that unadjusted inferences from imperfect database information can sometimes lead to severely biased estimates.

Statistical methods have been developed in the diagnostic testing literature, which can adjust prevalence estimates for misclassified data, including works by Walter and Irwig (1988), Joseph, Gyorkos, and Coupal (1995), Alonzo and Pepe (1999), Johnson, Gastwirth, and Pearson (2001), Black and Craig (2002), McInturff et al. (2004), Gustafson (2005), and many others. See Pepe (2003) and Gustafson (2004) for recent reviews. To our knowledge, these models have not been applied to administrative database studies, nor has the robustness of estimates from these methods been evaluated across a range of reasonable models. This is especially important in this context, because all models rely on difficult to verify assumptions and/or prior information.

In this article, we estimate the prevalence of OA using a government administrative database. We compare estimates across a variety of Bayesian latent class models, each using a different set of assumptions, to the naive estimator (unadjusted physician diagnoses) used by most researchers. We show that considerable uncertainty can surround prevalence estimates from administrative databases, even when models

that adjust for misclassification are used. Therefore, it is important for researchers to acknowledge that information from a database alone, however convenient, may not always be able to provide reliable estimates of some parameters. This is true even if a database is large, has little selection bias, and reasonable statistical methods provide inferences that are adjusted for potential database errors.

We begin in Section 2 by describing the database we will use and the three clues found in the database, each one an imperfect indicator of the presence of OA, that we use to ascertain OA status. We review various Bayesian latent class models for estimating prevalence from misclassified data in Section 3, and use these models to provide OA prevalence estimates in Section 4. Section 5 presents concluding remarks, including some suggestions for study designs based on further data collection, which can help to minimize the problems we discuss.

2. Indicators of Osteoarthritis from an Administrative Database

Assessing OA status in an individual is difficult, in part because of a lack of a definitive definition of the disease (Badley and Webster, 1995; Lawrence and Helmich, 1998). OA is often diagnosed on the basis of radiography, but many people with radiographic evidence of OA have no symptoms or disability, so it is unclear whether such persons should be considered as having OA. In this article, we define OA as present if, using the best available knowledge, a physician would classify the subject as OA positive. Thus, the target parameter of our investigation, the prevalence of OA in Quebeckers aged 65 years and over, is defined as the proportion of elderly Quebeckers who would be classified as OA positive by their physicians.

The Quebec government's Régie de l'Assurance Maladie du Québec (RAMQ) database records include physician diagnoses using the International Classification of Diseases 9th Edition (ICD-9) code, specialty of the billing physician, and medical procedure codes. Pharmacy claims records includes the drug identification number, drug dosage, dispensing date, number of days supplied, and prescribing physician specialty. Coverage is universal for those 65 years and older. We obtained Quebec-wide RAMQ data on all people in this age group in 2002 who were positive for at least one of our three tests, which we now describe in detail.

Our first test is classified as positive for anyone who received an ICD-9 diagnostic code for OA at least once during the year 2002, and is otherwise considered as negative. Positive subjects can have diagnostic codes for more than one condition, as long as at least one is for OA. Most researchers (e.g., Gabriel et al., 1995) have used this variable only, without adjustment for its imperfections, to ascertain disease status in database studies.

Our second test is positive for those who filled at least one prescription for acetaminophen or a nonsteroidal antiinflammatory drug (NSAID), but not methotrexate or plaquenil, medications specific to rheumatoid arthritis. Individuals with other prescriptions and those who did not fill any prescriptions are considered as test two negative. Since aspirin is frequently prescribed for its antiplaquetary effect, it was not included as an NSAID here.

Our third test is considered as positive if an individual received an injection procedure common to OA patients, an arthroplasty or a tibial osteotomy during 2002, even if they also received other surgical procedures. All other subjects are test three negative.

Subjects over age 65 years in 2002 and not listed in the RAMQ database as testing positive for any of the above three tests were considered as negative for all tests.

All of the above tests have serious limitations that virtually guarantee high rates of misclassification errors. While our definition assumes that physicians are able to diagnose OA patients when the disease is in fact present, many false positive and false negative cases are expected for our first test because of missing diagnostic codes, or early OA-like symptoms that can arise from conditions other than OA. A recent study (Wilchesky et al., 2004) comparing patient charts of general practitioners' to RAMQ records found that up to 30% of diagnostic records are missing from this database. Therefore, while many researchers (e.g., Gabriel et al., 1995; Robertson, Svenson, and Joffres, 1998; Romano, Schembri, and Rainwater, 2002; Papaioannou et al., 2003) have used unadjusted administrative database diagnostic codes to estimate prevalence, these estimates can be severely biased.

NSAIDs are the most commonly used medications for musculoskeletal disorders (Berger, 1994). Acetaminophen is recommended by both American and Canadian guidelines as a first line therapy for OA (Wegman et al., 2004). Both acetaminophen and NSAIDs are also used for relief of pain in other conditions such as headaches, back pain, and various injuries. Therefore, considering a subject who has taken chronic (30 or more days per year) OA medication while excluding those who have taken medications specific to other chronic musculoskeletal conditions (plaquenil and methotrexate, the most commonly prescribed disease modifying antirheumatic drugs) will lead to higher (although still imperfect) specificity. The RAMQ prescription claims database provides a reasonably accurate measure of drug exposure in the elderly population (Tamblyn, 1995), but imperfections still arise. For example, not all written prescriptions are filled, and some drugs are available without prescription. Substantial numbers of false positive and false negative cases are, therefore, expected for test two.

Arthroplasty (Thompson, 2001) and tibial osteotomy (Grelsamer, 1995) are both surgeries carried out to treat severe OA. Test three, therefore, will capture only the more severe cases of OA while missing milder cases, so that we expect to see many false negatives but fewer false positives compared to the first two tests.

The above "diagnostic test" definitions are of course not unique. Variations on the above criteria can be useful in assessing robustness of prevalence estimates across changes in test definitions. We investigated two main changes: for test one, we required two or more diagnoses of OA in 2002, rather than just one. Similarly, for test two, we required that an acetaminophen or an NSAID be prescribed at least twice in 2002, with total days supplied exceeding 30 days. In both cases, stricter criteria are expected to increase specificity at the expense of sensitivity.

It is obvious that none of our "tests" are perfect indicators of OA. In particular, it is clear that naive use of physician

diagnosis alone will likely provide a biased prevalence estimate. The question is: by combining the information from all three tests and using statistical methodology that adjusts for known database imperfections, can reliable prevalence estimates be derived?

3. Adjusting for Misclassification: Identifiable and Nonidentifiable Latent Class Models

We now discuss the inputs and assumptions required for each of the several related models. While these models are not new, they have not been previously applied to database studies, nor have prevalence estimates been compared across different models.

Let $D+$ and $D-$ denote true disease status, "diseased" and "non-diseased," respectively. Similarly, let $T+$ and $T-$ represent test positive and test negative outcomes on a given test. In the absence of a perfectly accurate test, estimating the prevalence θ will depend on the test characteristics, particularly the sensitivity $S = P(T+ \mid D+)$ and specificity $C = P(T- \mid D-)$, where $P(A \mid B)$ denotes the conditional probability of A given B. The probability of testing positive is the sum of the probabilities of being a true positive and the probability of being a false positive, that is, $P(T+) = p = \theta S + (1-\theta)(1-C)$. In the single test situation when S and C are exactly known, algebraically solving for θ, an expression for the prevalence, adjusted for imperfect sensitivity and specificity, is given by $\theta = \frac{p-(1-C)}{S+C-1}$. Note that in the case that $S = C = 1$ then $\theta = p$, and no adjustment is needed.

Test properties S and C are not usually exactly known, and so have to be estimated along with θ. This presents a nonidentifiable estimation problem, as three unknown parameters (θ, S, and C) must be estimated, but the dichotomous data of test positive and test negative subjects provide only one degree of freedom (Walter and Irwig, 1988; Joseph et al., 1995; Gustafson, 2005). Formally, a model expressed via the density function $f(x \mid \theta)$ is identifiable if $f(x \mid \theta_1) = f(x \mid \theta_2)$ for all x implies $\theta_1 = \theta_2$. In nonidentifiable problems, it is clear that the data alone cannot provide consistent estimators of θ. Estimates can nevertheless be derived via Bayesian methods, where prior information can separate out the likelihood of θ_1 from that of θ_2 when the data cannot distinguish between these values. In our problem, the prevalence estimate can be formed by averaging the adjusted estimate given above, $\theta = \frac{p-(1-C)}{S+C-1}$, over the prior distributions of S and C. Thus, the estimate depends not only on the data leading to p, but also on the prior distributions of the test properties. Of course, any estimates are then only as reliable as the prior distributions used, and these, in turn, can be quite difficult to ascertain due to the complex way in which most administrative databases are created.

When data for a single test are available, that is, when x positive tests are observed in n subjects, the likelihood function for (θ, S, C) is proportional to

$$\begin{aligned} l(x \mid \theta, S, C) &\propto p^x(1-p)^{n-x} \\ &= \{\theta S + (1-\theta)(1-C)\}^x \{\theta(1-S) + (1-\theta)C\}^{n-x}. \end{aligned}$$

From Bayes theorem, if the joint prior distribution of θ, S, and C is given by $f(\theta, S, C)$, the joint posterior density becomes

$$f(\theta, S, C \mid x) \propto f(\theta, S, C) l(x \mid \theta, S, C).$$

Posterior inferences can then either be derived through the Gibbs sampler (Joseph et al., 1995) or the SIR algorithm (Rahme, Joseph and Gyorkos, 2000).

On the other hand, when using three conditionally independent tests (Demissie et al. 1998), the problem is identifiable, as there are seven degrees of freedom (from the eight possible outcomes) from which to estimate the seven unknown parameters (S and C from each of three tests, and θ). In practice, this means that estimates of all parameters can be found by maximum likelihood or Bayesian methods with noninformative prior distributions, avoiding the reliance on finding good prior estimates of test properties. Unlike the one test situation, however, this method relies on the difficult to verify assumption of conditional independence between the three tests. Under conditional independence, the tests are assumed to be statistically independent of each other, conditional on the true disease status of the subject. This assumption may not hold, for example, because the records for the diagnosis of OA and any prescriptions will sometimes be derived from the same visit to a single physician.

To see how the likelihood function from the three test model is constructed, consider Table 1. When three test results are available for each subject, each result could be either positive or negative, as can the (latent) true status of each individual, leading to 16 possible combinations of observed and latent data. Let $Y_1, \ldots, Y_8$ be latent data that represents the number of true positive subjects out of $a, \ldots, h$, subjects in each possible category for the observed test results, respectively. We denote the likelihood function over the observed and latent data by $L(a, b, \ldots, h, Y_1, Y_2, \ldots, Y_8 \mid \theta, S_1, S_2, S_3, C_1, C_2, C_3)$. The likelihood function of the observed and latent data is proportional to the product of each entry in the likelihood contribution column of Table 1 raised to the power of the corresponding entry in the number of subjects column of the table.

Estimates are derived either by maximizing the likelihood function (Walter and Irwig, 1988), or by Bayesian methods, which will provide similar numerical methods to maximizing the likelihood if little prior information over (θ, S_1, S_2, S_3, C_1, C_2, C_3) is used, but has the added advantage of increased precision if reliable prior information is available, especially in small data sets (Joseph et al., 1995). In either case, analytic solutions are not feasible, so the EM algorithm or the Gibbs sampler are typically used in practice to maximize the likelihood or estimate Bayesian posterior distributions, respectively.

In the one test case we are essentially replacing the conditional independence assumption of the three test case with prior information over the test sensitivity and specificity. The situation for two tests is intermediate between the one and three test situations, with the likelihood function derived similarly to that of the three test case above, as given in Joseph et al. (1995). Similar to the three test case, conditional independence between the two tests used is required, but this is of course a weaker condition than requiring all three tests to be conditionally independent. Similar to the one test case, the problem is nonidentifiable (five parameters to estimate, but only three degrees of freedom), and so substantive prior information is required on at least two of the five parameters in order to obtain reasonable estimates. Similar to the one test

Table 1

Likelihood contributions of all possible combinations of observed and latent data for the case of three diagnostic tests. θ represents the prevalence, and S_i and C_i represent the sensitivity and specificity, respectively, of the ith test. The vector of observed numbers of subjects with each possible combination of test results is given by $(a, b, \ldots, h)$, and within each of these cells, we have the unobserved number of truly positive subjects, represented by the vector of latent data, $(Y_1, Y_2, \ldots, Y_8)$.

Truth	Test 1 result	Test 2 result	Test 3 result	Likelihood contribution per subject	Number of subjects
+	+	+	+	$\theta S_1 S_2 S_3$	Y_1
+	+	+	−	$\theta S_1 S_2 (1 - S_3)$	Y_2
+	+	−	+	$\theta S_1 (1 - S_2) S_3$	Y_3
+	+	−	−	$\theta S_1 (1 - S_2)(1 - S_3)$	Y_4
+	−	+	+	$\theta (1 - S_1) S_2 S_3$	Y_5
+	−	+	−	$\theta (1 - S_1) S_2 (1 - S_3)$	Y_6
+	−	−	+	$\theta (1 - S_1)(1 - S_2) S_3$	Y_7
+	−	−	−	$\theta (1 - S_1)(1 - S_2)(1 - S_3)$	Y_8
−	+	+	+	$(1 - \theta)(1 - C_1)(1 - C_2)(1 - C_3)$	$a - Y_1$
−	+	+	−	$(1 - \theta)(1 - C_1)\ (1 - C_2) C_3$	$b - Y_2$
−	+	−	+	$(1 - \theta)(1 - C_1) C_2 (1 - C_3)$	$c - Y_3$
−	+	−	−	$(1 - \theta)(1 - C_1) C_2 C_3$	$d - Y_4$
−	−	+	+	$(1 - \theta) C_1 (1 - C_2)(1 - C_3)$	$e - Y_5$
−	−	+	−	$(1 - \theta) C_1 (1 - C_2) C_3$	$f - Y_6$
−	−	−	+	$(1 - \theta) C_1 C_2 (1 - C_3)$	$g - Y_7$
−	−	−	−	$(1 - \theta) C_1 C_2 C_3$	$h - Y_8$

case, estimates are derived by combining the information in the prior distributions with that provided by the data through the likelihood function.

We will provide nine different estimates of the prevalence of OA: three using results from each test separately, from each of the three possible combinations of two tests, and from all three tests together. For three tests, we provide estimates both using and ignoring prior information, and for both our original and more strict test definitions.

In order to derive the prior distributions (Table 2), we consulted with clinicians familiar with the treatment of OA subjects and in filling in claims forms for the RAMQ database. We directly asked for estimates of proportions of both truly positive and truly negative subjects who would test positive or negative on each of our "tests." For example, to derive the sensitivity of physician diagnosis, we asked "Out of all patients in Quebec who truly have OA, what proportion do you think would have an OA diagnosis recorded in the RAMQ database in any given year?" These estimates were presented in the form of ranges, which we converted to beta distributions. For example, it was thought that 70% to 80% of subjects truly positive for OA in a given year would have an ICD-9 code for OA in the database in that year. This implies a mean value of about 75%, and a standard deviation of about 2.5%, so that four standard deviations (approximately a 95% interval) would cover the given range. This implies beta parameters of $\alpha = 55.5$, and $\beta = 18.5$ (see Joseph et al., 1995). Throughout, we used a uniform prior distribution for the prevalence of OA, our parameter of main interest.

Table 2

Equal tailed 95% probability ranges, and coefficients of the Beta prior distributions for the test parameters of OA "diagnostic tests." A uniform distribution was used for the prior distribution for the prevalence of OA.

	Sensitivity	Specificity
Test 1 (physician diagnosis)		
Range	70% to 80%	90% to 100%
Beta coefficients	$\alpha = 55.5$, $\beta = 18.5$	$\alpha = 17.1$, $\beta = 0.9$
Test 2 (medications)		
Range	70% to 80%	55% to 65%
Beta coefficients	$\alpha = 55.5$, $\beta = 18.5$	$\alpha = 57.0$, $\beta = 38.0$
Test 3 (medical acts)		
Range	20% to 30%	90% to 99%
Beta coefficients	$\alpha = 18.5$, $\beta = 55.5$	$\alpha = 17.1$, $\beta = 0.9$

We used the Gibbs sampler (for three tests) or the SIR algorithm (for the lower dimensional one or two test situations) to derive inferences. Both the Gibbs sampler and the SIR algorithm were run several times each to ensure stability of the results to Monte Carlo variations and different starting values. User-friendly software implementing the Gibbs sampler for diagnostic test data is available from the web page `www.medicine.mcgill.ca/epidemiology/Joseph/Bayesian-Software-Diagnostic-Testing.html`.

Table 3 provides a summary of our models' properties and assumptions. If three or more conditionally independent tests can be identified in the database, then the standard latent class model will be identifiable (Walter and Irwig, 1988; Joseph et al., 1995), so that OA prevalence can be accurately estimated from the data alone, but the conditional independence assumption may not always hold. For example, if a patient is truly OA positive and a physician correctly provides a diagnostic code for OA, they may also be more likely

Table 3

Properties and assumptions of the models used when data from one, two, or three diagnostic tests are available. DF represents degrees of freedom.

Number of tests	Number of unknown parameters	Number of DF in model	Assumes conditional independence?	Requires substantive prior input?
1	3	1	No	Yes
2	5	3	Yes	Yes
3	7	7	Yes	No

to prescribe an NSAID. Similarly, even if a physician incorrectly provides a diagnostic code for OA, they may still be more likely to prescribe an NSAID. Depending on the degree of correlation between the tests within subjects with and without OA, the violation of this assumption may or may not have a large effect on prevalence estimation (Dendukuri and Joseph, 2001; Black and Craig, 2002). While such models can be run, they are again nonidentifiable unless data from at least four tests are available and, therefore, will again rely on prior information about the degree of between-test correlations, and so are likely to add even more uncertainty. As the conditional independence assumption is difficult to verify in practice (but see Section 5 below), other models that rely less (two tests) or not at all (one test) on this assumption can be useful, although these typically require other difficult to verify inputs, such as substantive prior information about diagnostic test properties. As no method is perfect, considering results from a variety of models is useful, and final conclusions can be based on the relative confidence the researcher has in various assumptions required by each method. If prevalence estimates vary substantially depending on the model used, however, and if there is doubt about which set of modeling assumptions is most likely to be correct, researchers must admit that any final inferences remain uncertain.

4. Results

The data on our three "tests" for OA from the RAMQ database are given in Table 4. The naive prevalence estimate using only physician diagnosis without any adjustment for misclassification error is 10.1%, with 95% credible interval (CrI) 10.1–10.2. This is a very narrow interval, but it accounts only for uncertainty due to random variation. Not only is the extra variability due to misclassification errors ignored, but as Greenland (2005) points out, the meaning of random sampling is not clear in observational studies. How much confidence should we place in this seemingly very accurate estimate?

When this same physician diagnosis data are used, but the estimate is now adjusted by our prior distributions for the sensitivity and specificity of this "test" as given in Table 2, the prevalence becomes much less certain, at 11.5% (95% CrI 4.5–14.2). Similarly, when prescribed medications (NSAID or acetaminophen but not methotrexate or plaquenil) are used alone as a diagnostic test (test two), the estimated prevalence is 9.5% (95% CrI 3.3–22.2). When medical procedures (test three) was used, the prevalence estimate is 10.6% (95% CrI 5.2–18.6).

When information from all three diagnostic tests are combined and uniform priors are used across all parameters, the estimated prevalence of OA is 14.8% (95% CrI 14.5–15.1). Using informative prior distributions for all parameters (as given in Table 2), the estimated prevalence of OA matches that given above to at least one decimal place. This is not surprising given the large sample size of the data set in this identifiable problem. While the results for three tests, therefore, do not substantially depend on prior distributions, they do depend on the conditional independence assumption, and on the reasonableness of all three tests for OA, which together implicitly define OA positivity (Alonzo and Pepe, 1999). As expected, when only partial data are used (one test), the CrI are wider compared to the situation when three tests are used. The posterior estimates of the sensitivity and specificity using one test are not all within the range of the priors given in Table 2, partially explaining the discrepancies in prevalence estimates between the different analyses. We return to this point in the discussion.

For two tests, when combining diagnostic code and prescribed medication, the posterior prevalence of OA is 11.8% (95% CrI 8.6–14.8). When combining diagnostic code and medical procedures, OA prevalence is estimated by 9.8% (95% CrI 6.4–13.7). Finally, when combining prescribed medication

Table 4

Results of the three diagnostic tests for OA on elderly individuals resident of Quebec in 2002. *Test* 1 *represents physician diagnosis as determined by ICD-*9 *code, Test* 2 *indicated OA medications* (*but not lupus or RA medications*) *as per drug identification number code, and Test* 3 *represents OA related procedures, including injection procedures, arthroplasty or a tibial osteotomy.*

Test 1 Phys. diagnosis	Test 2 Medication	Test 3 Medical acts	Number of individuals observed (main test definitions)	Number of individuals observed (revised test definitions)
+	+	+	11,816	7104
+	+	−	57,222	19,595
+	−	+	3320	2570
+	−	−	25,651	9416
−	+	+	9610	11,208
−	+	−	260,923	204,008
−	−	+	5002	8866
−	−	−	595,415	706,192

Table 5

Marginal posterior medians (upper entry of each cell) and lower and upper limits of the posterior equal tailed 95% CrI (lower entry of each cell) for the prevalence (θ), the sensitivities (S_1, S_2, S_3), and the specificities (C_1, C_2, C_3) from each analyses (i.e., three tests, two tests, one test, and three tests with new definitions).

	θ	S_1	S_2	S_3	C_1	C_2	C_3
Prior distribution	50.0 0.0–100	75.0 70.0–80.0	75.0 70.0–80.0	25.0 20.0–30.0	95.0 90.0–100	60.0 55.0–65.0	95.0 90.0–100
One test							
Physician diagnosis alone	11.5 4.5–14.2	72.8 63.2–82.4		95.4	 92.8–99.9		
Prescribed medication alone	9.5 3.3–22.0		72.3 62.5–81.9			71.1 66.3–75.6	
Medical procedure alone	10.6 5.2–18.6			22.1 14.7–33.8			99.2 97.9–99.9
Two tests							
Combination of physician diagnosis and prescribed medication	11.8 8.6–14.8	75.1 68.4–81.4	76.1 73.9–78.4		98.9 98.1–99.1	70.5 70.1–71.3	
Combination of physician diagnosis and medical acts	9.8 6.4–13.7	74.1 63.2–83.2		23.5 15.9–31.1	96.7 94.3–99.6		99.2 98.6–99.3
Combination of prescribed medication and medical acts	10.0 6.8–16.4		77.6 72.2–83.3	24.0 18.2–38.2		70.6 68.2–72.5	99.6 99.3–99.9
Three tests—Original test definitions							
Three tests using informative priors	14.8 14.5–15.1	58.1 57.1–58.7	78.3 77.6–79.0	18.2 17.8–18.5	98.1 98.0–98.2	72.4 72.3–72.6	99.5 99.5–99.5
Three tests using noninformative priors	14.8 14.5–15.1	58.2 57.0–59.0	78.3 77.6–79.0	18.2 17.8–18.5	98.1 98.0–98.3	72.4 72.3–72.6	99.5 99.5–99.5
Three tests—stricter test definitions							
Three tests using noninformative priors	8.6 8.5–9.0	41.5 41.0–42.8	72.9 72.6–74.4	27.2 26.6–27.9	99.6 99.6–99.7	79.6 79.5–79.8	99.2 99.2–99.3

and medical procedures, the estimated prevalence is 10.0% (95% CrI 6.8–16.4). While all point estimates hover close to the 10.1% naive prevalence estimate using physician diagnosis alone, the CrI are roughly 60 to 80 times wider than the naive confidence interval. Even when the problem is identifiable (three test case), the CrI is approximately 10 times wider compared to the naive confidence interval. Clearly, in this case, most uncertainty arises from sources other than random variation.

In further sensitivity analyses, we changed the test definitions for tests one and two to be more restrictive (data given in last column of Table 4). Using data from all three tests and uniform prior distributions across all parameters, the estimated prevalence of OA was 8.6% (95% CI 8.5–9.0). As expected, S_1 and S_2 decreased while C_1 and C_2 increased compared to the analysis using less restrictive test one and test two definitions. Table 5 displays all results, including sensitivity and specificity estimates.

5. Discussion

Many researchers have derived prevalence estimates using a physician diagnosis field from a database, without any adjustment for the almost inevitable misclassification errors. The unadjusted prevalence estimate for OA using only physician diagnosis was 10.1%, compared to the 14.5% estimate using all three "tests," and adjusting for misclassification. This represents an almost 50% increase in the prevalence estimate. While nothing guarantees that the estimate of 14.5% is correct or even necessarily better than the naive 10.1% estimate, the latter is almost surely based on the incorrect assumption of perfect database information on diagnosis of OA, and has a confidence interval that is much too narrow. Calculating estimates across a variety of alternative models displays the impact not only of adjusting for misclassification error, but also shows how different modeling assumptions affect parameter estimates. We have found that the prevalence of OA varies between 3.3% and 22.0%, depending on the statistical model assumed and prior distribution chosen. This quite large interval shows that estimating prevalence from administrative databases can be highly problematic. To substantially shrink this interval, one needs to select a model as the "correct model," which is difficult since all rely on largely unverifiable assumptions or uncertain prior distributions on the sensitivity and specificity of the "tests."

While there has been some preliminary work on the reliability of administrative database items for research purposes (e.g., Wilchesky et al., 2004), sensitivity and specificity will rarely be exactly known in advance. Therefore, the three test statistical model in which prior distributions do not play a major role is useful, provided that the assumption of conditional independence holds. Similar latent class models can be developed for conditionally dependent tests (Yang and Becker, 1997; Dendukuri and Joseph, 2001; Pepe, 2003), but these models are nonidentifiable, so the prevalence estimate will again strongly depend on the choice of prior distribution for the correlation and other parameters. Therefore, one has the choice of an identifiable model with stronger assumptions, compared to a more complex and nonidentifiable model.

It will often be difficult to know if prior distributions selected are well calibrated. Our prior distributions for the sensitivities and specificities did not always closely match the posterior distributions for these parameters from the three test analysis. Prior information for S_2, C_1, and C_3 are relatively close to their corresponding posterior estimates, but priors for S_1, S_3, and C_2 are quite different from their estimates. This suggests that either physicians have only a very rough idea of these properties, or that some assumptions in our model, such as conditional independence, do not hold, or both.

We elicited independent prior information on each test property, as shown in Table 2. A more sophisticated design might elicit prior distributions over observable quantities, allowing for direct calibration of the prior distributions used. For example, one might elicit prior opinion about the proportions eventually observed in Table 4, asking physicians "What proportion of all subjects do you think will test positive on all three tests?" and so on. Not only could these prior quantities be converted into prior distributions on the test properties themselves [see similar work by Kynn (2001) on priors for logistic regression parameters obtained via questions about observable quantities], but comparing the prior predictions to the actual observed data would help to calibrate the priors obtained from the physicians. Physicians may be less skilled in eliciting multivariate prior quantities compared to single parameter elicitation. Eliciting the proportion who will be positive on all three tests involves estimating not only the prevalence of OA, but also the possibly correlated results from all three tests simultaneously. On the other hand, this allows physicians to account for a priori correlations among parameters. For example, a physician who expects high prevalence may also tend to provide high values for the sensitivity of one or more of the tests. Clearly, much important work remains to be done in effective prior elicitation methods.

While patients with mild OA would be treated with NSAIDs or acetaminophen, only more severe cases will have surgery. Therefore, another issue is that our tests may focus on slightly different subsets of subjects with OA. Prevalence estimates dropped from 14.5% under the original "test" definitions to 8.5% under the more strict definitions we used as part of our robustness study. In addition, the degree of conditional dependence among our tests may change as test definitions are modified, further compounding the problem of choosing one set of estimates over another.

There are several ways to improve robustness of estimates from such studies by considering designs that collect additional information on a subsample of subjects. For example, suppose it is possible to obtain data (say Z) on all medical tests, results of physical exams, x-rays, and so on that are relevant to physician diagnosis. If these results are complete (presumably a physician survey must be first done to derive a complete list of items upon which diagnoses would be based), then any actions taken by the physicians [e.g., $X =$ (entering an OA diagnosis in the database, prescriptions, medical acts)] would be independent conditional on Z. This would allow a three test model to assume between-test conditional independence given Z, providing good estimates of the sensitivity and specificity of each test to use in further prevalence estimates derived from the entire database. Alternatively, one could obtain a definitive diagnosis from a subset of patients through an in-depth investigation, providing a gold standard against which to compare the performance of the other "tests."

While we illustrate the problem via estimating prevalence of OA using databases, the problem carries over to other prevalence estimation problems (Hadgu, 1997; Ferreccio et al., 2003; Carabin et al., 2005), and potentially, to other research uses of information in databases (Mamdani et al., 2002; Ray et al., 2002; Solomon et al., 2004). More, generally, similar problems may arise in any observational study using imperfect data (Greenland 2005). Future work must include validation studies that provide estimates of the reliability of uncertain data items, such as those outlined above. For example, Bernatsky et al. (2005) added a chart review component to their database study, using this additional data to adjust their main parameter estimates.

References

Alonzo, T. A. and Pepe, M. S. (1999). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* **18,** 2987–3003.

Badley, E. M. and Webster, G. K. (1995). The impact of muscoloskeletal in the population: Are they just aches and pain? Finding from the 1990 Ontario Health Survey. *Journal of Rheumatology* **22,** 733–739.

Berger, R. G. (1994). Nonsteroidal anti-inflammatory drugs: Making the right choices. *Journal of the American Academy of Orthopaedic Surgeons* **2,** 255–260.

Bernatsky, S., Joseph, L., Bélisle, P., Boivin, J., Rajan, R., Moore, A., and Clarke, A. (2005). A Bayesian hierarchical model for estimating the properties of cancer ascertainment methods in cohort studies. *Statistics in Medicine* **24,** 2365–2379.

Black, M. and Craig, B. (2002). Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* **21,** 2653–2669.

Carabin, H., Marshall, C., Joseph, L., Riley, S., Olveda, R. and McGarvey, S. (2005). Estimating and modelling the dynamics of the intensity of infection with Schistosoma japonicum in villagers of Leyte, Philippines Part I: A Bayesian cumulative logit model. *American Journal of Tropical Medicine and Hygiene* **72,** 745–753.

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests. *Biometrics* **57,** 208–217.

Demissie, K., White, N., Joseph, L., and Ernst, P. (1998). Bayesian estimation of asthma prevalence, and comparison of exercise and questionnaire diagnostics in the absence of a gold standard. *Annals of Epidemiology* **8,** 201–208.

Ferreccio, C., Bratti, M. C., Sherman, M. E., Herrero, R., Wacholder, S., Hildesheim, A., Burk, R. D., Hutchinson, M., Alfaro, M., Greenberg, M. D., Morales, J., Rodriguez, A. C., Schussler, J., Eklund, C., Marshall, G., and Schiffman, M. (2003). A comparison of single and combined visual, cytologic, and virologic tests as screening strategies in a region at high risk of cervical cancer. *Cancer Epidemiology, Biomarkers and Prevention* **12,** 815–823.

Gabriel, S. E., Crowson, C. S., and O'Fallon, W. M. (1995). Costs of osteoarthritis: Estimates from a geographically defined population. *The Journal of Rheumatology* **43,** 23–25.

Green, J. and Wintfeld, N. (1993). How accurate are hospital discharge data for evaluating effectiveness of care? *Medical Care* **31,** 719–731.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society, Series A* **168,** 267–306.

Grelsamer, R. P. (1995). Current concepts review. Unicompartmental osteoarthrosis of the knee. *The Journal of Bone and Joint Surgery* **77,** 278–292.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology.* New York: Chapman and Hall.

Gustafson, P. (2005). On model expansion, model contraction, identifiability, and prior information: Two illustrative scenarios involving mismeasured variables (with discussion). *Statistical Science* **20,** 111–129.

Hadgu, A. (1997). Bias in the evaluation of DNA-amplification tests for detecting Chlamydia trachomatis. *Statistics in Medicine* **16,** 1391–1399.

Johnson, W., Gastwirth, J., and Pearson, L. (2001). Screening without a "Gold standard": The Hui-Walter paradigm revisited. *American Journal of Epidemiology* **153,** 921–924.

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

Kynn, M. (2001). ELICITOR: A novel interactive graphical approach to eliciting expert opinion for the logistic regression model with normal priors. In *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, B. Klein and L. Korsholm (eds), 267–274. Denmark: University of Southern Denmark.

Lawrence, R. C. and Helmick, C. G. (1998). Estimates of the prevalence of arthritis and selected muscoloskeletal disorders in the United States. *Arthritis & Rheumatism* **41,** 778–799.

Mamdani, M., Rochon, P. A., Juurlink, D. N., Kopp, A., Anderson, G. M., Naglie, G., Austin, P. C., and Laupacis, A. (2002). Observational study of upper gastrointestinal haemorrhage in elderly patients given selective cyclo-oxygenase-2 inhibitors or conventional non-steroidal anti-inflammatory drugs. *British Medical Association* **325,** 624.

McInturff, P., Johnson, W. O., Cowling, D., and Gardner, I. A. (2004). Modelling risk when binary outcomes are subject to error. *Statistics in Medicine* **23,** 1095–1109.

Papaioannou, A., Parkinson, W., Ferko, N., Probyn, L., Ioannidis, G., Jurriaans, E., Cox, G., Cook, R. J., Kumbhare, D., and Adachi, J. D. (2003). Prevalence of vertebral fractures among patients with chronic obstructive pulmonary disease in Canada. *Osteoporosis International* **14,** 913–917.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press.

Quan, H., Parsons, G. A., and Ghali, W. A. (2004). Assessing accuracy of diagnosis-type indicators for flagging complications in administrative data. *Journal of Clinical Epidemiology* **57,** 366–372.

Rahme, E., Joseph, L., and Gyorkos, T. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics* **49,** 119–228.

Ray, W. A., Stein, C. M., Daugherty, J. R., Hall, K., Arbogast, P. G., and Griffin, M. R. (2002). COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *The Lancet* **360,** 1071–1073.

Robertson, C. M., Svenson, L. W., and Joffres, M. R. (1998). Prevalence of cerebral palsy in Alberta. *The Canadian Journal of Neurological Sciences* **25,** 117–122.

Romano, P. S., Schembri, M. E., and Rainwater, J. A. (2002). Can administrative data be used to ascertain clinically significant postoperative complications? *American Journal of Medical Quality* **17,** 145–154.

Rushton, L. and Romaniuk, H. (1997). Comparison of the diagnosis of leukaemia from death certificates, cancer registration and histological reports—implications for occupational case-control studies. *British Journal of Cancer* **75,** 1694–1698.

Solomon, D. H., Schneeweiss, S., Glynn, R. J., Kiyota, Y., Levin, R., Mogun, H., and Avorn, J. (2004). Relationship between selective cyclooxygenase-2 inhibitors and acute myocardial infarction in older adults. *Circulation* **109,** 2068–2073.

Swerdlow, A. J., Douglas, A. J., and Vaughan, H. G. (1993). Completeness of cancer registration in England and Wales. *British Journal of Cancer* **67,** 326–329.

Tamblyn, R. (1995). The use of prescription claims databases in pharmacoepidemiological research: The accuracy and comprehensiveness of the prescription claims database in Quebec. *Journal of Clinical Epimemiology* **48,** 999–1009.

Thompson, N. W. (2001). Total knee arthroplasty without patellar resurfacing in isolated patellofemoral osteoarthritis. *The Journal of Arthroplasty* **16,** 607–612.

Walter, S. D. and Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology* **41,** 923–937.

Wegman, A., van der Windt, D., van Tulder, M., Stalman, W., and de Vries, T. (2004). Nonsteroidal antiinflammatory drugs or acetaminophen for osteoarthritis of the hip or knee? A systematic review of evidence and guidelines. *The Journal of Rheumatology* **31,** 344–354.

Wilchesky, M., Tamblyn, R. M., and Huang, A. (2004). Validation of diagnostic codes within medical services claims. *Journal of Clinical Epimemiology* **57,** 131–141.

Yang, I. and Becker, M. P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics* **53,** 948–958.

Received November 2005. *Revised May* 2006.
Accepted May 2006.