Rheumatology 2007;46:1814–1818 doi:10.1093/rheumatology/kem233

A population-based assessment of systemic lupus erythematosus incidence and prevalence—results and implications of using administrative data for epidemiological studies

S. Bernatsky[1,2], L. Joseph[1,3], C. A. Pineau[2], R. Tamblyn[3,4], D. E. Feldman[5] and A. E. Clarke[1,6]

Objectives. To estimate (i) systemic lupus erythematosus (SLE) incidence and prevalence using multiple sources of population-based administrative data; (ii) the sensitivity and specificity of case ascertainment methods; and (iii) variation in performance of each ascertainment approach, according to patient and physician characteristics.

Methods. We examined the physician billing and hospitalization databases of the province of Quebec (1994–2003) covering all health care beneficiaries (~7.5 million). We compared various approaches to ascertain SLE cases, using information from each database separately or combining sources; we then estimated the sensitivity and specificity of these alternative approaches. We used regression models to determine if sensitivity was independently influenced by patient or physician characteristics.

Results. Using billing data, we calculated SLE incidence at 3.0/100 000 person-years [95% confidence interval (CI) 2.6–3.4]; prevalence was 32.8/100 000 persons, in 2003. Results were similar using hospitalization data. However, only a proportion of prevalent cases were identified as having SLE by both methods. Combining cases from billing and hospitalization data, we found a prevalence of 51/100 000 in 2003. Our latent class regression model estimated a prevalence of 44.7/100 000 (95% CI 37.4–54.7). We found high specificity for SLE diagnoses across all strategies and data sources; sensitivity ranged from 42.1% to 67.6%, and was independently influenced by both patient and physician characteristics.

Conclusions. In observational studies, particularly with administrative databases, SLE incidence and prevalence estimates differ considerably, according to the approach for case ascertainment. In the absence of gold standards, statistical modelling can provide sensitivity and specificity estimates for different approaches.

Key words: Systemic lupus erythematosus, Incidence, Prevalence, Epidemiology.

In the rheumatic diseases, as in other conditions, administrative databases are being used increasingly for outcomes research. Unfortunately, such research has often relied on diagnostic algorithms for a condition without adequate acknowledgement of their limitations [1–3]. Recently, there have been strong recommendations for more validation studies of administrative sources such as physician claims databases [4, 5]. Importantly, there has been almost no research concerning the accuracy of diagnoses based on administrative data for studies of complex conditions such as systemic lupus erythematosus (SLE).

The objectives of our study were to estimate: (i) SLE incidence and prevalence in a population-based sample, using multiple sources of administrative data; (ii) the sensitivity and specificity of SLE case ascertainment approaches based on administrative data; and (iii) The variation in performance of each ascertainment approach, according to patient demographics and physician characteristics.

Methods

Our study used administrative data from the Régie d'assurance Maladie du Québec (RAMQ) and the Ministry of Health's Maintenance et Exploitation des Données pour l'Étude de la Clientèle Hospitalière (MEDECHO). For residents of the province of Quebec (~7.5 million individuals as of 2004 [6]), health service coverage is universal, although regulations require a waiting period of 3 months of residence before enrolment.

The RAMQ demographic file includes birthdate, sex, postal code and year of death for all beneficiaries. The RAMQ physician claims database documents physician services for all provincial beneficiaries including date, diagnosis and, when applicable, procedure codes. The MEDECHO hospital database captures data on hospital admissions for the province, including dates and discharge diagnoses (a primary diagnosis and up to 15 non-primary 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 physician claims database (which are based on independent billing information).

Information on physician characteristics (age, sex, practice location, specialty, year of graduation) is also maintained by RAMQ. The RAMQ and MEDECHO databases are linkable through the personal health insurance number, which is a 10-digit unique identifier for each beneficiary. These databanks have been validated for research and are extensively used for epideimiological studies [7]. We examined various approaches to ascertainment of incident and prevalent SLE cases in Quebec, based on data from 1989 to 2003. Approval for the study was obtained from the McGill University Ethics Review Board.

For physicians' claims data, we defined SLE cases according to an algorithm requiring a minimum of two claims for a diagnosis of SLE (ICD-9 code 710.0) at least 2 months apart but within a 2-yr span. Annual estimates of incident cases were ascertained over 1994–2003 by reviewing the administrative data from 1989 onward and identifying cases where there had been no previous claim for SLE. Similarly, using the MEDECHO data, we independently identified prevalent and incident SLE cases, by requiring at least one diagnosis (primary or non-primary) of SLE.

We estimated SLE incidence and prevalence, as well as the sensitivity and specificity of the data sources for SLE case

[1]Department of Clinical Epidemiology, McGill University Health Centre (MUHC), [2]Division of Rheumatology, MUHC, [3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, [4]Department of Medicine, McGill University, [5]École de réadaptation, Université de Montréal and [6]Division of Clinical Immunology/Allergy, MUHC.

Submitted 12 April 2007; revised version accepted 2 August 2007.

Correspondence to: S. Bernatsky, 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

ascertainment, using various approaches. This included consideration of information from each of the databases separately, as well as combining the data sources. For example, we considered billing data as a single source to calculate SLE prevalence and then compared this with hospitalization data. Estimates using a single source are presented under the naive assumption that this ascertainment method has no error, even though this assumption is almost surely not correct. This facilitates comparison between what researchers have typically used as estimates in the past and other more sophisticated methods we have used. For each approach, year-specific rates were calculated by dividing the number of SLE cases for each calendar year, by year-specific census figures (obtained from Statistics Canada) for the entire population of Quebec. The incidence rates were calculated for each year between 1994 and 2003 and then averaged; the 95% confidence intervals (CIs) of these estimates represent year-to-year variation. The prevalence estimates were calculated for the year 2003, based on the number of cases that had been identified during the study period (1994–2003) who remained alive as of December 31, 2003, with the denominator being the total number of Quebec residents at that time. No CIs are provided for the naïve prevalence estimates (generated under the assumption of no test error), since the estimates are based on the whole population of Quebec in 2003 and not a sample. Estimates stratified by age group and sex (here, the denominator is from the relevant census data for each demographic group) were also generated.

We generated alternative estimates of these parameters using Bayesian statistical approaches, which do not assume any of the data sources to be a gold standard. Bayesian methods are based on the idea that uncertainty about the unknown parameters can be represented by probability distributions. A Bayesian analysis begins with a prior probability distribution over all unknown parameters of interest. The prior distributions summarize all relevant previous information about these parameters. The prior distribution is then updated by new data, through the likelihood function. The combination of information contained in the prior distribution and the data results in 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 (for a detailed review see [8]). This approach can be used to incorporate information from imperfect data sources to produce parameter estimates that adjust for the imperfections in each test.

Such methods can be useful not only for calculating disease prevalence in the absence of a gold standard, but in estimating the sensitivity and specificity of the imperfect ascertainment methods [9–11]. If there are three methods of case ascertainment, we can estimate the seven unknown parameters (i.e. disease prevalence and the sensitivity and specificity of each of the three methods) without the need for substantive prior information. In other words, in this scenario, we can estimate all unknown parameters using information in the data only. For our purposes, the third source of information used to define a case of SLE (i.e. in addition to billing codes and hospitalization data) was procedure code data. The procedure code combinations were based on some key types of organ involvement in SLE, and their treatment. One code combination was kidney biopsy, plus cytotoxic therapy infusion (i.e. cyclophosphamide). The other was a skin biopsy, plus ophthalmology visual field evaluations (standard monitoring for the anti-malarial drugs, e.g. hydroxychloroquine).

Since the data in our study were found at multiple levels (patient level, physician level, and so on), we used a Bayesian hierarchical model, where the first level of the model took into account individual sampling variability, to calculate estimates for the whole population. The second level allowed variation of the performance of each case ascertainment approach, according to different patient characteristics (i.e. sex, age group and rural-*vs*-urban residence). In a third level, we allowed for the variability of each case ascertainment approach across different physician characteristics (i.e. specialty and year of graduation). A fourth level of the hierarchical model presented priors for all parameters that were not already taken care of by the hierarchy.

For the patient-level parameters from first level of our hierarchy (the incidence and prevalence parameters), we used β-prior distributions. The prior for SLE prevalence was a uniform distribution, $\beta(1, 1)$. This was an 'uninformative' prior, where all possible values for the parameter are equally likely; in this case, the prior does not influence the posterior distribution. For the sensitivities and specificities we also considered uninformative priors with uniform distributions, $\beta(1, 1)$, but alternatively used a $\beta(2, 4)$ distribution for sensitivities and a $\beta(4, 2)$ distribution for the specificities. Estimates were very similar for the two strategies and results are presented using the second strategy. For the higher levels of the hierarchy, year-specific estimates were obtained using logistic regression models where the outcomes were the first-level estimates. We assumed that the logarithm of the odds ratios (ORs) were distributed normally around an overall (across years) mean. Reported ORs are the exponentiated forms of these means. Highly dispersed normal prior distributions were assumed for the logistic regression coefficients.

Results

Based on our case definition of SLE from the physician billing database alone, we calculated an over-all population incidence (based on 219 incident cases) of 3.0 cases per 100 000 person-years (95% CI 2.6–3.4). Using the billing data definition alone we identified 2455 prevalent SLE cases in 2003, for a rate of 32.8 cases per 100 000 persons. Similar estimates were produced using hospital discharge database alone, where we calculated an SLE incidence (based on 203 incident cases) of 2.8 cases per 100 000 person-years (95% CI 2.6–3.0). Using the hospitalization data alone, we identified 2394 prevalent SLE cases in 2003, for a prevalence of 31.9 cases per 100 000. However, only a proportion (23.7%) of cases were identified as SLE by both methods; 43.8% of cases identified with the hospitalization data were not identified by our billing data methods, and 32.4% of cases identified with our billing data methods were not detected with the hospitalization data method. A composite reference standard, considering a subject to be an SLE case on the basis of either the two physician billing claims or a hospitalization, identified 3825 prevalent cases in 2003, for a prevalence of 51.0 cases per 100 000 for the over-all population. Figure 1 gives the SLE incidence and prevalence figures for age and sex groups, based on this composite reference case definition. Our Bayesian latent class model produced a prevalence estimate of 44.7 cases per 100 000 in 2003 (95% CI 37.4–54.7).

Table 1 presents estimates of sensitivity and specificity for methods using either billing codes alone, hospitalization data alone, the composite reference standard or using all three data sources without assuming a gold standard. We found high specificity (99.99, 95% CI 99.98–1.00) for SLE diagnosis across all strategies and data sources; regarding sensitivity, point estimates for physician billing ranged from 44.8% to 56.2%. For hospitalization data, sensitivity point estimates ranged from 42.1% to 67.6%.

In the hierarchical model (Table 2), the sensitivity of case ascertainment methods that relied on physician billing records appeared to be independently influenced by patient characteristics, sensitivity estimates were higher in females, and seemed to increase with age. Adjusting for age and sex, sensitivity was also independently affected by rural-*vs*-urban location; rural residence was associated with lower sensitivity of the case ascertainment methods.

Physician characteristics also influenced sensitivity estimates for case ascertainment methods, with effects seen for both speciality and era of graduation. The sensitivity for case

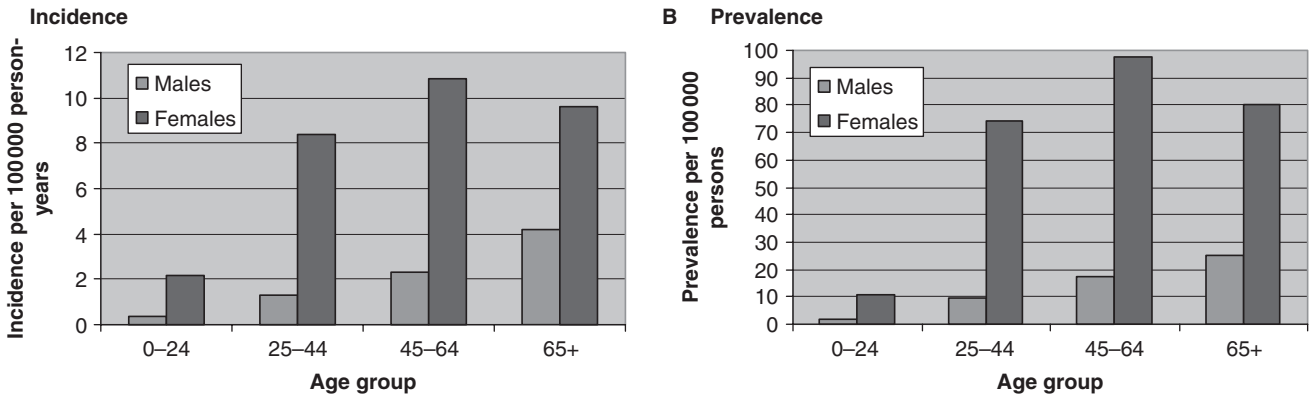

FIG. 1. Incidence (**A**) and prevalence (**B**) of SLE in Quebec 1994–2003, by sex and age group. Cases based on administrative data; included as a case on the basis of either physician billing records (≥2 diagnoses of SLE ≥8 weeks apart) or the provincial hospitalization database (≥1 discharge diagnoses of SLE).

TABLE 1. Sensitivity and specificity of various methods of SLE case ascertainment

	Comparator Sensitivity (95% Confidence/Credible Interval)			
Comparator	Only billing data	Only hospital data	Composite reference	No gold standard
Billing data	[a]	35.1 (34.0, 35.8)	56.2% (55.6, 56.8)	44.8% (43.1, 46.6)
Hospital data	42.1% (41.0, 42.8)	[a]	67.6% (67.0, 68.1)	58.4% (56.2, 60.5)
	Comparator Specificity (95% Confidence/Credible Interval)			
Only billing data	[a]	99.99% (99.98, 99.99)	100%[a]	99.99% (99.99, 100)
Only hospital data	99.99% (99.98, 99.99)	[a]	100%[a]	99.99% (99.99, 100)

Composite reference = case identified either using physician or hospitalization data; specificity of either source is thus 100%.

TABLE 2. Effect of patient and physician characteristics on sensitivity of SLE case ascertainment: estimates from Bayesian hierarchical model

	Odds ratio	95% CI
Patient characteristics (estimates from second level of hierarchy)		
Females		
Aged <25	0.66	0.54, 0.80
25–44	1.00	reference
45–64	1.01	0.97, 1.06
65+	1.41	1.32, 1.50
Males		
Aged <25	0.39	0.16, 0.96
25–44	0.55	0.45, 0.69
45–64	0.52	0.43, 0.64
65+	1.24	1.06, 1.46
Rural versus urban residence	0.64	0.59, 0.70
Physician characteristics (estimates from third level of hierarchy)		
Selected specialities (Rheumatologist, immunologist, nephrologist)	7.33	6.38, 8.42
Physicians graduating after 1980	0.96	0.84, 1.10
With patients aged		
<25	0.53	0.43, 0.64
25–44	1.28	1.1, 1.49
45–64	0.65	0.59, 0.72
65+	0.24	0.22, 0.26

ascertainment methods was highest when we considered only specialists who traditionally take on the care of SLE patients (rheumatologists, immunologists and nephrologists), adjusted for patient characteristics and physician graduation year. There was some evidence of lower sensitivity of SLE case ascertainment methods in the physician group graduating after 1980, but there appeared to be an interaction such that this effect was only apparent with certain patient age groups, i.e. those aged less than 25 yrs or those aged 45 or older (both men and women).

Discussion

SLE is an important autoimmune disorder occurring most frequently in women of childbearing age. Most of the data regarding SLE incidence and prevalence in North America are based on data from the USA. Much of this was published for much earlier eras (25–50 yrs ago), although there are a few US studies on unselected populations published in the last 10 yrs [12–15].

Our incidence estimates using a single source of data (either physician billing or hospitalization data) resembled the results of one of these recent studies (where the incidence rate was about 3 cases per 100 000 person-years) [13]. However, in two other relatively recent US studies, the estimated incidence was higher. Uramoto and colleagues [12] found a rate of 5.56 cases per 100 000 person-years (95% CI 3.93–7.19) and Naleway and colleagues [15] found a rate of 5.1 cases per 100 000 person-years (95% CI 3.6–6.6). The findings of these later two studies are closer to the incidence rate for Quebec if we used both billing and hospitalization data together (4.4 cases per 100 000 person-years, 95% CI 4.2–4.7). This suggests that one of these single sources of administrative data may not by itself be a very sensitive indicator of SLE incidence.

Regarding prevalence, to date, one attempt has been made to estimate SLE prevalence in Canada, based on medical record review in the province of Manitoba in 1980–96. This work estimated the prevalence of 20.6 cases per 100 000 (95% CI 18.2–23.3); the authors estimate that they may have missed ~15% of SLE cases in the population [16]. The prevalence estimates of this study are lower than that from other recent North American studies (see subsequently), possibly reflecting the fact that Manitoba, compared with other regions, may not have as extensive a population of black/African Americans, a group with particularly high SLE prevalence.

In the most recent US studies of unselected populations, where cases were obtained from clinical records, prevalence estimates range from 78.5 per 100 000 (95% CI 59.0–98.0) [13] to 124 per 100 000 (95% CI 40.0–289.0)[14]. These prevalence estimates are closer to what we generated using more than one source of

administrative data. This again suggests that a single source of administrative data may not by itself be a very sensitive means of identifying SLE cases.

Our methods would have underestimated SLE prevalence in Quebec if, for example, some patients did not receive lupus-related medical care over the 15-yr period over which we identified prevalent cases. Though this may occur rarely, it would likely result in the under-ascertainment of milder forms of SLE. Of course, for some physician visits or hospitalizations, an underlying diagnosis of SLE may not be recorded (if the medical care was primarily for a reason separate from SLE). In their study of Medicare data, Katz and colleagues [18] compared physician claims data for a diagnosis of SLE with medical records, and estimated sensitivity to be 85% (95% CI 73–97%). That study focused on the billing data of rheumatologists only. A very recent study by Nightingale and colleagues [19] discussed the issue of sensitivity for prevalent SLE cases, and pointed out that the likelihood of detecting such cases increases with increasing time of a patient's contribution to an observational database, producing an apparently increasing prevalence over time that is in fact not true.

Any ascertainment method, as well as missing some true cases, will also misclassify as those cases who do not have SLE. One study has previously documented that, in a cohort assembled using hospitalization records, many of the subjects (>50%) did not, upon chart review, fulfil American College of Rheumatology and/or clinical criteria for SLE [17]. We are unable to comment definitively on the number of false positives in our own study, but it is likely to be significant. We hypothesize that, in the case of hospitalization data, the specificity of an SLE diagnosis might be increased if only primary discharge diagnoses were used, but that would of course decrease sensitivity. Using only primary discharge diagnoses, the SLE incidence rate based on MEDECHO hospitalization data alone fell to 1.0 case per 100 000 person-years, and similarly the prevalence for 2003 (based on hospitalization data alone) fell to 13.8 cases per 100 000.

Alternatively, we performed repeat analyses for our billing code data, retaining only SLE cases if there was diagnosis for SLE contributed by a rheumatologist (assuming that this could be considered a 'confirmed' SLE case). Our estimates for SLE incidence and prevalence were relatively similar to the primary analyses using billing data alone (being only slightly lower), suggesting that the primary algorithm for billing code data (two ICD-9 codes at least 8 weeks apart, within 2 yrs) largely does detect 'confirmed' SLE cases. Chart review would be an alternate method of verification, and we are undertaking a new study to address this.

We believe a key point of our work is that, for observational studies, particularly when using administrative databases to identify cohorts of SLE patients, careful consideration of alternate algorithmic definitions of SLE and the data source used are needed. Also needed are better statistical methods that account for imperfect ascertainment. The point estimates for incidence and prevalence can differ considerably depending on which approach for SLE detection is used. Perhaps even more importantly, different methods of cohort assembly using administrative data may create subject pools with varying demographic or clinical features, which may influence the conclusions drawn.

Ultimately, it is not possible to say that one method or approach to case ascertainment is correct or necessarily better than another. However, with appropriate statistical methods, there can be some adjustments made for the presence of error in each data source.

With our Bayesian hierarchical model, we found that the sensitivity of case ascertainment methods that relied on physician billing and/or hospitalization records was independently influenced by both patient and physician characteristics. It is logical that sensitivity estimates might be higher in adult female patients, since that is the age group where physician suspicion of SLE should be highest. The finding that rural residence was associated with lower sensitivity of the case ascertainment methods is interesting, but potentially complex to explain. Though one might suggest that rural residence would be associated with less contact with specialists, and thus in part explain the lower sensitivity, our estimates were adjusted for specialty of physician. It is possible that incomplete adjustment for physician or patient factors, and residual confounding related to these, may be responsible for the lower sensitivity for patients in rural areas.

That our case ascertainment methods show higher sensitivity for specialists who traditionally oversee the care of SLE patients (rheumatologists, immunologists and nephrologists) intuitively makes sense. On the other hand, the lower sensitivity for physicians graduating after 1980 suggests that, when it comes to making an SLE diagnosis, experience plays an important role.

In summary, administrative databases may be useful sources of information regarding observational studies of patients with SLE. However, one must carefully consider both the strategies, data sources and algorithms used as well as variations in sensitivity for these approaches, related to both patient and physician characteristics. In the ongoing work, we are examining these issues as they relate to other rheumatic conditions as well.

Rheumatology key messages

- Administrative databases may be useful in SLE research, but reliance on a single source (e.g. only billing data or only hospitalization data) is not optimally sensitive, and does not account for inherent error.
- Appropriate statistical methods can adjust for the presence of error in each data source.

Acknowledgements

S.B. holds CIHR Junior Investigator and Canadian Arthritis Network Scholar Awards and is supported by the McGill University Department of Medicine and Research Institute and the Fonds de la recherche en santé du Québec (FRSQ). A.E.C. and L.J. each hold an FRSQ National Scholar Award. C.A.P. is supported by the McGill University Health Center Research Institute. D.E.F. is supported by the Arthritis Society. R.T. is a CIHR Scientist.

Funding: This work was funded by operational grant funding from the Canadian Institutes of Health Research [CIHR] and support from the Singer Family Fund for Lupus Research.

Disclosure statement: The authors have declared no conflicts of interest.

References

1 Insinga RP, Glass AG, Rush BB. Diagnoses and outcomes in cervical cancer screening: a population-based study. Am J Obstet Gyne 2004;191:105–13.
2 Luthi JC, Flanders WD, Pitts SR, Burnand B, McClellan WM. Outcomes and the quality of care for patients hospitalized with heart failure. Int J Qual Health Care 2004;16:201–10.
3 Williams LS, Ghose SS, Swindle RW. Depression and other mental health diagnoses increase mortality risk after ischemic stroke. Am J Psychiatry 2004;161:1090–5.
4 Losina E, Barrett J, Baron JA, Katz JN. Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. J Clin Epidemiol 2003;56:515–9.
5 Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004;57:131–41.
6 Statistique Canada. Institut de la statistique du Québec. Population, Québec et Canada. Recensements du Canada et estimations de la population 2004.
7 Tamblyn R, Lavoie G, Petrella L, Monette J. The use of prescription claims databases in pharmacoepidemiological research: the accuracy and comprehensiveness of the prescription claims database in Quebec. J Clin Epidemiol 1995;48:999–1009.
8 Ashby D. Bayesian statistics in medicine: a 25 year review. Stat Med 2006;15:3589–631.
9 Ladouceur M, Rahme E, Pineau C, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. Biometrics 2007;63:272–9.

10 Bernatsky S, Joseph L, Bélisle P *et al.* Bayesian modelling of imperfect ascertainment methods in cancer studies. Stat Med 2005;24:2365–79.
11 Joseph L, Gyorkos T, 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.
12 Uramoto K, Michet CJ, Thumboo J, Sunku J, O'Fallon W, Gabriel S. Trends in the incidence and mortality of systemic lupus erythematosus, 1950-1992. Arthritis Rheum 1999;42:46–50.
13 McCarty DJ, Manzi S, Medsger TA Jr, Ramsey-Goldman R, LaPorte RE, Kwoh CK. Incidence of systemic lupus erythematosus. Race and gender differences. Arthritis Rheum 1995;38:1260–70.
14 Hochberg MC, Perlmutter DL, Medsger TA *et al.* Prevalence of self-reported physician-diagnosed systemic lupus erythematosus in the USA. Lupus 1995;4:454–6.
15 Naleway AL, Davis ME, Greenlee RT, Wilson DA, McCarty DJ. Epidemiology of systemic lupus erythematosus in rural Wisconsin. Lupus 2005;14:862–6.
16 Peschken C, Esdaile J. Systemic lupus erythematosus in North American Indians: a population based study. J Rheumatol 2000;27:1884–91.
17 Lofstrom B, Backlin C, Sundstrom C, Ekbom A, Lundberg I. Lymphoma subtypes and prognosis in patients with systemic lupus erythematosus. Arthritis Rheum 2005;9(Suppl.):S1960.
18 Katz JN, Barrett J, Liang MH *et al.* Sensitivity and positive predictive value of Medicare Part B physician claims for rheumatologic diagnoses and procedures. Arthritis Rheum. 1997;40:1594–600.
19 Nightingale AL, Farmer RDT, de Vries CS. Systemic lupus erythematosus prevalence in the U.K.: methodological issues when using the General Practice Research Database to estimate frequency of chronic relapsing-remitting disease. Pharmacoepidemiol Drug Safety 2007;16:144–51.