

# Estimating the prevalence of polymyositis and dermatomyositis from administrative data: age, sex and regional differences

S Bernatsky,<sup>1,2</sup> L Joseph,<sup>1,3</sup> C A Pineau,<sup>2</sup> P Bélisle,<sup>1</sup> J F Boivin,<sup>3</sup> D Banerjee,<sup>4</sup> A E Clarke<sup>1,4</sup>

<sup>1</sup> Division of Clinical Epidemiology, McGill University Health Centre (MUHC), Montreal, Quebec, Canada; <sup>2</sup> Division of Rheumatology, MUHC, Montreal, Quebec, Canada; <sup>3</sup> Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada; <sup>4</sup> Division of Clinical Immunology/Allergy, MUHC, Montreal, Quebec, Canada

Correspondence to:  
Dr 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;  
[sasha.bernatsky@mail.mcgill.ca](mailto:sasha.bernatsky@mail.mcgill.ca)

Accepted 26 July 2008  
Published Online First  
18 August 2008

## ABSTRACT

**Objective:** To estimate the prevalence of polymyositis and dermatomyositis using population-based administrative data, the sensitivity of case ascertainment approaches and patient demographics and these parameters.

**Methods:** Cases were ascertained from Quebec physician billing and hospitalisation databases (approximately 7.5 million beneficiaries). Three different case definition algorithms were compared, and statistical methods were also used that account for imperfect case ascertainment, to generate estimates of disease prevalence and case ascertainment sensitivity. A hierarchical Bayesian latent class regression model was developed to assess patient characteristics with respect to these parameter estimates.

**Results:** Using methods that account for the imperfect nature of both billing and hospitalisation databases, the 2003 prevalence of polymyositis and dermatomyositis was estimated to be 21.5/100 000 (95% credible interval (CrI) 19.4 to 23.9). Prevalence was higher for women and for older individuals, with a tendency for higher prevalence in urban areas. Prevalence estimates were lowest in young rural men (2.7/100 000, 95% CrI 1.6 to 4.1) and highest in older urban women (70/100 000, 95% CrI 61.3 to 79.3). Sensitivity of case ascertainment tended to be lower for older versus younger individuals, particularly for rheumatology billing data. Billing data appeared more sensitive in ascertaining cases in urban (vs rural) regions, whereas hospitalisation data seemed most useful in rural areas.

**Conclusions:** Marked variations were found in the prevalence of polymyositis and dermatomyositis according to age, sex and region. These methods allow adjustment for the imperfect nature of multiple data sources and estimation of the sensitivity of different case ascertainment approaches.

Autoimmune myopathies are potentially debilitating (or even life-threatening) diseases characterised by muscle inflammation (myositis) with subsequent weakness. Two classic forms are polymyositis and dermatomyositis. Attention has been drawn to the paucity of epidemiological studies related to systemic autoimmune rheumatic disease including autoimmune myopathies.<sup>1</sup> Administrative databases, containing information such as physician billing and hospitalisation data, are potentially an efficient means for disease surveillance, particularly in relatively rare conditions. However, the optimal approach to extracting information is unknown. Rather than naively assuming that billing and hospitalisation data are

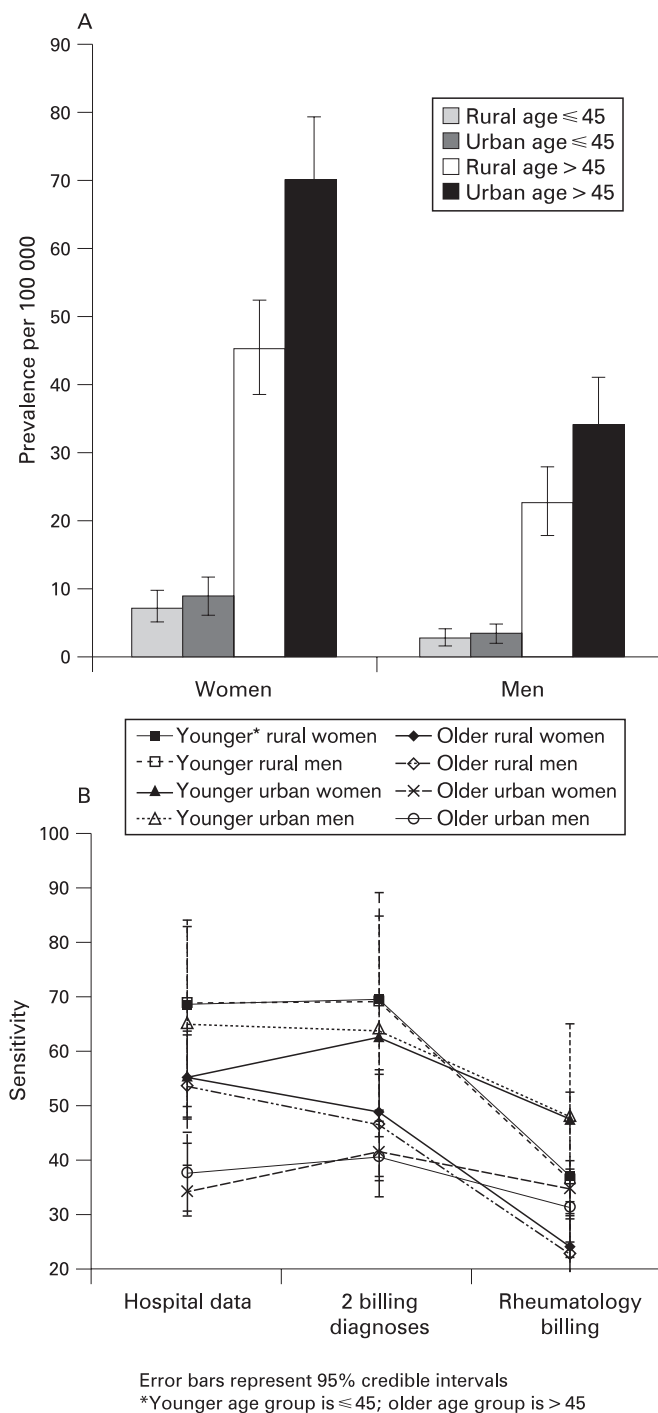
completely accurate, users should consider adjusting for inherent measurement error (and resultant misclassification).

Our objective was to describe the prevalence of polymyositis/dermatomyositis in a large general population, using administrative data. As well as naive estimates (that do not consider errors in the data), we used methods to adjust for possible misclassification within the data sources. In addition, we estimated case ascertainment sensitivity and assessed the effects of patient characteristics on both prevalence and sensitivity. Our research was approved by the McGill University ethics review board.

## METHODS

We ascertained cases of polymyositis and dermatomyositis from the physician billing and hospitalisation databases covering the province of Quebec (approximately 7.5 million individuals) for the period 1989–2003, to determine disease prevalence as of 2003. The billing database documents physician services for all provincial beneficiaries; only one diagnostic code is allowed per visit. The hospital database records hospitalisations, including a primary diagnosis and up to 15 non-primary diagnoses per hospitalisation. All discharge diagnoses are abstracted from the chart by medical records clerks and are not necessarily the same diagnoses as in the billing database (which is based on independent claims information). For both databases, diagnoses are provided as International Classification of Diseases, version 9 (ICD-9) codes.

For our study, we used case definitions based on ICD-9 codes for dermatomyositis (710.3) and polymyositis (710.4). In the billing data, cases were first defined according to an algorithm requiring two or more claims by any physician for 710.3 or 710.4, 2 months or less apart but within a 2-year span. In a second alternative algorithm, we defined cases as those in which there was one or more billing code (710.3–710.4) contributed by a rheumatologist. For the hospitalisation data, we defined a case as any hospitalisation including an ICD-9 code of 710.3–710.4 as a primary or non-primary discharge diagnosis. All Quebec citizens seeking healthcare are captured in billing data and all those with an inpatient stay are captured in the hospitalisation database. However, as our three diagnosis definitions differed significantly from each other, some individuals would be detected by one definition, but not by another. We did not exclude paediatric-onset disease in our estimates.



**Figure 1** (A) Prevalence of autoimmune myopathies (dermatomyositis and polymyositis) in Quebec, 2003. (B) Sensitivity of administrative database diagnoses for autoimmune myopathies (dermatomyositis and polymyositis).

For the “naive” prevalence estimates, generated under the assumption of no error (perfect sensitivity and specificity), we applied the three case definition algorithms (ie, two for billing data and one for hospitalisation data) separately; in each case we divided the number of identified cases by the appropriate population denominator (obtained from Statistics Canada). We then generated a fourth prevalence estimate combining all cases from the first billing code algorithm and the hospitalisation data (our “composite” case definition). The prevalence estimates were calculated for 31 December 2003, based on the number of

cases that had been identified during the study period (1989–2003) who remained alive as of 31 December 2003. No confidence interval is provided for the naive prevalence estimates, because the estimates are based on the entire population of the province of Quebec (not on a sample), assuming no ascertainment error within a given method.

Studies using administrative databases have often implicitly relied on such “naive” algorithms for various disease definitions, without necessarily considering the sensitivity and specificity of the case ascertainment approach. This issue is currently a focus of interest for database researchers,<sup>2,3</sup> who recognise that any method of case ascertainment contains some error. Therefore, in addition to calculating naive estimates (simply using the algorithms above and assuming no error), we generated estimates adjusted for the imperfect sensitivity and specificity of both billing and hospitalisation data. Given that there is, in this context, no perfect case definition, we used a previously developed Bayesian latent class model that does not assume any data source to be a gold standard.<sup>4,5</sup> Bayesian statistical methods are based on the central Bayes’ Theorem.<sup>6</sup> They allow estimation of an unknown parameter (for example, disease prevalence), by combining previous information (eg, existing data or expert opinion on the properties of the case ascertainment definitions) with new data. A “prior distribution” can be informative (ie, have a strong influence on results) or non-informative. A reliable body of existing evidence (eg, past data) allows one to input informative previous information; otherwise, one may use a non-informative prior distribution (for example, one may assume that any value within a given range is equally likely—a “uniform” prior distribution).

Without a perfect gold standard for a method of case ascertainment, the true sensitivity and specificity of a single data source or diagnostic approach cannot be directly estimated; neither can disease prevalence. These parameters are all “latent”; that is, they are not observed directly, but may be derived from existing data. The specific model used depends on whether the sources of information available can be considered as conditionally independent or conditionally dependent. Conditional independence implies that given disease status (positive or negative, but unobserved, thus latent), the result of one test or data source does not correlate with the results of the others. As two of our ascertainment methods are derived from a similar source (billing data), conditional independence was an unreasonable assumption. Therefore, in our model we estimated the dependence between these tests and further adjusted for this parameter.

With only three tests and possible dependence between them, our problem becomes non-identifiable, which means in practice that informative prior distributions need to be elicited over a subset of the parameters in order to estimate all quantities of interest.<sup>7</sup> For our model, given seven unknown parameters (prevalence and the sensitivities and specificities of the three different case ascertainment methods), we needed to define informative priors for two of them. Based on previous work on case ascertainment of systemic autoimmune rheumatic diseases using administrative data,<sup>2,4</sup> we expected the specificities of all methods to be very high. Therefore, for our primary analyses we set informative beta (248.3, 1.65) prior distributions for the specificities of our two billing data case ascertainment approaches. This prior corresponds to specificities of 99% (95% credible interval (CrI) 98 to 100). We set alternative specificity priors corresponding to specificities of 98% (96 to 100) and 94% (88 to 100). As the results using these different sets of previous inputs were similar, we only report results from

## Extended report

the first set. Very diffuse prior distributions were used for all other parameters.

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. Our use of latent class Bayesian regression models in this context has been described previously,<sup>4,5,7</sup> although with different case definitions in other diseases. In brief, the levels of our hierarchical model accounted for: (1) sampling variability in prevalence across the population and accounting for errors in each test, assumed to follow binomial distributions in which the probability of a positive test includes terms for the sensitivity and specificity of each method of ascertainment (thus adjusting for both false positives and false negatives); (2) variation in prevalence according to patient demographics (age, sex and rural vs urban residence), input as a logistic regression model on the binomial probabilities from the first level; (3) variation in the sensitivity of case ascertainment according to patient demographics (age, sex and rural vs urban residence), again input as a logistic regression model, this time on the sensitivities.

For all our Bayesian estimates, we produced 95% CrI; these represent the values between which there is a 95% probability of containing the parameter of interest, given the data at hand and the previous information used.<sup>1</sup> All programming was carried out using WinBUGS.<sup>8</sup>

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

	Adjusted OR (95% CrI)*
Prevalence in 2003	
Female	2.52 (1.08 to 3.99)
Age ≤45 years	8.58 (2.87 to 13.61)
Urban vs rural residence	1.40 (0.80 to 2.01)
Interaction, age and sex	18.16 (6.32 to 30.57)
Effects on sensitivity, hospitalisation data†	
Female	0.91 (0.44 to 1.97)
Age ≤45 years	0.49 (0.25 to 1.23)
Urban vs rural residence	0.58 (0.36 to 1.43)
Interaction, age and sex	0.44 (0.19 to 1.36)
Effects on sensitivity, billing first algorithm‡	
Female	1.03 (0.50 to 2.07)
Age ≤45 years	0.43 (0.19 to 0.96)
Urban vs rural residence	0.78 (0.49 to 1.34)
Interaction, age and sex	0.44 (0.16 to 1.04)
Effects on sensitivity, billing second algorithm‡	
Female	1.02 (0.56 to 1.88)
Age >45 years	0.53 (0.28 to 1.04)
Urban vs rural residence	1.58 (0.98 to 2.31)
Interaction, age and sex	0.58 (0.28 to 1.19)

\*Bayesian credible intervals (CrI) represent the values between which there is a 95% probability of containing the parameter of interest, given the data and previous information input. †Case ascertainment based on one or more hospitalisation discharge diagnostic codes (primary or non-primary). ‡For billing data, the first algorithm was based on two or more diagnostic codes (≥8 weeks apart and within 2 years) contributed by any physician; the second algorithm was based on one or more diagnostic codes contributed by a rheumatologist. OR, odds ratio.

<sup>1</sup>The definition of credible intervals actually corresponds to how many interpret frequentist confidence intervals, but a 95% confidence interval in fact represents the concept that 95% of the confidence intervals generated with a large number of repeated samples would include the true value of the parameter.

## RESULTS

Regarding “naive” estimates, when a case of dermatomyositis/polymyositis was defined based on cases identified from billing data only, the prevalence estimate was 10.2 cases per 100 000 with the requirement of two or more physician codes for dermatomyositis/polymyositis (≤2 months apart but within 2 years) and 8.1 per 100 000 with the algorithm requiring only one or more rheumatologist code for dermatomyositis/polymyositis. The prevalence was 10.5 per 100 000 if the case definition was limited to only cases identified from the hospitalisation data. Approximately half of the cases identified in hospitalisation data were not identified with the first billing data algorithm, and the same proportion of cases identified by that billing algorithm was not identified in the hospitalisation database. For our “composite” case definition based on either two or more physician visits or one or more hospitalisation, the prevalence estimate for 2003 was 15.6 per 100 000.

In the primary analyses using the Bayesian latent class hierarchical model that adjusts for the imperfect nature of the databases, we estimated the prevalence of dermatomyositis/polymyositis in Quebec, for the year 2003, to be 21.5 per 100 000 (95% CrI 19.4 to 23.9). Prevalence was higher for women than for men, and for older individuals (fig 1A, table 1). Prevalence estimates were lowest in young rural men (2.7/100 000, 95% CrI 1.6 to 4.1) and highest in older urban women (70/100 000, 95% CrI 61.3 to 79.3). With multivariate adjustment in our regression model, the effect of sex and age on prevalence remained, and there was also evidence of an interaction between age and sex (table 1).

Sensitivity of case ascertainment tended to be lower in older versus younger individuals (table 1, fig 1B), although wide credible intervals preclude definitive conclusions. In the regression model (table 1), this effect was most clearly seen with the billing data algorithm requiring two or more physician visits for dermatomyositis/polymyositis, with the adjusted odds ratio (0.43, 95% CrI 0.19 to 0.96), suggesting that using this algorithm, dermatomyositis/polymyositis cases in older individuals would be less likely to be detected (compared with younger individuals). As illustrated in fig 1B, hospitalisation data tended to be more sensitive than billing data for case ascertainment among individuals living in rural areas, particularly for older individuals. For older rural residents, the sensitivity estimates of case ascertainment relying on rheumatology billing were particularly low (23–24%), whereas sensitivity estimates for hospitalisation data and physician billing data in rural areas were approximately 69–70% for younger individuals and 47–55% in older individuals. The confidence intervals for estimates, however, were generally wide and overlapping.

The demographics and drug exposure histories of the subject pool appeared to be consistent with what one would expect for individuals with myositis. For example, in patients identified through rheumatology billing data, the average age was 60 years (SD 18, median 61) and the sample was 68.5% female. We did not have drug exposure information on all individuals; however, the province provides drug insurance for all residents aged 65 years of older (as well as a subset of the younger population without private drug insurance). In these individuals we noted that 80% of identified myositis cases were exposed to glucocorticoids or steroid-sparing agents following diagnosis. Additional individuals may have been treated with infusion therapies such as intravenous gamma globulin, methylprednisolone, or cyclophosphamide (information not available from the administrative drug database).



## DISCUSSION

Earlier results<sup>9</sup> of pooled analyses placed the prevalence of autoimmune myopathies at approximately five per 100 000. Those pooled estimates do not reflect the effects of variations among populations in terms of sex, age, time and region. Given that autoimmune myopathies affect demographic subgroups differentially, an overall population prevalence estimate might be somewhat uninformative. Variations in the frequency of autoimmune myopathies by age and sex have been estimated in earlier studies.<sup>10 11</sup> However, much of the existing data was generated decades ago; the ageing of populations in the developed world probably means that current prevalence is higher than previous estimates.<sup>12</sup> One hypothesis is that with increasing awareness of systemic autoimmune disorders, these conditions are diagnosed and treated more promptly; this has not been specifically proved.<sup>13</sup>

Although not definitive, our estimates suggest that different case ascertainment approaches may be more sensitive in certain demographic subgroups. For example, case ascertainment based on hospitalisation data alone would tend to pick up a greater proportion of rural (vs urban) cases. This might be explained by the fact that people in rural areas generally have poorer access to specialists. One implication is that a study relying on a single ascertainment method (eg, recruitment from rheumatology offices) may correspond to a very different sample than another study using a different method (eg, that examined only cases from hospital records). The sample differences might relate not only to demographics, but potentially also to clinical factors (eg, disease severity).

One limitation of our case ascertainment approach was that we relied on the ICD-9 classification system, which has specific codes for polymyositis and dermatomyositis, but does not have a specific code for a related condition, inclusion body myositis (IBM). Although this condition is distinct from the other two, there is an autoimmune component. Traditionally IBM was said to be quite rare, but recent studies now suggest a greater prevalence than has been appreciated, as high as 22/100 000 for males and 10/100 000 for females, with a prevalence of up to 35.3/100 000 for people aged 50 years and older.<sup>14</sup> It is felt that IBM is often misdiagnosed as polymyositis or other conditions and is only suspected retrospectively after a poor response to initial therapy.<sup>15</sup> It is likely therefore that some of the polymyositis cases we identified through our administrative datasets were actually IBM.

In other studies of myositis epidemiology, stricter, clinical criteria for case identification were used. That approach may miss some cases of autoimmune myopathy, as has been noted.<sup>16</sup> On the other hand, physician coding is not always accurate or complete, either as a result of diagnostic uncertainty or inattention to coding details. Also, the existence of only one diagnostic code per visit for Quebec billing data limits the sensitivity of this data source. Patients often have multiple comorbidities, which physicians may use as the billing diagnosis, instead of the autoimmune myopathy itself. Regarding the specificity of the diagnostic codes that we used, "myopathy" is a non-specific term related to any muscle pathology; however, autoimmune myopathies have distinct ICD-9 code labels for dermatomyositis (710.3) and polymyositis (710.4), compared with non-autoimmune myopathies of genetic (eg, muscular dystrophy) or other origin, which also have specific codes (359.4–359.9). Hospitalisation data also contain some error, both in terms of false positives and false negatives, as others have shown.<sup>17</sup> This again underlines the importance of

analytical methodology that accounts in some way for error in any approach.

Sensitivity of case ascertainment using billing data may vary with the training and experience of physicians and their access to diagnostic tests. Our hierarchical model allowed for the variability of case ascertainment across different patient characteristics. This included rural versus urban residence, which is one of the primary determinants of access to specialists and tests. Differences in case ascertainment according to physician characteristics can potentially be studied by adding another level to the hierarchical model. Unfortunately, our current work was not adequately powered to produce precise estimates of physician effects directly. However, our earlier work suggested that the sensitivity of case ascertainment methods for rheumatic diseases using physician billing data may be higher among relevant specialists (eg, rheumatologists) compared with non-specialists. We note that including 710.3–710.4 billing activity by neurologists in our second case definition did not vary the results appreciably.

In many developed countries, health systems are experiencing stress related to ageing populations and the increasing prevalence of chronic diseases. Some, including Canada, have looked to the use of administrative databases for disease surveillance, to aid in resource planning. An example is the Quebec Infostructure de recherche integree en santé (IRIS), whose mandate includes developing surveillance activities for all major chronic disease groups, including arthritis and other rheumatic conditions. In this context, the sensitivity and specificity of case ascertainment is one issue, but the stability of these parameters over time is another.

Agencies considering the use of administrative data for disease surveillance need to consider other issues, such as privacy. Privacy risks to patients can be minimised with the use of anonymised data and optimal security measures. As the potential for societal benefits (eg, improved patient care and public health) is high, some endorse waivers of informed consent for minimal-risk observational studies, on the grounds that requiring consent makes it difficult or impossible to complete a valid study.<sup>18</sup> Another issue is whether physicians should be actively encouraged to ensure that diagnostic information is adequately recorded in order to assist surveillance efforts. There are at present no incentives for Quebec physicians to bill accurately. Research efforts that highlight the potential limitations of billing data are one means of bringing the issue to the forefront. The increasing computerisation of medical records will make extraction of clinical data more feasible (using clinical vocabulary maps, eg, SNOMED CT) to validate billing and hospitalisation diagnoses in the future.

That we were able to establish expected demographic patterns, in terms of age and sex, suggests that case ascertainment of autoimmune myopathies using administrative data has some face validity. However, no method of case ascertainment is completely accurate and there is obviously always error in administrative data, as demonstrated by the widely varying estimates using any one of our sources. Adequate attention should thus be paid to account for imperfect sensitivity and specificity of these data sources. The Bayesian hierarchical latent class regression model we used provides a means for doing this. With such methods, regardless of their limitations, administrative databases may still be useful for surveillance of diseases such as autoimmune myopathies.

**Funding:** This study was funded by the Canadian Institutes of Health Research (CIHR). SB is a Canadian Arthritis Network scholar and is supported by the CIHR, the Fonds de

## Extended report

la Recherche en Santé du Québec (FRSQ) and the McGill University Health Centre (MUHC) Research Institute and Department of Medicine. CAP is supported by the MUHC Research Institute and Department of Medicine. AEC and LJ are FRSQ national scholars.

**Competing interests:** None.

**Ethics approval:** This research was approved by the McGill University ethics review board.

## REFERENCES

- Jacobson DL**, Gange SJ, Rose NR, Graham NM. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. *Clin Immunol Immunopathol* 1997;**84**:223–43.
- 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**:515–19.
- Wilchesky M**, Tamblyn RM, Huang A. ValOne recent identification of diagnostic codes within medical services claims. *J Clin Epidemiol* 2004;**57**:131–41.
- Ladouceur M**, Rahme E, Pineau CA, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics* 2007;**63**:272–9.
- Bernatsky S**, Joseph L, Belisle P, Boivin JF, Rajan R, Moore A, *et al.* Bayesian modelling of imperfect ascertainment methods in cancer studies. *Stat Med* 2005;**24**:2365–79.
- Ashby D**. Bayesian statistics in medicine: a 25 year review. *Stat Med* 2006;**25**:3589–631.
- Dendukuri N**, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001;**57**:158–67.
- Spiegelhalter D TA**. *Best N WinBUGS version 1.2 user manual*. Cambridge, UK: MRC Biostatistics Unit, 1999.
- Cooper GS**, Stroehla BC. The epidemiology of autoimmune diseases. *Autoimmun Rev* 2003;**2**:119–25.
- Lynn SJ**, Sawyers SM, Moller PW, O'Donnell JL, Chapman PT. Adult-onset inflammatory myopathy: North Canterbury experience 1989–2001. *Intern Med J* 2005;**35**:170–3.
- Marie I**, Hatron PY, Levesque H, Hachulla E, Hellot MF, Michon-Pasturel U, *et al.* Influence of age on characteristics of polymyositis and dermatomyositis in adults. *Medicine (Baltimore)* 1999;**78**:139–47.
- Lyon MG**, Bloch DA, Hollak B, Fries JF. Predisposing factors in polymyositis-dermatomyositis: results of a nationwide survey. *J Rheumatol* 1989;**16**:1218–24.
- Airio A**, Kautiainen H, Hakala M. Prognosis and mortality of polymyositis and dermatomyositis patients. *Clin Rheumatol* 2006;**25**:234–9.
- Phillips BA**, Zilko PJ, Mastaglia FL. Prevalence of sporadic inclusion body myositis in Western Australia. *Muscle Nerve* 2000;**23**:970–2.
- Dalakas MC**. Sporadic inclusion body myositis—diagnosis, pathogenesis and therapeutic strategies. *Nat Clin Pract Neurol* 2006;**2**:437–47.
- Troyanov Y**, Targoff IN, Tremblay JL, Goulet JR, Raymond Y, Senecal JL. Novel classification of idiopathic inflammatory myopathies based on overlap syndrome features and autoantibodies: analysis of 100 French Canadian patients. *Medicine (Baltimore)* 2005;**84**:231–49.
- Lofstrom B**, Backlin C, Sundstrom C, Ekbohm A, Lundberg IE. A closer look at non-Hodgkin's lymphoma cases in a national Swedish systemic lupus erythematosus cohort: a nested case-control study. *Ann Rheum Dis* 2007;**66**:1627–32.
- Tu JV**, Willison DJ, Silver FL, Fries JF. Impracticability of informed consent in the Registry of the Canadian Stroke Network. *N Engl J Med* 2004;**350**:1414–21.