Identifying Rheumatoid Arthritis Cases within the Quebec Health Administrative Database

Zeinab F. Slim, Cristiano Soares de Moura, Sasha Bernatsky, and Elham Rahme

ABSTRACT. ***Objective.*** Our objective was to calculate rheumatoid arthritis (RA) point prevalence estimates in the CARTaGENE cohort, as well as to estimate the sensitivity and specificity of our ascertainment approach, using physician billing data. We investigated the effects of using varying observation windows in the Régie de l'assurance maladie du Québec (RAMQ) health services administrative databases, alone or in combination with self-reported diagnoses and drugs.

Methods. We studied subjects enrolled in the CARTaGENE cohort, which recruited 19,995 participants from 4 metropolitan regions in Québec from August 2009 to October 2010. A series of Bayesian latent class models were developed to assess the effects of 3 factors: the number of years of billing data, the addition of self-reported information on RA diagnoses and drugs, and the adjustment for misclassification error.

Results. The 3-year 2010 point prevalence estimate among cohort members aged 40–69 years, using physician billing plus self-report, adjusting for misclassification error in each source, was 0.9% [95% credible interval (CrI) 0.7–1.2] with RAMQ sensitivity of 84.0% (95% CrI 74.0–93.7) and a specificity of 99.8% (95% CrI 99.6–100.0). Our results show variations in the prevalence point estimates related to all 3 factors investigated.

Conclusion. Our study illustrates that multiple data sources identify more RA cases and thus a higher prevalence estimate. RA point prevalence estimates using billing data are lower if fewer years of data are used. (First Release August 1 2019; J Rheumatol 2019;46:1570–6; doi:10.3899/jrheum.181121)

Key Indexing Terms:
BAYESIAN LATENT CLASS MODELS PREVALENCE QUEBEC SELF-REPORT DATA
CANADIAN PROVINCIAL HEALTH ADMINISTRATIVE DATA RHEUMATOID ARTHRITIS

Rheumatoid arthritis (RA) is a type of chronic autoimmune disease, and like most chronic diseases, it is caused by a constellation of potential factors, including environmental and genetic risk factors[1]. Surveillance data can provide insights into the epidemiology of RA. Additionally, prevalence data derived from surveillance can assist in making future projections and studying geographic variations[2]. Having unbiased prevalence estimates is essential to improving care and outcomes. In Canada, the provincial government health insurance is nearly universal and administrative databases such as those collected by the Régie de l'assurance maladie du Québec (RAMQ) have been an attractive resource for prevalence studies on RA[3]. Methods for estimating RA prevalence in these databases rely on physician billing and/or hospitalization International Classification of Diseases (ICD) codes[4]. Prevalence estimates of RA obtained from administrative health databases have varied depending on several factors such as case definitions[5] and the size of the observation window available for analysis in the health administrative database[6,7]. Any ascertainment method within health administrative databases may miss some true cases and misclassify others.

An additional source of data for RA surveillance is self-reported data collected from large survey databases[8,9,10,11]. Ascertainment of RA based on the patient's self-reported data should be done with caution because misclassification is a concern. Supplementing this ascertainment method with medication information such as disease-modifying antirheumatic drugs (DMARD) improved the accuracy of self-reported RA in some studies[12]. DMARD are the cornerstone of RA treatment and according to national and international guidelines, all RA patients with active disease should be offered DMARD therapies. Of course, a

From the Department of Epidemiology, Biostatistics and Occupational Health, McGill University; Division of Clinical Epidemiology, Research Institute of the McGill University Health Centre; Department of Medicine, McGill University, Montreal, Quebec, Canada.

Z.F. Slim, PhD, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, and Division of Clinical Epidemiology, Research Institute of the McGill University Health Centre; C. Soares de Moura, PhD, Research Associate, Centre for Outcomes Research and Evaluation, Division of Clinical Epidemiology, Research Institute of the McGill University Health Centre; S. Bernatsky, MD, PhD, Professor of Medicine, Division of Clinical Epidemiology, Research Institute of the McGill University Health Centre, and Department of Medicine, McGill University; E. Rahme, PhD, Associate Professor of Medicine, Division of Clinical Epidemiology, Research Institute of the McGill University Health Centre, and Department of Medicine, McGill University.

Address correspondence to Dr. S. Bernatsky, Centre for Outcomes Research and Evaluation of Research Institute of the McGill University Health Centre, 5252 Boul. de Maisonneuve Ouest, Office 3F.51, Montreal, Quebec H4A 3S5, Canada. E-mail: sasha.bernatsky@mcgill.ca

Accepted for publication February 13, 2019.

small number of patients with RA cannot take these drugs (if their RA is in remission — a relatively rare event — or for other reasons), so there could be false-positive and false-negative RA cases using this method as well. What makes the situation more challenging in large population-based surveillance studies is the absence of a gold standard to validate self-report or health administrative data sources.

Few RA prevalence estimates are available in Quebec or even in Canada. One prior study estimated RA prevalence for Quebec, using only physician billing and hospitalization diagnostic codes for the period 1992–2008; this accounted for misclassification error in administrative data[3]. However, additional studies may be helpful to elucidate the effects of the observation window within health administrative databases, the use of self-reported information, and the adjustment for misclassification error in all ascertainment methods on RA prevalence estimates. This study's specific objective was to calculate, within 11 different observation windows in physician billing data, 2010 RA point prevalence estimates (unadjusted and adjusted for misclassification error) among the CARTaGENE cohort of adults aged 40–69 years, as well as to estimate the sensitivity and specificity of our ascertainment approach, using administrative data (alone or combined with self-reported data)[13].

MATERIALS AND METHODS

Study setting, sources of data, ascertainment of RA cases, and time frame. This study took place in the context of a large established cohort entitled CARTaGENE, which recruited 19,995 participants (aged 40–69 yrs) from August 2009 to October 2010 from 4 metropolitan regions in Québec (Montreal, Sherbrooke, Québec City, and Saguenay, constituting 55.7% of the Quebec population). Participants were randomly selected from the provincial health insurance FIPA files (*fichier administratif des inscriptions des personnes assurées*), which include the entire population because health insurance coverage in Quebec is universal. Individuals were excluded if they were not registered in the FIPA files (such as the military), resided outside the selected regions in 2009, lived in First Nations reserves or longterm healthcare facilities, or were in prison. Participants were invited to an interview and completed a self-administered sociodemographic and lifestyle questionnaire as well as an interviewer-administered health questionnaire. Participation rate was 25.6% and there were regional variations in the participation rates, with the Saguenay region having the highest participation rate (33.9%) and the Montreal northern suburbs having the lowest (21.8% for Laval and 21.2% for the North Shore). Data on demographic and socioeconomic factors, lifestyle habits, mental health, individual and family history of disease, medical care history such as visits to a doctor or a nurse, and current medications were collected[8]. Further details on the CARTaGENE cohort can be found elsewhere[8]. The CARTaGENE research cohort has been linked to RAMQ data using patients' unique provincial health insurance numbers. The RAMQ medical service database has information on physician outpatient visits, including diagnoses coded according to the International Classification of Diseases, 9th revision (ICD-9) during the time interval of data collection.

Our study included all CARTaGENE participants who were interviewed between 2009 and 2010. Individuals with incomplete or missing information concerning RA diagnosis and current DMARD use were excluded. Therefore, our reported estimates may be considered as estimates of 2010 point prevalence for RA, in which the point represents the end of 2010 and the denominators are those individuals enrolled in CARTaGENE by the end of the data collection phase. Survey-based RA cases were defined using the self-reported information on RA diagnosis as well as current use of either conventional DMARD (hydroxychloroquine, sulfasalazine, methotrexate, leflunomide, azathioprine, cyclosporine, gold, and cyclophosphamide) and/or the biologic DMARD (infliximab, adalimumab, etanercept, abatacept, and rituximab). RAMQ-based RA cases were defined using physicians' claims data according to an algorithm requiring 2 or more RA diagnoses by any physician at least 2 months apart but within a 2-year span, or at least 1 RA diagnosis by a rheumatologist.

RAMQ data were available for our study subjects from January 1, 1998, to December 31, 2010. Eleven successive nested observation windows that ranged from a minimum of 3 years (2008–2010) to a maximum of 13 years (1998–2010) were constructed by adding successively one earlier year to the years under observation (2008–2010; 2007–2010 … 1998–2010). Therefore, all time windows ended in December 2010 and were used to calculate the point prevalence of 2010.

Statistical methods. In our analyses, we considered both the self-reported and physician claims ascertainment methods to be imperfect. In such case (i.e., in the absence of a gold standard), the true RA status can be thought of as "missing." By knowing the values of the sensitivity and specificity of the imperfect ascertainment method, a latent class analysis can be used to adjust the prevalence for misclassification errors. We used a Bayesian latent class analysis to summarize the existing information about each variable (sensitivity, specificity, and prevalence) in the form of prior distributions. Then, the prior information was updated by the data through Bayes' theorem to result in posterior distributions of these variables[14,15,16,17,18].

More specifically, the number of subjects who are categorized as having RA according to each imperfect ascertainment method is a mix of true-positive and false-positive individuals. The Bayesian latent class model links the observed results of each method to the unobserved truth of RA status using the following formula: (total sample size)*[(prevalence of RA*sensitivity of the ascertainment method) + (1 – prevalence)(1 – specificity of the ascertainment method)][18].

Informative prior distributions were used over the sensitivity and specificity of RAMQ based on the subjective opinions of 8 experts in the field as well as on a published validation study of provincial administrative data, which used primary care records as reference standard[19]. We varied the prior distributions of the sensitivity and specificity of the physician claim ascertainment method ranging from 60% to 90% and 82% to 99%, respectively. Informative prior distributions over the prevalence ranging from 0% to 8% were chosen based on the literature. For the sensitivity and specificity of self-reported data, "uninformative" prior distributions [e.g., β (1,1)] were used. For all variables, a β prior distribution was used[18].

In a Bayesian latent class model, the likelihood function relating the observed and latent data to the unknown variables for one ascertainment method (i.e., RAMQ) is as follows:
$L(a,b,X,Y/\ \pi, Se, Sp) = [\pi Se]^X\ [\pi(1 - Se)]^Y\ [(1 - \pi)(1 - Sp)]^{a-X}\ [(1 - \pi)(Sp)]^{b-Y}$, where "a" and "b" are the observed number of individuals with positive and negative results on the ascertainment method (here RA diagnoses in RAMQ), respectively; X and Y are the latent truly positive; π is the prevalence of RA; and Se and Sp are the sensitivity and specificity of the ascertainment method, respectively. In the case where RAMQ was combined with self-reported sources, the likelihood contributions of all possible combinations of observed and latent data are provided in Table 1. The likelihood is proportional to the product of each entry in the last column raised to the power of the corresponding entry in the first column of the table.

To address the potential issue that self-reported RA diagnosis and DMARD use may be dependent, even conditional, on the true disease status in the model combining the 3 methods, conditional correlation between the 2 CARTaGENE self-reported sources of information in RA subjects and in non-RA subjects were incorporated[20].

The unadjusted (naive) estimates of RA prevalence were estimated based on RAMQ billing codes for each time window in administrative data. These estimates were obtained by dividing the number of those diagnosed with RA by the total sample size. The unadjusted prevalence estimates were calculated using the Bayesian method for single proportions. Uninformative β

Table 1. Likelihood contribution of observed and latent data when combining Régie de l'assurance maladie du Québec (RAMQ) ascertainment method with self-reported RA diagnosis and DMARD use.

N	Truth	RAMQ Result	Self-RA Diagnosis Result	DMARD Use Result	Likelihood Contribution
Y1	+	+	+	+	πSe1Se2Se3
Y2	+	+	+	–	πSe1Se2(1–Se3)
Y3	+	+	–	+	πSe1(1–Se2)Se3
Y4	+	+	–	–	πSe1(1–Se2)(1–Se3)
Y5	+	–	+	+	π(1–Se1)Se2Se3
Y6	+	–	+	–	π(1–Se1)Se2(1–Se3)
Y7	+	–	–	+	π(1–Se1)(1–Se2)Se3
Y8	+	–	–	–	π(1–Se1)(1–Se2)(1–Se3)
a-Y1	–	+	+	+	(1–π)(1–Sp1)(1–Sp2)(1–Sp3)
b-Y2	–	+	+	–	(1–π)(1–Sp1)(1–Sp2)Sp3
c-Y3	–	+	–	+	(1–π)(1–Sp1)Sp2(1–Sp3)
d-Y4	–	+	–	–	(1–π)(1–Sp1)Sp2Sp3
e-Y5	–	–	+	+	(1–π)Sp1(1–Sp2)(1–Sp3)
f-Y6	–	–	+	–	(1–π)Sp1(1–Sp2)Sp3
g-Y7	–	–	–	+	(1–π)Sp1Sp2(1–Sp3)
h-Y8	–	–	–	–	(1–π)Sp1Sp2Sp3

A–H are the observed results of 3 ascertainment methods. Y1–Y8 are the latent truly positive subjects. π is the prevalence of RA. Se1, Se2, Se3, Sp1, Sp2, and Sp3 are the sensitivity and specificity of 3 ascertainment methods. RA: rheumatoid arthritis; DMARD: disease-modifying antirheumatic drug.

prior distribution [e.g. β (1,1)], where all values are equally likely, was used over the unknown unadjusted prevalence variable. In this case, the posterior prevalence estimates (unadjusted for misclassification error) are expected to be numerically the same as those obtained using frequentist method[18] (i.e., dividing the number of those diagnosed with RA using billing codes by the total sample size).

Posterior estimates for each variable were determined based on a sample from the posterior distribution using Gibbs sampling with the WinBUGS statistical freeware (version 1.4.3, MRC Biostatistics Unit). Each model was assessed after a burn-in of 5000 iterations and a further 30,000 iterations for use in inferences[21]. The mean and 2.5–97.5 percentile values (95% credible intervals; CrI) for each variable were extracted.

Approval for the study was obtained from McGill University Ethics Review Board (approval number: A04-M47-12B), CARTaGENE as well as Commission d'accès à l'information du Québec (approval number: 100 49 57). Additionally, participants signed a written informed consent to publish the material.

RESULTS

The baseline characteristics of the study cohort were evaluated, including age, sex, geographical region, education, and current working status. Just over half of the sample was female, and the overwhelming majority lived in Montreal. The full profile of the participants is presented in Table 2.

Using only self-reported RA diagnosis, without any adjustment for misclassification, the RA prevalence estimate was 2.9% (564 out of 19,704) with 95% CrI 2.6–3.1. The naive estimate from DMARD use was lower at 0.9% (182 out of 19,704) with 95% CrI 0.8–1.1. Adjusting for misclassification error decreased the point prevalence estimate to 1.3% (95% CrI 0.07–3.2) for self-RA diagnosis and 0.4% (95% CrI 0.02–1.1) for current DMARD use.

We found 197 RA cases using only 3 years of physician billing, unadjusted for misclassification error. When more years were used, the number of RA cases continued to increase, up to 321 when looking back 13 years.

Table 2. Demographics of CARTaGENE participants who have complete self-reported information.

Demographics	Total, n = 19,704
Age, yrs, mean ± SD	54.2 ± 7.9
Sex	
Male	9552 (48.5)
Female	10,152 (51.5)
Region	
Montreal	15,001 (76.1)
Quebec	2997 (15.2)
Saguenay	789 (4.0)
Sherbrooke	917 (4.6)
Education	
High school and less	5096 (25.9)
College	6239 (31.7)
University	8262 (41.9)
Employment status	
Currently working	12,835 (65.1)
Retired	4363 (22.1)
Unable to work	789 (4.0)
Unemployed	1030 (5.2)
Caregiving at home	554 (2.8)

Values are n (%) unless otherwise specified.

The unadjusted 2010 RA prevalence point estimate based on 3 years of RAMQ data alone was 1.0% (197 RA cases out of 19,704) with 95% CrI 0.9–1.2. Using 5 years of data, the prevalence point estimate increased by 20%. When using 13 years of RAMQ data, there was a 60% increase in the unadjusted prevalence point estimate (1.6%, 95% CrI 1.5–1.8) compared to the estimate from using 3 years of data (Table 3).

Adjusting for misclassification error using the Bayesian latent class model, RA prevalence point estimate was 0.4% (95% CrI 0.03–1.1) for the shortest observation window.

Table 3. RA prevalence of the different combinations of ascertainment methods for the 11 observation periods, CARTaGENE cohort, Quebec, 2009–2010.

	Unadjusted Prevalence Using RAMQ Data Alone	Adjusted Prevalence Using RAMQ Alone	Adjusted Prevalence Using RAMQ and the Self-reported Data
3-yr period (Jan 1, 2008–Dec 31, 2010)	1.0 (0.9–1.2)	0.4 (0.03–1.1)	0.9 (0.7–1.2)
4-yr period (Jan 1, 2007–Dec 31, 2010)	1.1 (1.0–1.2)	0.5 (0.03–1.2)	0.9 (0.7–1.2)
5-yr period (Jan 1, 2006–Dec 31, 2010)	1.2 (1.0–1.3)	0.5 (0.03–1.3)	1.0 (0.8–1.3)
6-yr period (Jan 1, 2005– Dec 31, 2010)	1.2 (1.1–1.4)	0.5 (0.03–1.3)	1.0 (0.8–1.3)
7-yr period (Jan 1, 2004– Dec 31, 2010)	1.3 (1.1–1.4)	0.5 (0.03–1.4)	1.0 (0.8–1.4)
8-yr period (Jan 1, 2003– Dec 31, 2010)	1.3 (1.2–1.5)	0.6 (0.03–1.5)	1.0 (0.8–1.4)
9-yr period (Jan 1, 2002– Dec 31, 2010)	1.4 (1.2–1.6)	0.6 (0.03–1.5)	1.1 (0.8–1.5)
10-yr period (Jan 1, 2001– Dec 31, 2010)	1.5 (1.3–1.6)	0.6 (0.04–1.6)	1.1 (0.8–1.5)
11-yr period (Jan 1, 2000– Dec 31, 2010)	1.5 (1.3–1.7)	0.6 (0.04–1.6)	1.1 (0.9–1.5)
12-yr period (Jan 1, 1999– Dec 31, 2010)	1.6 (1.4–1.8)	0.7 (0.04–1.7)	1.2 (0.9–1.6)
13-yr period (Jan 1, 1998– Dec 31, 2010)	1.6 (1.5–1.8)	0.7 (0.04–1.7)	1.2 (0.9–1.6)

Values are posterior means (lower and upper limits of the posterior equal-tailed 95% credible intervals. RA: rheumatoid arthritis; RAMQ: Régie de l'assurance maladie du Québec.

Additionally, the adjusted prevalence was lower than the unadjusted prevalence estimates for all observation windows. The adjusted estimates across all time windows showed an increasing trend but remained lower than the RAMQ-based unadjusted estimate. The CrI around the adjusted point estimate using RAMQ alone were much wider than the CrI around the unadjusted estimates, which is expected because adjustment accounts for misclassification.

As for the combined RAMQ and self-reported information, the different combinations of the observed data are presented in Supplementary Table 1 (available from the authors on request). For all observation windows, the adjusted point estimates derived from combining RAMQ with self-reported data were lower than the unadjusted estimates and higher than the adjusted estimates using RAMQ alone. When combining administrative and self-reported data, adding more years of administrative data increased the adjusted point estimates (Table 3) in a similar fashion to when administrative data were used alone. The CrI were all overlapping. Figure 1 shows the increasing trends in the point estimates (unadjusted and adjusted, with administrative data alone and then adding self-reported data).

The results for the sensitivity estimates of case ascertainment across varying time windows (with administrative data alone and combining with self-reported data) are shown in Table 4. The sensitivity of case ascertainment using

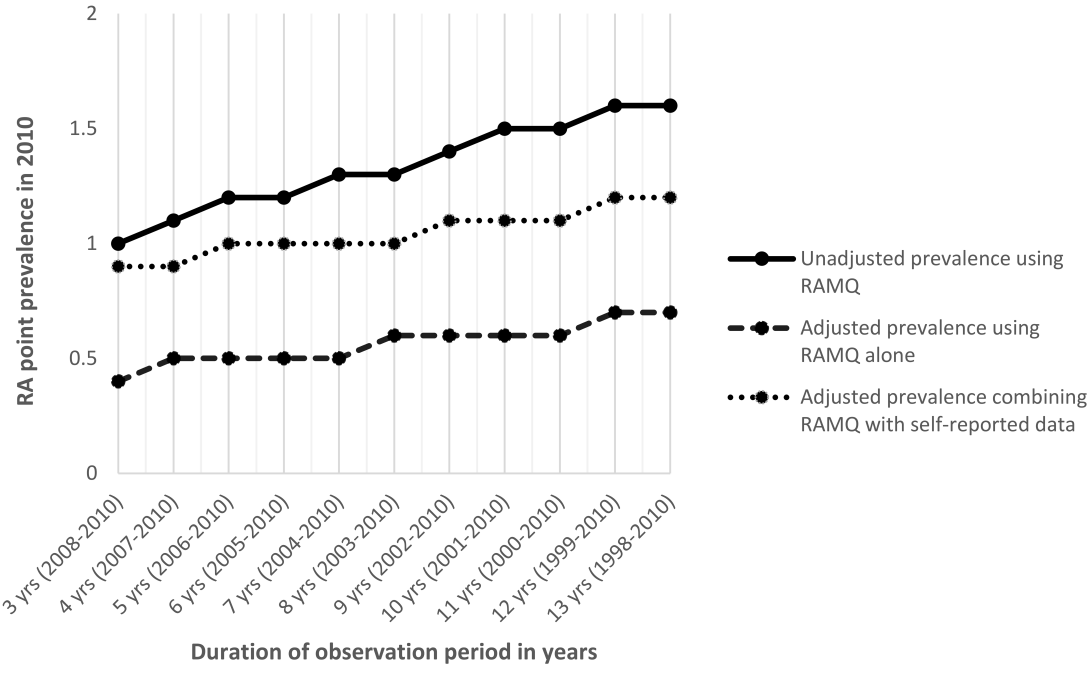

Figure 1. RA point prevalence estimates by the duration of observation period within Régie de l'assurance maladie du Québec (RAMQ). RA: rheumatoid arthritis.

Table 4. Sensitivity and specificity of Régie de l'assurance maladie du Québec (RAMQ) ascertainment method alone and when combining it with self-reported data for the 11 observation periods, CARTaGENE cohort, Quebec (2009–2010).

	RAMQ Alone		RAMQ Combined with Self-reported Data	
	Sensitivity, %	Specificity, %	Sensitivity, %	Specificity, %
3-yr period (Jan 1, 2008–Dec 31, 2010)	78.1 (58.3–92.6)	99.3 (99.0–99.8)	84.0 (74.0–93.7)	99.8 (99.6–100.0)
4-yr period (Jan 1, 2007–Dec 31, 2010)	78.0 (58.5–92.5)	99.3 (98.9–99.8)	84.0 (74.1–93.6)	99.7 (99.5–99.9)
5-yr period (Jan 1, 2006–Dec 31, 2010)	78.1 (58.5–92.7)	99.2 (98.8–99.8)	85.4 (76.3–94.1)	99.7 (99.5–99.9)
6-yr period (Jan 1, 2005–Dec 31, 2010)	78.1 (58.4–92.7)	99.2 (98.8–99.8)	85.5 (76.3–94.2)	99.6 (99.5–99.9)
7-yr period (Jan 1, 2004–Dec 31, 2010)	78.1 (58.4–92.6)	99.2 (98.7–99.8)	85.0 (75.7–93.9)	99.6 (99.5–99.7)
8-yr period (Jan 1, 2003–Dec 31, 2010)	78.1 (58.6–92.6)	99.2 (98.9–99.6)	85.0 (75.7–93.9)	99.5 (99.4–99.8)
9-yr period (Jan 1, 2002–Dec 31, 2010)	78.1 (58.4–92.7)	99.1 (98.6–99.7)	85.0 (75.6–94.0)	99.5 (99.3–99.8)
10-yr period (Jan 1, 2001–Dec 31, 2010)	78.1 (58.4–92.7)	99.0 (98.5–99.7)	85.0 (76.3–94.1)	99.5 (99.3–99.8)
11-yr period (Jan 1, 2000–Dec 31, 2010)	78.2 (58.7–92.7)	99.0 (98.5–99.7)	85.4 (76.3–94.1)	99.4 (99.2–99.8)
12-yr period (Jan 1, 1999–Dec 31, 2010)	78.1 (58.5–92.7)	98.9 (98.4–99.7)	85.4 (76.2–94.0)	99.4 (99.2–99.8)
13-yr period (Jan 1, 1998–Dec 31, 2010)	78.2 (58.8–92.7)	98.9 (98.3–99.7)	85.5 (76.3–94.1)	99.4 (99.1–99.8)

Values are posterior means (lower and upper limits of the posterior equal-tailed 95% credible intervals).

RAMQ data alone was unchanged (78%) for all observation windows. However, complementing the RAMQ billing codes case ascertainment method with self-reported data sources on RA diagnosis and current DMARD use increased the point estimate for sensitivity from 78.1% (95% CrI 58.3–92.6) to 84.0% (95% CrI 74.0–93.7) for the shortest time window. Our estimates of the sensitivity of RAMQ data versus the self-reported data remained relatively steady over time. The specificity of RAMQ ascertainment method alone as well as combining it with self-reported data was high (99%) and stable throughout all time windows.

DISCUSSION

In this study, a series of Bayesian latent class models were developed to assess the effects of 3 factors (i.e., the length of observation window within administrative data, the inclusion of self-reported information on RA, and adjustment for misclassification error in administrative data) on RA prevalence estimates in the CARTaGENE sample. Our results show variations in the prevalence point estimates related to all 3 factors. There was negligible change in the sensitivity estimates for case ascertainment using administrative data with more years of observation, but a noticeable gain in sensitivity when additional information from self-reported information on RA diagnosis and current DMARD use were added to the model. The 3-year 2010 point prevalence estimate among adults aged 40–69 years using the 3 ascertainment methods and adjusting for misclassification error in each method was 0.9% (95% CrI 0.7–1.2).

Previous studies of the effect of increasing years of administrative data on rheumatic diseases prevalence estimates found trends similar to ours (i.e., higher prevalence estimates with more years of data)[6,7,22,23,24]. However, ours is the only one that adjusted for the imperfect data sources. As evident from our study, the inclusion of self-reported RA data reduced the trend for incomplete ascertainment with few years of administrative data. RA is a dynamic chronic disease, characterized by unpredictable flares and remissions of disease activity[25]. During periods of remission, patients may not seek medical treatment, at least for RA. So, extracting ICD codes for a short observation window in RAMQ may miss some cases, specifically those patients in remission or with mild disease activity who happen not to use health services in the years under observation. Since 1 diagnostic code is allowed per physician visit in Quebec, RA patients with comorbidities may escape detection based on ICD codes within short observation windows if the code reported by the physician is for comorbidity and not RA.

Ng, *et al* studied the effect of the number of years of administrative data observed on estimates of SLE prevalence and recommended the use of long time windows to avoid underascertainment[6]. However, using longer observation windows could lead to overestimation of RA prevalence if misclassification error is not accounted for. This highlights the importance of carefully thinking about both sensitivity and specificity. Moreover, using longer time windows within health administrative databases has some drawbacks when the interest is in more recent prevalence estimates because temporal changes such as diagnostic drift have occurred over time[26]. For example, the American College of Rheumatology (ACR) criteria for RA diagnosis have changed 3 times in the last 50 years[27]. The most recent are the 2010 ACR/European League Against Rheumatism classification criteria[28]. These changes in diagnostic criteria could alter RA prevalence estimates when longer time windows are analyzed.

The sensitivity of case ascertainment using administrative data alone was about 78% and remained steady throughout all time windows in our study. Supplementing administrative data with patient self-reported RA diagnosis and current use of DMARD increased the point estimate for sensitivity to about 85% (although CrI overlapped). This finding may be important for investigators who may have access to only a few years of administrative data, if they have additional sources of information on RA status. The importance of using

multiple data sources is corroborated by recommendations from other researchers working on chronic disease surveillance[26,29,30]. In the absence of other data sources, lengthening the number of years of RAMQ data increases RA prevalence point estimates, but with overlapping CrI across all observation windows.

One potential limitation in our study is the use of current DMARD consumption as an ascertainment method. Prior DMARD use was not available in the data. If ever DMARD use was assessed instead, then a better identification of RA cases (i.e., increase in the sensitivity estimate) would have been likely with the 3 ascertainment methods. Current DMARD use identifies only those with active disease. Although the low sensitivity of this ascertainment method was accounted for in the prior distribution, it is possible that accounting for ever DMARD use would have improved the collection of RA cases and further reduced the misclassification error by identifying those who were in remission during the survey.

Additionally, our adjusted results using health administrative data alone were not that precise even with such a large sample size. The difficulty in getting accurate prior information on the sensitivity and specificity can affect the precision of the posterior intervals. However, the precision was improved with additional information on RA status from self-reported data.

In our study, we did not use hospitalization RA codes. In fact, the Canadian working group on rheumatic disease definitions for surveillance using administrative data has done analyses of billing data with or without hospitalization data, and their consensus (based on analyses from each province) was that hospitalization data does not increase sensitivity of RA ascertainment.

The strengths of our study were the use of a very large cohort of individuals with both self-reported and administrative data on RA. Both data sources were adjusted for misclassification error in the absence of gold standard, which reflects a real-life challenge because few RA ascertainment approaches are considered 100% accurate. To the authors' knowledge, this is the first study to date to combine self-reported data and Canadian provincial health administrative data to estimate an adjusted RA prevalence.

Our study illustrates that when using administrative data, RA point prevalence estimates are lower if few years of data are observed, and that multiple data sources can help identify more RA cases.

REFERENCES

1. Thacker SB, Stroup DF, Rothenberg RB. Public health surveillance for chronic conditions: a scientific basis for decisions. Stat Med 1995;14:629-41.
2. Gordis L. The occurrence of disease: I. Disease surveillance and measures of morbidity. In: Epidemiology. Fifth ed. Philadelphia: Elsevier Saunders; 2014:49-52.
3. Bernatsky S, Dekis A, Hudson M, Pineau CA, Boire G, Fortin PR, et al. Rheumatoid arthritis prevalence in Quebec. BMC Res Notes 2014;7:937.
4. Malone DC, Billups SJ, Valuck RJ, Carter BL. Development of a chronic disease indicator score using a Veterans Affairs Medical Center medication database. J Clin Epidemiol 1999;52:551-7.
5. Widdifield J, Labrecque J, Lix L, Paterson JM, Bernatsky S, Tu K, et al. Systematic review and critical appraisal of validation studies to identify rheumatic diseases in health administrative databases. Arthritis Care Res 2013;65:1490-503.
6. Ng R, Bernatsky S, Rahme E. Observation period effects on estimation of systemic lupus erythematosus incidence and prevalence in Quebec. J Rheumatol 2013;40:1334-6.
7. Nightingale A, Farmer R, de Vries CS. Systemic lupus erythematosus prevalence in the UK: methodological issues when using the general practice research database to estimate frequency of chronic relapsing-remitting disease. Pharmacoepidemiol Drug Saf 2007;16:144-51.
8. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, et al; CARTaGENE Project. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol 2013;42:1285-99.
9. Chaaya M, Slim ZN, Habib RR, Arayssi T, Dana R, Hamdan O, et al. High burden of rheumatic diseases in Lebanon: A COPCORD study. Int J Rheum Dis 2012;15:136-43.
10. Gariepy G, Rossignol M, Lippman A. Characteristics of subjects self-reporting arthritis in a population health survey: distinguishing between types of arthritis. Can J Public Health 2009;100:467-71.
11. Centers for Disease Control and Prevention (CDC). Prevalence of self-reported arthritis or chronic joint symptoms among adults—United States, 2001. MMWR Morb Mortal Wkly Rep 2002;51:948-50.
12. Walitt BT, Constantinescu F, Katz JD, Weinstein A, Wang H, Hernandez RK, et al. Validation of self-report of rheumatoid arthritis and systemic lupus erythematosus: The Women's Health Initiative. J Rheumatol 2008;35:811-8.
13. Slim Z. Estimating rheumatoid arthritis prevalence and care quality in a large sample from the Quebec population [dissertation]. Montreal: McGill University; 2018:111 pp.
14. Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11:iii, ix-51.
15. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard—a systematic review. Am J Epidemiol 2014;179:423-31.
16. Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev Vet Med 2005;68:19-33.
17. Enøe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev Vet Med 2000;45:61-81.
18. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 1995;141:263-72.
19. Widdifield J, Bombardier C, Bernatsky S, Paterson JM, Green D, Young J, et al. An administrative data validation study of the accuracy of algorithms for identifying rheumatoid arthritis: the influence of the reference standard on algorithm performance. BMC Musculoskelet Disord 2014;15:216.
20. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001;57:158-67.
21. Weichenthal S, Joseph L, Bélisle P, Dufresne A. Bayesian estimation

of the probability of asbestos exposure from lung fiber counts. Biometrics 2010;66:603-12.
22. Wiréhn A-BE, Karlsson HM, Carstensen JM. Estimating disease prevalence using a population-based administrative healthcare database. Scand J Public Health 2007;35:424-31.
23. Powell KE, Diseker RA, Presley RJ, Tolsma D, Harris S, Mertz KJ, et al. Administrative data as a tool for arthritis surveillance: estimating prevalence and utilization of services. J Public Health Manag Prac 2003;9:291-8.
24. Kopec JA, Rahman MM, Berthelot JM, Le Petit C, Aghajanian J, Sayre EC, et al. Descriptive epidemiology of osteoarthritis in British Columbia, Canada. J Rheumatol 2007;34:386-93.
25. Kvien TK. Epidemiology and burden of illness of rheumatoid arthritis. Pharmacoeconomics 2004;2 Suppl 1:1-12.
26. Ward MM. Estimating disease prevalence and incidence using administrative data: some assembly required. J Rheumatol 2013;40:1241-3.
27. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
28. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010; 62:2569-81.
29. Cricelli C, Mazzaglia G, Samani F, Marchi M, Sabatini A, Nardi R, et al. Prevalence estimates for chronic diseases in Italy: Exploring the differences between self-report and primary care databases. J Public Health Med 2003;25:254-7.
30. Bernatsky S, Lix L, Hanly J, Hudson M, Badley E, Peschken C, et al. Surveillance of systemic autoimmune rheumatic diseases using administrative data. Rheumatol Int 2011;31:549-54.