BMJ Open

Case definitions in Swedish register data to identify systemic lupus erythematosus

Elizabeth V Arkema,[1] Andreas Jönsen,[2] Lars Rönnblom,[3] Elisabet Svenungsson,[4] Christopher Sjöwall,[5] Julia F Simard[1,6,7]

To cite: Arkema EV, Jönsen A, Rönnblom L, *et al*. Case definitions in Swedish register data to identify systemic lupus erythematosus. *BMJ Open* 2016;**6**:e007769. doi:10.1136/bmjopen-2015-007769

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-007769).

Received 23 January 2015
Accepted 23 November 2015

For numbered affiliations see end of article.

Correspondence to
Dr Elizabeth V Arkema;
Elizabeth.Arkema@ki.se

ABSTRACT

Objective: To develop and investigate the utility of several different case definitions for systemic lupus erythematosus (SLE) using national register data in Sweden.

Methods: The reference standard consisted of clinically confirmed SLE cases pooled from four major clinical centres in Sweden (n=929), and a sample of non-SLE comparators randomly selected from the National Population Register (n=24 267). Demographics, comorbidities, prescriptions and autoimmune disease family history were obtained from multiple registers and linked to the reference standard. We first used previously published SLE definitions to create algorithms for SLE. We also used modern data mining techniques (penalised least absolute shrinkage and selection operator logistic regression, elastic net regression and classification trees) to objectively create data-driven case definitions. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for the case definitions identified.

Results: Defining SLE by using only hospitalisation data resulted in the lowest sensitivity (0.79). When SLE codes from the outpatient register were included, sensitivity and PPV increased (PPV between 0.97 and 0.98, sensitivity between 0.97 and 0.99). Addition of medication information did not greatly improve the algorithm's performance. The application of data mining methods did not yield different case definitions.

Conclusions: The use of SLE International Classification of Diseases (ICD) codes in outpatient clinics increased the accuracy for identifying individuals with SLE using Swedish registry data. This study implies that it is possible to use ICD codes from national registers to create a cohort of individuals with SLE.

Strengths and limitations of this study

- The use of objective data mining techniques, a large sample size and physician-diagnosed systemic lupus erythematosus (SLE) cases as the 'gold standard' were unique strengths of this study.
- This study supports the use of national register data, especially outpatient non-primary specialist care, to identify a cohort of individuals with SLE. This can aid future work in SLE research, minimising misclassification and improving statistical power.
- Although the Swedish healthcare setting and the linkage of several population-based registers are important strengths, it may limit comparability with other data sources and settings.
- Validation efforts, in Sweden and internationally, should confirm the use of the case definitions identified in this study.

INTRODUCTION

Systemic lupus erythematosus (SLE) is a relatively uncommon, complex and heterogeneous disease, making it a challenge to conduct adequately powered studies based on it. Identification of individuals with SLE from large health administrative databases is an important and practical approach to increase a SLE study power. In Sweden, a wealth of register data exists that could potentially be used for this purpose, but the extent of SLE misclassification is unknown.

The case definition used when identifying patients with SLE in administrative or register data for epidemiological studies can greatly affect the number and type of cases included. Not only is SLE clinically complex, but the lack of definitive tests and diagnostic criteria complicate the diagnostic process. Among a small set of individuals identified on the basis of a single inpatient admission, nearly half could not be confirmed by medical record review in two validation studies.[1 2] However, in other studies that required multiple SLE-specific visits, the use of administrative data was shown to be valid when using the American College of Rheumatology (ACR) criteria as the gold standard.[3 4] Previous reports of the accuracy of SLE case definitions have used different

reference standards and did not include a sample from the general population, thus decreasing their utility in different settings.

Our aim was to investigate different case definitions—some generated using data mining techniques and some obtained and/or modified from previous studies—to determine which case definition was the most accurate at identifying prevalent SLE in health register data.

METHODS

Setting and design

Patients with SLE (cases) were identified from four major Swedish clinical cohorts in Linköping, Lund, Stockholm and Uppsala. Their recruitment and data collection have been described elsewhere.[5 6] Each centre contributed data available through 2010: Linköping's clinical lupus register in northeastern Gothia (KLURING) at Linköping University Hospital (N=207), Lund University's Department of Rheumatology (N=330), the Karolinska University Hospital cohorts (Karolinska Lupus Cohort, sometimes referred to as Stockholm, N=443) and the Uppsala Lupus Cohort at Uppsala University Hospital (N=179). These patient cases were clinically confirmed, met at least four of the 1982 ACR criteria,[7] agreed to participate in their cohorts and were 17 years of age or older. We restricted the recruitment to cases alive and living in Sweden with a diagnosis of SLE before 1 January 2010.

Individuals without SLE (non-cases) who were 17 years of age or older and living in Sweden as on 1 January 2010 were identified from the Total Population Register in a separate large matched cohort. The non-case population is derived from a large matched cohort that is the predecessor to the SLINK cohort,[8] and were required to not have had an ICD code for SLE at the time each matched case was diagnosed. Briefly, each register-identified SLE case, including the present study's cases, was matched on birth year, sex and county of residence to five individuals selected from the general population. The entire pool of general population comparators was eligible for inclusion in the present study (n=81 974), but subject to the same exclusions as the cases. To increase power, matching from the original SLINK cohort was not preserved; instead, matching factors were considered in the analyses. Furthermore, for this study, only individuals who utilised the healthcare system (inpatient or outpatient care for any reason) in one of the four clinical cohort counties were included (see figure 1).

The final study population consisted of 929 clinically confirmed SLE cases and 24 267 non-cases (a cohort

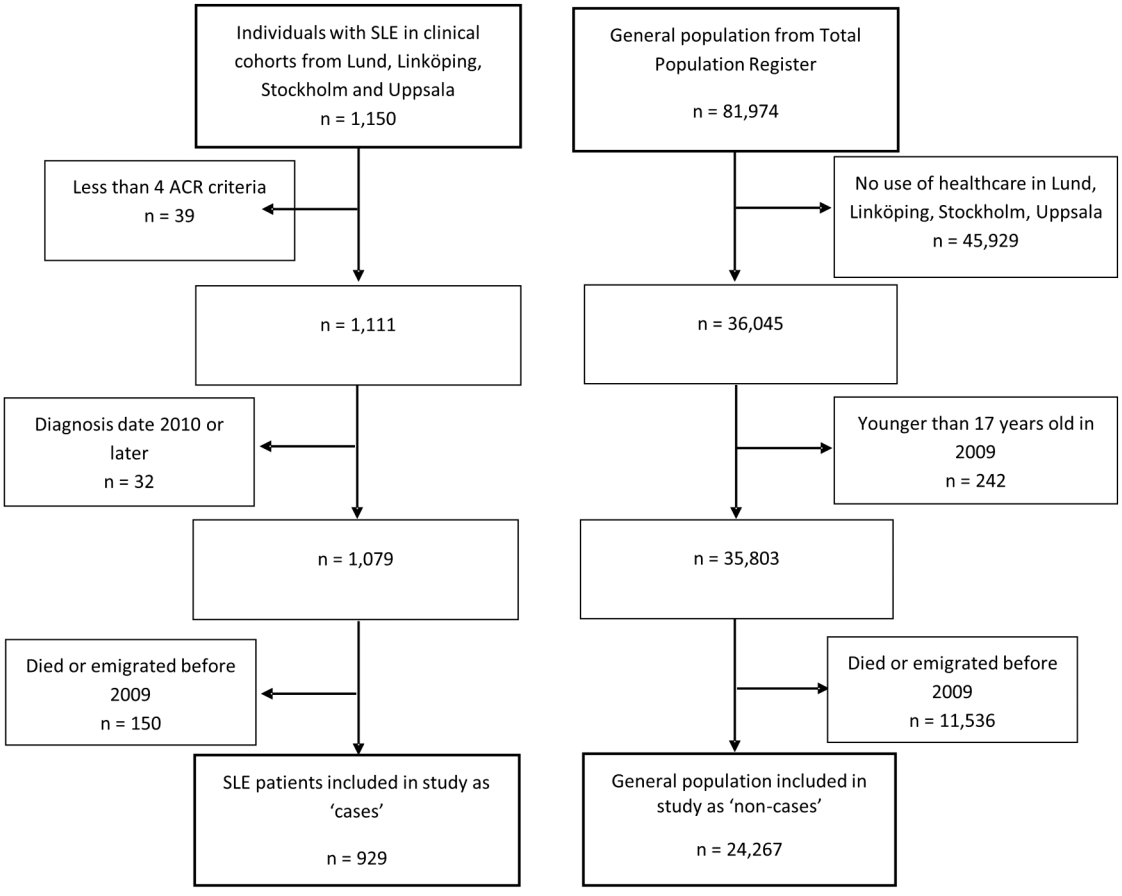

Figure 1 Flow chart showing how the final group of SLE cases and general population non-cases were identified through exclusion criteria. ACR, American College of Rheumatology; SLE, systemic lupus erythematosus.

prevalence of 3.8%, with a ratio of cases to non-cases of 26:1). These cases and non-cases served as our reference standard.

Register-derived data

The cases and non-cases were linked to multiple registers using a unique personal identification number. The registers include comorbidity and demographic data from which candidate variables were derived to create the algorithms used in this study (for a complete list of variables considered, see online supplementary appendix). We identified discharge diagnoses for SLE and several comorbidities via ICD codes listed in the inpatient (1964–2009) or outpatient (2001–2009) registers, and determined whether each individual had a history of the disease. We used the Prescription Drug Register (2005–2009) to identify any medication dispensing for several conditions. We defined diseases or conditions by presence of either a discharge diagnosis or documented medication dispensing data. The data on biopsies (lip, renal or salivary gland biopsy) were collected from the inpatient and the day surgery registers (1997–2009). Comorbidities in relatives (any ICD code for an autoimmune disease in a child, parent or sibling) were obtained through linkage of the Multi-Generation Register (available information on relatives of individuals registered in Sweden since 1961 and born 1932, through 2009). From the Medical Birth Register, which contains information on all births in Sweden 1973–2009, we obtained an indicator of maternal SLE diagnosis as well as any ICD code for SLE listed at the time of the birth of a child or during prenatal care (women only).

Number of hospitalisations and outpatient visits listing an SLE ICD code were considered. We also included information from the patient register on where the diagnosis was made. An indicator for a 'specialised clinic' was defined as an SLE discharge diagnosis in rheumatology, internal medicine, paediatrics, dermatology or nephrology. Lastly, several demographic variables (age, country of birth, education level) were obtained from the Total Population Register.

Case definition identification

We identified several definitions using different approaches. Our first approach used traditional methods that involved creating case definitions based on previously published SLE definitions and clinical experience. The second approach used modern data mining techniques to objectively create case definitions that are data-driven. In both approaches, we examined SLE case definitions in men and women separately.

Traditional approach

We used previously published SLE case definitions and calculated measures of accuracy. The following case definitions were examined: any hospitalisation listing SLE (inpatient register), any outpatient visit listing SLE, any visit to either outpatient or inpatient, or any visit to a specialised clinic. We also expanded these definitions by including an element of time (two or more visits within 2 years) to examine the effect of the timing of healthcare visits. Furthermore, we examined variations of the above case definitions.

Data mining approach

We used classification trees, penalised least absolute shrinkage and selection operator (LASSO) regression, and elastic net regression. Classification trees classify individuals based on binary variables that separate cases from non-cases through a series of repeated stratifications. Each branch of a classification tree represents a stratification based on the value of the variable selected using a splitting criterion. The splitting criterion identifies which variable and what cut point splits the data in the best possible way (making each resulting node of the tree more homogeneous with respect to cases and non-cases). For continuous variables, the criterion considers each possible value as a cut point and selects the one that best distinguishes cases from non-cases, thereby, creating a binary variable. The R package RPART[9] was used to construct a large classification tree including all of the available variables (see online supplementary appendix). We built the tree using the Gini index to split nodes, with a minimum of two in any terminal node.[10] We set the complexity parameter to zero so that there was no limit on amount of improvement of splits (no prepruning). From the resulting tree, we then pruned down the 'branches' to where the cross-validated misclassification rate was the smallest. The optimal tree was chosen using 10-fold cross-validation by identifying the subtree that minimised the cross-validated misclassification rate.

The LASSO selects variables by shrinking the coefficients of unimportant predictors to zero.[11] This method is useful when there are a large number of predictors to determine which variables have the strongest effects. We applied LASSO to a logistic regression model including the variables in the online supplementary appendix (R GLMNET[12]). We obtained β coefficients with the shrinkage parameter value that minimised the 10-fold cross-validated misclassification error using α set to 1.

Lastly, we employed elastic net regression which uses a mix between LASSO and ridge regression penalties and allows correlated variables to remain in the model.[13] We conducted a grid search over a range of α (between 0 and 1, by intervals of 0.1), optimised λ through cross validation for each α, and selected the α with the λ with lowest mean-squared error.

To minimise overfitting, each data mining derived algorithm was first developed from randomly selected two-thirds of the data (the training sample) and then applied to the remaining 1/3 of data (the test sample).

Statistical analysis

For each SLE case definition considered, we calculated the sensitivity, specificity, positive predictive value (PPV)

and negative predictive value (NPV). Our a priori objective was to determine which algorithm would provide the highest PPV, highest sensitivity, and highest specificity so that we could minimise false positives, exclude true negatives, and identify as many cases as possible.

RESULTS

Description of reference standard

SLE cases (n=929) were on average 50.8 years old with a mean age at diagnosis of 34.3 years and 88% were female (table 1). The non-cases from the general population (N=24 267) were 57.9 years old on average and 86% were female.

Results of using traditional/previously identified SLE case definitions

Table 2 lists the sensitivity and predictive values for various case definitions of SLE overall, and for men and women separately. Defining SLE using only hospitalisation data resulted in the lowest sensitivity (0.792). Using SLE-coded outpatient visit information performed better (1 or more outpatient visit sensitivity 0.981, PPV 0.972). Overall, the case definitions that used SLE codes derived from the outpatient register resulted in very high sensitivity (0.91–0.99) and PPV (0.97–0.98). Requiring visits to occur within 1 or 2 years and/or the addition of a medication (DMARD, NSAID or glucocorticoid) decreased performance in all three measures.

Registry data algorithms identified by data mining methods

The data mining models identified the following case definitions in the sex-stratified training sets. For males, the classification tree method identified any SLE-coded outpatient visit as the only predictor. The LASSO model also identified any SLE outpatient visit, and the elastic net regression resulted in any outpatient and any inpatient visit with SLE discharge diagnoses. For the latter approach, the selected αs were 0.6 for the model for women and 0.5 for the model for men.

For females, the classification tree method identified one or more outpatient and at least one inpatient visit with SLE as the discharge diagnosis with the best combination of predictors. The LASSO model identified any SLE-coded outpatient visit as being the best predictor for SLE in this population. Using elastic net regression, the number of SLE outpatient visits, any SLE inpatient visit, and any DMARD dispensing were identified as the best set of predictors. This set of predictors was the only one that was not originally included in the set of traditionally-identified case definitions. The sensitivity and specificity of this algorithm in the test set was 0.602 and 1.00, respectively. The positive predictive value was 0.990, and the negative predictive value was 0.985.

Table 1 Clinical characteristics for SLE cases

Characteristics	SLE cases (N=929)
Age in 2009, mean (SD) (years)	50.8 (16.0)
Age at SLE diagnosis, mean (SD) (years)	34.3 (15.0)
Female, N (%)	819 (88.2)
ACR criteria, N (%)	
Malar rash	519 (55.9)
Discoid rash	217 (23.4)
Photosensitivity	610 (65.7)
Oral ulcer	249 (26.8)
Arthritis	734 (79.0)
Serositis	386 (41.6)
Renal disorder	312 (33.6)
Neurological disorder	83 (8.9)
Haematological disorder	588 (63.3)
Immunological disorder	628 (67.6)
ANA positive, N (%)	912 (98.2)
Total ACR Criteria, mean (SD)	5.6 (1.4)

ACR, American College of Rheumatology; ANA, antinuclear antibodies; SLE, systemic lupus erythematosus.

Sensitivity analysis

The denominators in our PPV calculations represent only those who seek care in inpatient or outpatient clinics from the general population. Therefore, the true prevalence is actually lower than the prevalence in the study population. To illustrate the impact of disease prevalence on the observed predictive values,[14] we recalculated PPVs under different prevalence scenarios. The specificity was held constant and the number of individuals in the non-SLE comparator was artificially increased (thereby decreasing SLE prevalence in the sample). By artificially decreasing the prevalence of SLE from 3.8% in the sample to 0.38% of the sample, the PPV dropped from 97.6% to 80.1% in the female sample when using the definition of at least two SLE visits and at least one visit coded for SLE in a specialist clinic. We also standardised using Heston's approach, which artificially sets the prevalence in the sample to 50%, and calculated PPV >99%.[15]

DISCUSSION

The Swedish healthcare registers have immense resources that are used to study several rheumatic diseases such as rheumatoid arthritis (RA) and ankylosing spondylitis.[16 17] We demonstrate that data from the National Patient Register, specifically the outpatient register, can be used to identify a cohort of individuals with SLE. Using a set of algorithms designed through clinical experience and also more objective data mining techniques, we sought to find the case definition that best discriminated between cases and non-cases. Our findings show that both approaches worked well in determining SLE case definitions that performed with high accuracy by keeping the SLE definition simple enough to be useful for other data sources.

Table 2 Accuracy measurements of multiple algorithms to define prevalent SLE in four counties in Sweden 2010

	Overall			Women			Men		
Algorithms	Sensitivity	PPV	NPV	Sensitivity	PPV	NPV	Sensitivity	PPV	NPV
≥1 hospitalisation	0.792	0.979	0.992	0.789	0.976	0.992	0.818	1.000	0.994
≥1 hospitalisation OR outpatient visit	0.991	0.970	1.000	0.994	0.967	1.000	0.973	1.000	0.999
≥1 outpatient visit	0.981	0.972	0.999	0.984	0.969	0.999	0.955	1.000	0.998
≥1 visit in a specialist clinic, hospitalisation OR outpatient visit	0.988	0.975	1.000	0.990	0.971	1.000	0.973	1.000	0.999
≥2 hospitalisations	0.646	0.982	0.987	0.648	0.980	0.986	0.627	1.000	0.988
≥2 hospitalisations OR outpatient visits	0.981	0.975	0.999	0.983	0.972	0.999	0.964	1.000	0.999
≥2 outpatient visits	0.970	0.977	0.999	0.974	0.974	0.999	0.936	1.000	0.998
≥2 hospitalisations OR outpatient visits, ≥1 in a specialist clinic	0.981	0.979	0.999	0.983	0.976	0.999	0.964	1.000	0.999
≥2 hospitalisations within 1 year	0.531	0.980	0.982	0.525	0.977	0.982	0.573	1.000	0.986
≥2 hospitalisations within 2 years	0.566	0.980	0.984	0.565	0.977	0.983	0.573	1.000	0.986
≥2 outpatient visits within 1 year	0.941	0.978	0.998	0.947	0.975	0.998	0.891	1.000	0.996
≥2 outpatient visits within 2 years	0.963	0.977	0.999	0.968	0.974	0.999	0.927	1.000	0.998
≥2 hospitalisations OR outpatient visits within 1 year	0.961	0.975	0.999	0.962	0.972	0.999	0.955	1.000	0.998
≥2 hospitalisations OR outpatient visits within 2 years	0.976	0.975	0.999	0.979	0.972	0.999	0.955	1.000	0.998
≥2 hospitalisations OR outpatient visits, ≥1 in a specialist clinic within 1 year	0.961	0.978	0.999	0.962	0.975	0.999	0.955	1.000	0.998
≥2 hospitalisations OR outpatient visits, ≥1 in a specialist clinic within 2 years	0.976	0.978	0.999	0.979	0.976	0.999	0.955	1.000	0.998
≥2 hospitalisations within 1 year and medication*	0.508	0.981	0.982	0.502	0.979	0.981	0.545	1.000	0.985
≥2 hospitalisations within 2 years and medication*	0.538	0.980	0.983	0.537	0.978	0.982	0.545	1.000	0.985
≥2 outpatient visits within 1 year and medication*	0.889	0.978	0.996	0.893	0.975	0.996	0.864	1.000	0.995
≥2 outpatient visits within 2 years and medication*	0.906	0.977	0.996	0.910	0.974	0.996	0.882	1.000	0.996
≥2 hospitalisations OR outpatient visits within 1 year and medication*	0.907	0.976	0.996	0.907	0.973	0.996	0.909	1.000	0.997
≥2 hospitalisations OR outpatient visits within 2 years and medication*	0.918	0.976	0.997	0.919	0.973	0.997	0.909	1.000	0.997
≥2 hospitalisations OR outpatient visits, ≥1 in a specialist clinic within 1 year and medication*	0.907	0.978	0.996	0.907	0.975	0.996	0.909	1.000	0.997
≥2 hospitalisations OR outpatient visits, ≥1 in a specialist clinic within 2 years and medication*	0.918	0.978	0.997	0.919	0.975	0.997	0.909	1.000	0.997

Specificity was 0.999 or 1.000 for all algorithms.
*Medication includes any dispensing for a disease-modifying antirheumatic drug, glucocorticoid, or non-steroidal anti-inflammatory drug listed in the prescription drug register (see online supplementary appendix for detailed list).
NPV, negative predictive value; PPV, positive predictive value; SLE, systemic lupus erythematosus.

Studies attempting to identify SLE cases from Swedish registry data have resulted in varying degrees of classification accuracy, but these have not included information from the outpatient register, which is where many SLE cases are diagnosed and managed.[1 2] We found that reliance on only the inpatient register resulted in the lowest accuracy. Furthermore, previous studies in Sweden included SLE cases diagnosed over 20 years ago. In contrast, we provide more up-to-date estimates of the performance of multiple case definitions in a more current time period, perhaps reflecting differences in how physicians diagnose and manage SLE today. We also found that the case definition for males was simpler than for females, which may not be entirely unexpected. Given that SLE is more often considered a disease of women in their childbearing years, it may be that males will have more obvious and severe manifestations before the diagnosis of SLE is considered.

Previous studies from the US and Canada[3 4 18 19] have shown that it is possible to accurately identify SLE using administrative data, but due to different sampling techniques and reference standards, comparing their measures of accuracy to other data sources and settings is difficult. This applies also to our findings, as the Swedish setting and the linkage of several population-based registers is a unique and important strength, but may not be comparable to other data sources.

Our methods of examining the accuracy of administrative data case definitions are similar to other studies in RA[20] with the addition of a data mining approach[21] to objectively identify case definitions. Traditional approaches and case definitions based on clinical experience proved to be as useful as determining an algorithm using data mining techniques. Unlike Liao *et al*,[21] we did not have access to non-codified Electronic Medical Record data for the present study. Future work to determine the most accurate case definitions for disease in administrative data may benefit from using objective data mining techniques to complement their findings; however, these should not completely replace traditional and common sense methods. In this study, we penalised the data mining methods to avoid overfitting (through use of test and training data sets, and cross-validation). This may have resulted in the statistical algorithms not performing the same as the non-statistical case definition methods, which we did not subject to the same scrutiny. Furthermore, although data mining techniques are often considered objective, some subjectivity is introduced by the identification and selection of patient groups, and the variables available for the classification.

The availability of well-defined clinical patients diagnosed and confirmed by physicians using ACR criteria was a unique strength of this study. However, we cannot exclude the possibility that the high PPVs might reflect the coverage of the clinical cohorts. The cases included were diagnosed in specialist clinics, so that a clinic-based diagnosis alone is a strong predictor for SLE as part of the design. When one or more visits to a specialist was used as the algorithm definition, the sensitivity was 0.99 and PPV was 0.98.

Another unique feature of this study was the inclusion of a sample of individuals from the general population as the reference standard for non-cases that was part of a larger register-based control group. These individuals were required to have no history of an SLE diagnosis at the time their matched SLE cases had been diagnosed (index date), but could receive an SLE diagnosis after the index date. The non-SLE reference sample may have included some people misclassified as non-SLE, but due to the relative rarity of SLE this is unlikely to have affected the numbers of false negatives greatly. An error analysis of the 23 misclassified non-SLE revealed that nearly all of these cases appeared to be treated for extended periods of time in rheumatology clinics for SLE with comorbidities and complications characteristic of SLE: hypertension, sicca syndrome, nephritis and end-stage renal disease. A strength of using this convenience sample of non-SLE individuals was that it was possible to estimate PPV, NPV and specificity. These estimates, though, may only be generalisable to data sets sampled in the same way.[22] Our results should be interpreted with this in mind and to show how, in addition to an ICD code for SLE, adding other variables changes the accuracy of the algorithm. As many register-based cohorts are sampled in this way,[16 17] these estimates are especially useful for future use in these studies. Lastly, we were limited to identifying case definitions to the four counties where clinical cases were available. If the accuracy of the case definition varies in different parts of the country and the clinical care is different, we could not detect these differences here.

In this study, we strategically included more cases, which falsely elevates the disease prevalence in the study population. The PPV decreases with a decreasing prevalence of disease in the population, as was shown in our sensitivity analyses. Therefore, our estimated PPV in a cohort with SLE prevalence of 3.8% describes the algorithm's accuracy in these data, but may not be as high in other data sets. Prevalence can also impact the data-driven methods used in our study, resulting in class imbalance which may have led to the identification of fewer variables.

We report here several different algorithms, which will prove to be useful depending on what data are available and for what purpose these will be used. For our goal of creating a cohort of patients with SLE to conduct epidemiological investigations of the causes and consequences of this relatively rare disease, the SLE definitions including outpatient visits were the best. This form of case identification does not aim to identify all SLE in the country, but rather captures patients with prevalent SLE seeking care via non-primary care. Thus our study represents the type of SLE cases that are usually studied in most clinical scientific studies.

The next step is to conduct a validation of cases, including those identified in other counties to see how well these algorithms perform. In the future, pooling SLE cases which are identified with similar algorithms from other countries with similar population-based registers will increase the power of epidemiological analyses in SLE research.

Author affiliations
[1]Clinical Epidemiology Unit, Department of Medicine, Solna Karolinska Institute, Stockholm, Sweden
[2]Department of Clinical Sciences, Lund University, Lund, Sweden
[3]Department of Medical Sciences, Science for Life Laboratories, Uppsala University, Uppsala, Sweden
[4]Rheumatology Unit, Department of Medicine, Karolinska University Hospital, Karolinska Institute, Stockholm, Sweden
[5]Rheumatology/AIR, Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
[6]Division of Epidemiology, Department of Health Research and Policy, Stanford School of Medicine, Stanford, California, USA
[7]Division of Immunology and Rheumatology, Department of Medicine, Stanford School of Medicine, Stanford, California, USA

Acknowledgements The authors would like to thank Lorene Nelson for her helpful feedback.

Contributors EVA and JFS designed this study and performed the analysis. EVA, AJ, LR, ES, CS and JFS interpreted the results, drafted the paper and approved the final version.

Funding EVA is supported by the Strategic Research Programme in Epidemiology at Karolinska Institute. LR acknowledges the Swedish Research Council, the Swedish Rheumatism Association, the King Gustaf V 80-year Foundation and Combine. ES acknowledges ALF funding from Stockholm County Council, the Swedish Research Council, the Swedish Heart-Lung Foundation, the Swedish Rheumatism Foundation, and the King Gustaf V 80-year Foundation. CS acknowledges the County Council of Östergötland, the Swedish Society for Medical Research, the Swedish Rheumatism Association, the Swedish Society of Medicine and the King Gustaf V 80-year Foundation. JF Simard is partly supported by the Strategic Research Programme in Epidemiology at Karolinska Institute.

Competing interests None declared.

Ethics approval Karolinska Institute.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Lofstrom B, Backlin C, Sundstrom C, *et al*. 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.
2. Lofstrom B, Backlin C, Sundstrom C, *et al*. Myeloid leukaemia in systemic lupus erythematosus—a nested case-control study based on Swedish registers. *Rheumatology (Oxford)* 2009;48:1222–6.
3. Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. *Lupus* 2010;19:741–3.
4. Bernatsky S, Joseph L, Pineau CA, *et al*. A population-based assessment of systemic lupus erythematosus incidence and prevalence—results and implications of using administrative data for epidemiological studies. *Rheumatology (Oxford)* 2007;46:1814–18.
5. Leonard D, Svenungsson E, Sandling JK, *et al*. Coronary heart disease in systemic lupus erythematosus is associated with interferon regulatory factor-8 gene variants. *Circ Cardiovasc Genet* 2013;6:255–63.
6. Frodlund M, Dahlstrom O, Kastbom A, *et al*. Associations between antinuclear antibody staining patterns and clinical features of systemic lupus erythematosus: analysis of a regional Swedish register. *BMJ Open* 2013;3:e003608.
7. Tan EM, Cohen AS, Fries JF, *et al*. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1982;25:1271–7.
8. Arkema EV, Simard JF. Cohort profile: systemic lupus erythematosus in Sweden: the Swedish Lupus Linkage (SLINK) cohort. *BMJ Open* 2015;5:e008259.
9. Therneau T, Atikinson B, Ripley B. 2013. rpart: Recursive partitioning. R package version 4.1-1. http://CRAN.R-project.org/package=rpart.
10. Breiman L, Friedman J, Stone CJ, *et al*. *Classification and regression trees*. CRC Press, 1984.
11. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B Methodol* 1996;58:267–88.
12. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
13. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B Stat Methodol* 2005;67:301–20.
14. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ* 1994;309:102.
15. Heston TF. Standardizing predictive values in diagnostic imaging research. *J Magn Reson Imaging* 2011;33:505; author reply 506–7.
16. Neovius M, Simard JF, Askling J, *et al*. Nationwide prevalence of rheumatoid arthritis and penetration of disease-modifying drugs in Sweden. *Ann Rheum Dis* 2011;70:624–9.
17. Askling J, Klareskog L, Blomqvist P, *et al*. Risk for malignant lymphoma in ankylosing spondylitis: a nationwide Swedish case-control study. *Ann Rheum Dis* 2006;65:1184–7.
18. Bernatsky S, Linehan T, Hanly JG. The accuracy of administrative data diagnoses of systemic autoimmune rheumatic diseases. *J Rheumatol* 2011;38:1612–16.
19. 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.
20. Widdifield J, Labrecque J, Lix L, *et al*. Systematic review and critical appraisal of validation studies to identify rheumatic diseases in health administrative databases. *Arthritis Care Res (Hoboken)* 2013;65:1490–503.
21. Liao KP, Cai T, Gainer V, *et al*. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010;62:1120–7.
22. Widdifield J, Bombardier C, Bernatsky S, *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.