Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 61, No. 3, March 15, 2009, pp 400–404
DOI 10.1002/art.24339

ORIGINAL ARTICLE

Scleroderma Prevalence: Demographic Variations in a Population-Based Sample

S. BERNATSKY,[1] L. JOSEPH,[1] C. A. PINEAU,[1] P. BELISLE,[1] M. HUDSON,[2] AND A. E. CLARKE[1]

Objective. **To estimate the prevalence of systemic sclerosis (SSc) using population-based administrative data, and to assess the sensitivity of case ascertainment approaches.**

Methods. **We ascertained SSc cases from Quebec physician billing and hospitalization databases (covering ~7.5 million individuals). Three case definition algorithms were compared, and statistical methods accounting for imperfect case ascertainment were used to estimate SSc prevalence and case ascertainment sensitivity. A hierarchical Bayesian latent class regression model that accounted for possible between-test dependence conditional on disease status estimated the effect of patient characteristics on SSc prevalence and the sensitivity of the 3 ascertainment algorithms.**

Results. **Accounting for error inherent in both the billing and the hospitalization data, we estimated SSc prevalence in 2003 at 74.4 cases per 100,000 women (95% credible interval [95% CrI] 69.3–79.7) and 13.3 cases per 100,000 men (95% CrI 11.1–16.1). Prevalence was higher for older individuals, particularly in urban women (161.2 cases per 100,000, 95% CrI 148.6–175.0). Prevalence was lowest in young men (in rural areas, as low as 2.8 cases per 100,000, 95% CrI 1.4–4.8). In general, no single algorithm was very sensitive, with point estimates for sensitivity ranging from 20–73%.**

Conclusion. **We found marked differences in SSc prevalence according to age, sex, and region. In general, no single case ascertainment approach was very sensitive for SSc. Therefore, using data from multiple sources, with adjustment for the imperfect nature of each, is an important strategy in population-based studies of SSc and similar conditions.**

INTRODUCTION

Systemic sclerosis (SSc; also called scleroderma) is a multisystem disease characterized by tissue thickening and fibrosis, often with involvement of internal organs. Prevalence estimates vary widely, from 7 cases per million to 489 cases per million (1). Reported prevalence estimates in North America have varied from 13.8 cases per 100,000 from 1950–1979 (2) to 28.6 cases per 100,000 in 1985 (3). These variations may reflect several issues. There are probably true variations in SSc prevalence over time, space, and demographics (e.g., age and race). However, differences in case ascertainment methods or disease classification systems can also have a large effect on estimates (2–8).

To our knowledge, there is little research to date concerning the accuracy of administrative data for studies of complex systemic autoimmune rheumatic diseases such as scleroderma. Therefore, the objective of our study was to estimate SSc prevalence using population-based administrative data, and to assess the sensitivity of various case ascertainment approaches. We used recently developed methods (9) to adjust for possible misclassification within the administrative data sources. Our research was approved by the McGill University Ethics Review Board and by Quebec's Commission d'accès à l'information.

MATERIALS AND METHODS

Our data sources were the databases for the hospitalization (Ministry of Health's Maintenance et Exploitation des Données pour l'Étude de la Clientèle Hospitalière [MEDECHO]) and physician billing (Régie d'Assurance

Supported by the Canadian Institutes for Health Research. Dr. Bernatsky's work was supported by the McGill University Health Centre Department of Medicine and Research Institute and Fonds de Recherche en Santé du Québec. Dr. Bernatsky is recipient of a Canadian Institutes of Health Research Junior Investigator award and operational grant and a Canadian Arthritis Network Scholar award. Drs. Joseph and Clarke are recipients of Fonds de la Recherche en Santé du Quebec National Scholar awards. Dr. Pineau's work was supported by the McGill University Health Center Department of Medicine and Research Institute. Dr. Hudson is recipient of a Canadian Arthritis Network Scholar award and a Canadian Institutes of Health Research Career award.

[1]S. Bernatsky, MD, PhD, L. Joseph, PhD, C. A. Pineau, MD, FRCPC, P. Belisle, MSc, A. E. Clarke, MD: McGill University Health Centre, Montreal, Quebec, Canada; [2]M. Hudson, MD, MPH: McGill University, Montreal, Quebec, Canada.

Address correspondence to S. Bernatsky, MD, PhD, Division of Clinical Epidemiology, Research Institute of the McGill University Health Centre, 687 Pine Avenue West, V-Building, Montreal, Quebec, H3A 1A1, Canada. E-mail: sasha.bernatsky@mail.mcgill.ca.

Submitted for publication July 18, 2008; accepted in revised form December 3, 2008.

Maladie du Québec [RAMQ]) of all residents of the province of Quebec (~7.5 million individuals). We used data from 1989–2003 to determine the prevalence of existing cases in 2003. The MEDECHO database maintains data on hospitalization dates and discharge diagnoses (a primary diagnosis and ≤15 nonprimary diagnoses per hospitalization). All discharge diagnoses are abstracted from the chart by medical records clerks and are not necessarily the same as the diagnoses recorded in the RAMQ database (which are based on independent claims information). The RAMQ database records information on physician services in the province, including the date and diagnostic code relevant to each physician visit (a single diagnostic code is allowed per visit). For both billing and hospitalization data, diagnoses are provided as International Classification of Diseases, Ninth Revision (ICD-9) codes.

In the hospitalization data, we defined an SSc case as any hospitalization including an ICD-9 code of 710.1 (scleroderma) as a primary or nonprimary discharge diagnosis. In the billing data, cases were first defined according to an algorithm requiring ≥2 SSc diagnostic codes (by any physician) ≥2 months apart but within a 2-year span. A second alternative algorithm defined a case on the basis of ≥1 SSc diagnostic codes during a visit to a rheumatologist.

The sensitivity and specificity of case ascertainment in administrative databases is currently a focus of interest (10). We have recently developed methods to generate prevalence estimates from physician billing and hospitalization databases, adjusting for the imperfect sensitivity and specificity of these administrative data sources. These methods involve a Bayesian latent class model that does not rely on a gold standard; the 3 parameters of interest (prevalence, sensitivities, and specificities of the 3 different case ascertainment methods) cannot be observed directly, but must be estimated statistically.

The Bayesian statistical approach is based on the idea that uncertainty about the unknown parameters can be represented by probability distributions. One begins with prior probability distributions, summarizing all previous relevant information about all unknown parameters of interest (7 in our study, as indicated above). The prior distribution is then updated by new data through the likelihood function. By combining information in the prior distribution with the data, one obtains a posterior distribution, which represents what one should now believe about the parameter values given the initial background and the new data. The methodology is underpinned by Bayes theorem, a mathematical rule for updating prior beliefs in the light of new data (11). This approach is one way to incorporate information from imperfect data sources to produce parameter estimates that adjust for the imperfections in each test.

In our current work, our model must account for possible statistical dependence between our ascertainment methods. This requires informative prior input for at least 2 of these 3 parameters. Based on previous work on case ascertainment using administrative data (12), we expected the specificities of all methods to be very high. Therefore, for our primary analyses we set informative beta (89, 1.2) prior distributions for the specificities of our 3 case ascertainment approaches. This prior distribution corresponds to specificities of 98% (potential values ranging from approximately 96–100%). In sensitivity analyses, we used a different beta prior distribution corresponding to specificities of 99% (potential values ranging from approximately 98–100%).

We then developed a latent class Bayesian hierarchical regression model to provide estimates of disease prevalence and the sensitivities of case ascertainment, and to assess the effects of patient characteristics on prevalence. The first level of the model accounted for sampling variability in prevalence, and for errors in each of the 3 case definitions. These were represented by binomial distributions in which the probability of a positive test includes terms for the sensitivity and specificity of each method of ascertainment. We also added a term to estimate the possible dependence of the 2 billing data algorithms. The second level of the model accounted for variations in prevalence according to patient demographics (age, sex, and rural versus urban residence), which were derived from a logistic regression model on the binomial probabilities from the first level. The third level of the hierarchical model accounted for variation in the sensitivity of case ascertainment according to patient demographics (age, sex, and rural versus urban residence), again derived from a logistic regression model, this time on the sensitivities.

For our estimates, we constructed 95% credible intervals (95% CrI), which represent the values between which there is a 95% probability of containing the parameter of interest, given the data and the prior information used. All programming was carried out using WinBUGS software (MRC Biostatistics Unit, University of Cambridge, Cambridge, UK). WinBUGS uses a Gibbs sampler algorithm, which is an iterative algorithm that draws random samples from the marginal posterior distributions of all parameters of interest. To ensure convergence of the algorithm, we discarded 5,000 initial burn-in iterations, followed by 100,000 iterations that were used for inferences. To further verify that the Gibbs sampler converged, we ran each analysis several times and carefully reviewed the output for each parameter for uniformity across runs.

RESULTS

The use of different choices of prior distributions all resulted in very similar posterior densities, so we present only the primary results, where the prior corresponds to specificities of 98%. Accounting for error inherent in both the billing and the hospitalization data, we estimated SSc prevalence in the province of Quebec in 2003 to be 44.3 cases per 100,000 (95% CrI 41.1–47.6). Prevalence differed greatly between sexes, with 74.4 cases per 100,000 women (95% CrI 69.3–79.7) and 13.3 cases per 100,000 men (95% CrI 10.2–14.8). Prevalence was higher for older individuals, particularly in urban women, for whom the prevalence was 161.2 cases per 100,000 (95% CrI 148.6–175.0). Prevalence was lowest in young men (as low as 2.8 cases per 100,000, 95% CrI 1.4–4.8 in rural areas). In general, no single algorithm was very sensitive, with point estimates for sensitivity ranging from 20% (for rheumatology billing in rural men) to 73% (for physician billing data in young rural men).

Table 1. Effects of demographics on SSc prevalence and case ascertainment sensitivity estimates: Bayesian latent class hierarchical model*

	Adjusted OR	95% CrI†
Effects on prevalence in 2003		
Female	5.0‡	1.4–9.5‡
Age >45 years	6.8‡	1.8–13.1‡
Urban vs. rural residence	1.2	0.6–2.4
Interaction, age, and sex	34.9‡	7.1–75.7‡
Effects on sensitivity		
Hospitalization data§		
Female	0.7	0.4–1.4
Age >45 years	0.9	0.4–1.6
Urban vs. rural residence	0.6‡	0.4–0.9‡
Interaction, age, and sex	1.1	0.5–2.1
Physician data¶		
Female	0.9	0.4–1.9
Age >45 years	0.8	0.4–1.6
Urban vs. rural residence	0.8	0.5–1.5
Interaction, age, and sex	1.0	0.3–2.3
Rheumatologist data#		
Female	1.2	0.6–2.1
Age >45 years	0.9	0.4–1.6
Urban vs. rural residence	1.5‡	1.0–2.2‡
Interaction, age, and sex	1.3	0.6–2.5

* SSc = systemic sclerosis; OR = odds ratio; 95% CrI = 95% credible interval.
† Bayesian CrI represents the values between which there is a 95% probability of containing the parameter of interest, given the data and prior information input.
‡ Significant values.
§ Case ascertainment based on ≥1 hospitalization discharge diagnostic codes (primary or nonprimary).
¶ Algorithm based on ≥2 diagnostic codes for SSc (≥8 weeks apart and within 2 years) contributed by any physician.
#Algorithm based on ≥1 diagnostic codes for SSc (International Classification of Diseases, Ninth Revision code 710.1) contributed by a rheumatologist.

The effects, in terms of adjusted odds ratios, of demographics on SSc prevalence and on the sensitivity of hospitalization data and billing data for SSc ascertainment in the population, using information from both sources and not assuming a gold standard, are presented in Table 1. The first part of Table 1 illustrates not only the higher prevalence in women and in older individuals, but also the interaction, whereby prevalence was particularly high in older women compared with all other groups. The rest of Table 1 shows the sensitivities of the 3 methods of case ascertainment, indicating that the sensitivity of hospitalization data for SSc case ascertainment was higher in rural (versus urban) areas; conversely, the sensitivity of rheumatology diagnostic codes was lower in rural areas.

DISCUSSION

Our prevalence estimates were higher than results from the US, including recent work by Mayes et al (5), in which SSc prevalence for the Detroit tricounty metropolitan area, using data from 1989–1991, was estimated at 27.6 (95% confidence interval 24.5–31.0) cases per 100,000. In fact, our estimates are more in line with the estimates of Maricq et al, whose study is, to our knowledge, the one truly population-based SSc prevalence study that has been performed (3).

There has been some suggestion of increased SSc prevalence over time. Increased survival may be one reason for this; in one study, patients in a post-1985 cohort had significantly better 10-year survival compared with a pre-1985 cohort (13). Other data suggests this same phenomenon (5,7). Increased survival over time could be related to increasing recognition of milder cases of SSc, or to better management of life-threatening complications of the disease, particularly renal crisis and pulmonary hypertension. Another reason for the relatively high prevalence estimate in our study was that we ascertained cases over a fairly lengthy period (10 years). For example, the work of Mayes et al retrospectively ascertained SSc cases using multiple sources of data, but the period of ascertainment spanned only 3 years. In addition, we accounted for the imperfect sensitivities of the ascertainment methods; this, in reducing the false-negative rate, increases estimated prevalence.

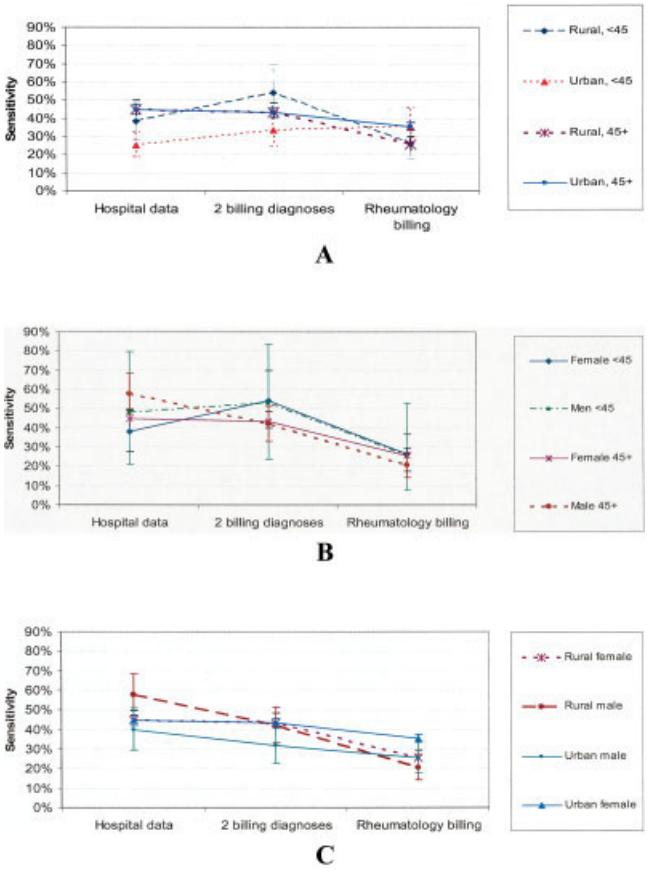

Figure 1. Estimates of the sensitivity of 3 different case ascertainment approaches, **A,** women only, **B,** rural residents only, and **C,** older individuals only, without considering any one as a gold standard. Error bars represent Bayesian credible intervals. Case ascertainment based on ≥1 hospitalization discharge diagnostic codes (primary or nonprimary), ≥2 diagnostic codes for systemic sclerosis (SSc; International Classification of Diseases, Ninth Revision code 710.1) ≥8 weeks apart and within 2 years contributed by any physician, and ≥1 diagnostic codes for SSc contributed by a rheumatologist.

Use of administrative databases has both benefits and potential limitations. One difficulty with the use of administrative data, such as physician billing, is the possibility of imprecisely used diagnostic codes leading to compromises in specificity. On the other hand, clinical studies may rely on very specific classification systems such as American College of Rheumatology criteria, but such criteria may actually exclude many patients diagnosed with SSc by experienced clinicians (14,15). This illustrates the universal phenomenon that increasing specificity decreases sensitivity.

Medical chart review has been held as a gold standard for diagnosis, but charts are frequently unavailable (more than 25% in the Mayes et al article) or the documentation of clinical features may be too incomplete to allow diagnosis (5). Our case ascertainment methods using administrative data represent one way to avoid nonrandomly missing data that may exist in chart review studies. The price for this is some lack of specificity. If assembling a cohort in this way includes a significant number of persons without the disease of interest (in our case, SSc), one might hypothesize that attempts to study outcomes, such as morbidity or mortality, might be biased toward the null value. However, it is difficult to predict the exact direction or magnitude of such effects.

Comparison of different case definitions for scleroderma and consideration of different data sources (physician billing versus hospitalization databases) leads to the following observations. First, as others have shown, it is clear that for epidemiologic studies of SSc, a single approach to case ascertainment is likely to miss a substantial number of cases. For example, in a regional assessment of scleroderma prevalence and incidence Mayes et al showed that a search of hospital discharge records recorded most, but not all, cases in their catchment area (5). Like Mayes et al, Alamanos et al (16) used both inpatient and outpatient data in an attempt to improve completeness of capture, but in addition Mayes et al used a third source (a limited registry kept by a scleroderma support group) and estimated, by capture-recapture methods, that even after combining 3 data sources, 12% of scleroderma cases would be missed.

We suspect that reliance on hospitalization records alone might introduce considerable bias, because a cohort assembled in this way is potentially more likely to capture persons with the most severe SSc. Furthermore, hospitalized patients tend to have more comorbidity than the general population, so epidemiologic studies of this kind of cohort, compared with the general population, might produce biased results. Just as importantly, when using hospitalization data alone there may not be a good record of events that occurred prior to hospitalization; even data on disease onset may be inaccurate. On the other hand, using physician billing data alone can produce limitations when, as in our case, a physician is only allowed one diagnostic field per visit. This means that a patient who has multiple comorbid conditions may escape detection with a case ascertainment approach that relies on physician billing if the physician following the patient tends to use the diagnostic code for a comorbid state (e.g., pulmonary hypertension) instead of the underlying condition of interest to the researchers (in our case, SSc). Combining administrative data sources is thus an attractive solution.

One remaining limitation of our work is that the ICD-9 coding does not differentiate between limited and diffuse SSc. This is important because these 2 subsets may have different case ascertainment properties that will not be captured by our methods. Furthermore, attempting to describe the epidemiology and outcomes of the 2 SSc subgroups separately is problematic when relying on administrative data diagnoses based on ICD-9 coding.

From the work by Mayes et al, it appears that any given method of case ascertainment will result in false-positives as well as false-negatives; if these cancel out, a prevalence estimate may be close to the actual value, even using imperfect methods. However, a cohort assembled based on an imperfect ascertainment method (which is always the case) will never contain all actual cases in a given population. Reliance on primarily administrative database sources may be a valid option for some epidemiologic studies, but researchers, and their audiences, should keep in mind that all data sources are imperfect.

AUTHOR CONTRIBUTIONS

Dr. Bernatsky had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study design. Bernatsky, Joseph.
Acquisition of data. Bernatsky, Clarke.
Analysis and interpretation of data. Bernatsky, Joseph, Pineau, Belisle, Hudson, Clarke.
Manuscript preparation. Bernatsky, Joseph, Pineau, Belisle, Hudson, Clarke.
Statistical analysis. Joseph, Belisle.

REFERENCES

1. Chifflot H, Fautrel B, Sordet C, Chatelus E, Sibilia J. Incidence and prevalence of systemic sclerosis: a systematic literature review. Semin Arthritis Rheum 2008;37:223–35.
2. Michet CJ, McKenna CH, Elveback LR, Kaslow RA, Kurland LT. Epidemiology of systemic lupus-erythematosus and other connective-tissue diseases in Rochester, Minnesota, 1950 through 1979. Mayo Clin Proc 1985;60:105–13.
3. Maricq HR, Weinrich MC, Keil JE, Smith EA, Harper FE, Nussbaum AI, et al. Prevalence of scleroderma spectrum disorders in the general population of South Carolina. Arthritis Rheum 1989;32:998–1006.
4. Steen VD, Oddis CV, Conte CG, Janoski J, Casterline GZ, Medsger TA Jr. Incidence of systemic sclerosis in Allegheny County, Pennsylvania: a twenty-year study of hospital-diagnosed cases, 1963–1982. Arthritis Rheum 1997;40:441–5.
5. Mayes MD, Lacey JV Jr, Beebe-Dimmer J, Gillespie BW, Cooper B, Laing TJ, et al. Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. Arthritis Rheum 2003;48:2246–55.
6. Medsger TA, Masi AT, Rodnan GP, Benedek TG, Robinson H. Survival with systemic sclerosis (scleroderma): life-table analysis of clinical and demographic factors in 309 patients. Ann Intern Med 1971;75:369–76.
7. Medsger TA Jr, Masi AT. The epidemiology of systemic sclerosis (scleroderma) among male U.S. veterans. J Chronic Dis 1978;31:73–85.
8. Kurland LT, Hauser WA, Ferguson RH, Holley KE. Epidemiologic features of diffuse connective tissue disorders in

Rochester, Minn, 1951 through 1967, with special reference to systemic lupus erythematosus. Mayo Clin Proc 1969;44:649–63.
9. Ladouceur M, Rahme E, Pineau CA, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. Biometrics 2007;63:272–9.
10. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004;57:131–41.
11. Ashby D. Bayesian statistics in medicine: a 25 year review. Stat Med 2006;25:3589–631.
12. Bernatsky S, Joseph L, Pineau CA, Tamblyn R, Feldman DE, Clarke AE. A population-based assessment of systemic lupus erythematosus incidence and prevalence: results and implications of using administrative data for epidemiological studies. Rheumatology (Oxford) 2007;46:1814–8.
13. Ferri C, Valentini G, Cozzi F, Sebastiani M, Michelassi C, la Montagna G, et al. Systemic sclerosis: demographic, clinical, and serologic features and survival in 1,012 Italian patients. Medicine (Baltimore) 2002;81:139–53.
14. Lonzetti LS, Joyal F, Raynauld JP, Roussin A, Goulet JR, Rich E, et al. Updating the American College of Rheumatology preliminary classification criteria for systemic sclerosis: addition of severe nailfold capillaroscopy abnormalities markedly increases the sensitivity for limited scleroderma [letter]. Arthritis Rheum 2001;44:735–6.
15. Giordano M, Valentini G, Migliaresi S, Picillo U, Vatti M. Different antibody patterns and different prognoses in patients with scleroderma with various extent of skin sclerosis. J Rheumatol 1986;13:911–6.
16. Alamanos Y, Tsifetaki N, Voulgari PV, Siozos C, Tsamandouraki K, Alexiou GA, et al. Epidemiology of systemic sclerosis in northwest Greece 1981 to 2002. Semin Arthritis Rheum 2005;34:714–20.

DOI 10.1002/art.24478

Submissions Invited for Themed Issue of *Arthritis Care & Research*: Drug Safety in the Rheumatic Diseases

Arthritis Care & Research is soliciting manuscripts for a themed issue addressing drug safety in the treatment of rheumatic diseases, including but not limited to biologic agents. Manuscripts covering a broad range of topics related to the major theme are invited; for example, update on safety issues related to a specific drug or biologic agent, issues related to classes of treatments (e.g., anti–tumor necrosis factors [anti-TNFs]) and types of events (e.g., opportunistic infections in patients receiving anti-TNF agents), and issues related to different methodologies for assessing safety. Submissions may also describe more general issues related to treatment safety such as new or evolving methods of assessing or discussing safety, or benefit or safety/benefit ratio with patients. Manuscripts from a wide range of disciplines relevant to safety are welcome.

The issue will include regular submission as well, but a certain number of pages will be reserved for manuscripts accepted in response to this solicitation. Manuscripts will be subject to the usual review process and all types of manuscripts (e.g., original articles, contributions from the field, case studies, trainee rounds, reviews) are included in this solicitation.

The deadline for submission is October 1, 2009. For further information, contact the editors of *Arthritis Care & Research*, Edward H. Yelin, PhD (Ed.Yelin@ucsf.edu) or Patricia P. Katz, PhD (Patti.Katz@ucsf.edu).