PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort Profile: Filling the Gaps in SARDs Research - Collection and Linkage of Administrative Health Data and Self-Reported Survey Data for a General Population-Based Cohort of Individuals with and without Diagnoses of Systemic Autoimmune Rheumatic Disease (SARDs) from British Columbia, Canada
AUTHORS	McCormick, Natalie; Reimer, Kathryn; Famouri, Ali; Marra, Carlo; Aviña-Zubieta, J. Antonio

VERSION 1 - REVIEW

REVIEWER	Elizabeth Arkema Karolinska Institutet Sweden
REVIEW RETURNED	19-Sep-2016

GENERAL COMMENTS	This manuscript describes the creation of a cohort of individuals with and without systemic autoimmune diseases using both administrative health data and questionnaire data to study the economic burden of SARDs. This cohort-in-progress has the potential to have the best of both worlds – the strengths of using population-based data as well as unique and more detailed information from self-reported data. The manuscript is well written, the study is well designed and will provide useful and interesting information. I have a few questions and comments: 1. In the strengths and limitations, 1st bullet point, I am unsure that this is the first population-based cohort of SARDs. One might argue that cohorts derived from Swedish register data (e.g. papers by Hemminki and Sundquist) are also population-based cohorts of these diseases, although their register-based cohort is more inclusive of other diseases and does not incorporate survey data. Re-wording this bullet point so it is more similar to what is stated in the 1st sentence under “further details” would improve this point. 2. In a supplement or in the paper on page 8, could you provide the ICD codes used to identify SARDs? 3. The current data are cross-sectional, but with the addition of the second survey, the data will be prospective. It may not be very informative of change over time with only two time points. Are more than two surveys planned? Is the purpose of these data to look at changes over time or is it to describe the current (cross-sectional) economic burden of SARDs? 4. The patients included in the survey are a mix of incident and prevalent cases. This could be a limitation to your study as the economic burden due to SARDs may change over time since diagnosis. Is information on how much time has passed since diagnosis available in these data? 5. Are SARD-diagnosed individuals more likely to participate in the survey? In the flow chart, there were 135 consenting participants,

	but it does not say what percentage were SARDs. 6. Will medication data be collected, including non-prescription medications? 7. Are questions included in the survey related to disease severity and phenotypes? These may be important for future stratified analyses.

REVIEWER	Dr Mark D Atkinson Swansea University, Swansea, United Kingdom
REVIEW RETURNED	20-Feb-2017

GENERAL COMMENTS	This is a very interesting paper and it is always good to see what comparison there can be between data sources. Although currently there is no linked data, I look forward to future results. I am, however, a little unclear as to what opportunity there will be to compare diagnoses as declared by the participants and shown by the administrative data (both for SARDs and comorbidities). Would it be possible therefore to have more detail (on page 8, paragraph 1) about the content of administrative data files. For example do they have diagnoses from primary care and outpatient clinics as well as from hospital discharge records? Also, do the administrative databases have any information relating to smoking/alcohol, eg Nicotine Replacement Therapy prescriptions, referrals to smoking cessation or alcohol clinics etc? You say on page 12 that each of the two surveys comprises 6 sections. So which is the survey outlined on pp 13 and 14? is this the first survey or are both summarised here? Maybe the 1st survey is section 1 since the summary in Table 1 suggests that only section 1 is being summarised. Please could you clarify this.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Elizabeth Arkema
Institution and Country: Karolinska Institutet, Sweden Competing Interests: None declared

This manuscript describes the creation of a cohort of individuals with and without systemic autoimmune diseases using both administrative health data and questionnaire data to study the economic burden of SARDs. This cohort-in-progress has the potential to have the best of both worlds – the strengths of using population-based data as well as unique and more detailed information from self-reported data. The manuscript is well written, the study is well designed and will provide useful and interesting information. I have a few questions and comments:

1. In the strengths and limitations, 1st bullet point, I am unsure that this is the first population-based cohort of SARDs. One might argue that cohorts derived from Swedish register data (e.g. papers by Hemminki and Sundquist) are also population-based cohorts of these diseases, although their register-based cohort is more inclusive of other diseases and does not incorporate survey data. Re-wording this bullet point so it is more similar to what is stated in the 1st sentence under “further details” would improve this point.

Thank you for pointing this out; we agree it would be more appropriate to emphasise our study's place in the Canadian context. As per the Reviewer's suggestion, we have re-worded the first bullet point of the Strengths and Limitations section.

2. In a supplement or in the paper on page 8, could you provide the ICD codes used to identify SARDs?

As per the Reviewer's suggestion, we now provide the ICD codes in a supplementary file.

3. The current data are cross-sectional, but with the addition of the second survey, the data will be prospective. It may not be very informative of change over time with only two time points. Are more than two surveys planned? Is the purpose of these data to look at changes over time or is it to describe the current (cross-sectional) economic burden of SARDs?

The initial purpose of these data are to describe the current economic burden of SARDs cross-sectionally. The follow-up survey will allow us to assess how different sociodemographic and health behaviour variables, and generic measures of health status, may be associated with changes in employment status and costs over time. At this point we have only sought ethical approval to administer two surveys, though we realise that two time points may not be informative of change over time. To address this, we plan to invite those members of our cohort who have consented to future contact to participate in future surveys.

4. The patients included in the survey are a mix of incident and prevalent cases. This could be a limitation to your study as the economic burden due to SARDs may change over time since diagnosis. Is information on how much time has passed since diagnosis available in these data?

We agree that the varying disease durations of our cohort members will be a limitation when analysing the economic burden, and now acknowledge this in the manuscript (page 21, paragraph 1, final sentence).

In the surveys we are collecting data on month and year of diagnosis, and month and year when symptoms first started. This information will allow us to determine disease duration, and potentially stratify patients by period of time since diagnosis. This is now mentioned on page 13 (first bullet point) when describing the components of the survey.

5. Are SARD-diagnosed individuals more likely to participate in the survey? In the flow chart, there were 135 consenting participants, but it does not say what percentage were SARDs.

At the beginning, we were concerned that SARD-diagnosed individuals would be more likely to participate, as they would have more incentive to do so than those without a SARD diagnosis. However, to our surprise, half of the 127 respondents (64/127=50.4%) did not report any SARD diagnosis. This is now indicated in the flow chart. We had originally included this information in the manuscript itself (second sentence of page 17), though not the flow chart, and we thank the Reviewer for this suggestion, as the reader may otherwise have missed this important, and unexpected, finding.

6. Will medication data be collected, including non-prescription medications?

In Section 1 we ask about the use (ever, and in the past six moths) of immunosuppressive medications and biologics. We also ask about the amount of money spent over the past six months on prescription and non-prescription medications, and vitamins and supplements. To minimise the

length of the survey, and corresponding burden on participants, we are not collecting additional medication data in these first two surveys, but may do so in the future.

7. Are questions included in the survey related to disease severity and phenotypes? These may be important for future stratified analyses.

The data on use of immunosuppressive and biologic medications (ever and recent) will be used as a measure of disease severity. We also ask participants who report a systemic sclerosis diagnosis to indicate whether they have the limited or diffuse form of this disease. Apart from that, we are not collecting data on disease severity in these first two surveys, mainly due to the complexity of including disease-specific instruments in a survey completed by people who may have any (or none) of ten different SARD diagnoses. Still, we agree these are important data to collect, especially in longitudinal analyses, and may collect it in future surveys.

Reviewer: 2
Reviewer Name: Dr Mark D Atkinson
Institution and Country: Swansea University, Swansea, United Kingdom Competing Interests: None declared

This is a very interesting paper and it is always good to see what comparison there can be between data sources. Although currently there is no linked data, I look forward to future results. I am, however, a little unclear as to what opportunity there will be to compare diagnoses as declared by the participants and shown by the administrative data (both for SARDs and comorbidities).

Thank you very much for your comments, and for your inquiry about comparing diagnoses. When the survey responses are linked with the administrative data, we will be able to compare respondents' self-reported diagnoses (both for SARDs and comorbidities) with the inpatient or outpatient diagnoses recorded in the administrative data (up to 25 per inpatient admission, and up to five per outpatient encounter) at any time during the follow-up period. This period will span from January 1, 1990 (or date of registration in the provincial medical insurance plan, if later than January 1990) to the earliest of death, de-registration from the provincial medical plan, or December 31, 2013.

Would it be possible therefore to have more detail (on page 8, paragraph 1) about the content of administrative data files. For example do they have diagnoses from primary care and outpatient clinics as well as from hospital discharge records?

The administrative data files contain diagnoses from all provincially-funded outpatient encounters including visits to primary care and specialist physicians, and outpatient interventions and investigations such as laboratory tests. Up to five diagnoses are recorded for each outpatient encounter, and up to 25 diagnoses for each inpatient admission. We have now added this information to paragraph 1 of page 8.

Also, do the administrative databases have any information relating to smoking/alcohol, eg Nicotine Replacement Therapy prescriptions, referrals to smoking cessation or alcohol clinics etc?

The information contained in Canadian administrative databases on smoking, alcohol, and other health behaviours is very limited. While there are ICD codes for obesity and alcohol abuse, these conditions tend to be under-coded. For example, in a recent Canadian evaluation of ICD-10 codes for obesity[1], their specificity was high (99%), but their sensitivity was only 7.8% overall, and they differentially identified those with more-severe obesity (sensitivity=12.6%) than less-severe (sensitivity=6.8%). Similarly, a high specificity (92%), but low sensitivity (50%), for the ICD codes for

alcohol abuse was reported by another Canadian group[2].

1 Martin B-J, Chen G, Graham M, et al. Coding of obesity in administrative hospital discharge abstract data: accuracy and impact for future research studies. BMC Health Serv Res 2014;14. doi:10.1186/1472-6963-14-70
2 Durand M, Wang Y, Venne F, et al. Diagnostic accuracy of algorithms to identify hepatitis C status, AIDS status, alcohol consumption and illicit drug use among patients living with HIV in an administrative healthcare database: Validation of Administrative Healthcare Database for Study of HIV. Pharmacoepidemiol Drug Saf 2015;24:943–50. doi:10.1002/pds.3808
Identifying smokers on the basis of prescriptions for nicotine replacement therapy is an interesting idea, but we believe the sensitivity would also be low. It would only identify smokers who had been dispensed nicotine replacement therapy, and not those who did not seek out this therapy, or who had purchased it without a prescription. As well, this approach would not inform about current smoking status (current vs. former) or pack-years of smoking.

You say on page 12 that each of the two surveys comprises 6 sections. So which is the survey outlined on pp 13 and 14? is this the first survey or are both summarised here?

Thank you for the opportunity to clarify the content of the two surveys. Each survey is comprised of the six sections on pages 13 and 14. The second survey is nearly identical to the first, with a couple of exceptions:

a) Section 1 does not include questions on items that would not change (or would not be expected to change) since Survey #1, such as sex, year of birth, or educational attainment;
b) Section 5 is shorter and asks about change in employment status since Survey #1.

We now include these details when describing the content of the two surveys on pages 12-14.

Maybe the 1st survey is section 1 since the summary in Table 1 suggests that only section 1 is being summarised. Please could you clarify this.

The first survey included all six sections described on pages 13 and 14. In Table 1 we present the baseline characteristics of the survey respondents, which were derived from data collected in Section 1 of the first survey. Data collected in the other sections of the survey (including the productivity data from Sections 4 and 5) will be reported in future papers.

VERSION 2 – REVIEW

REVIEWER	Elizabeth Arkema Karolinska Institutet, Sweden
REVIEW RETURNED	18-Apr-2017

GENERAL COMMENTS	The authors addressed all of my comments, thank you!

REVIEWER	Mark D Atkinson Swansea University Medical Scool, Swansea, United Kingdom
REVIEW RETURNED	12-Apr-2017

GENERAL COMMENTS	You have provided more detail about the administrative data.

	You have provided expanded information about the questions in the two surveys and this is now much clearer. The addition of ICD9 and ICD10 codes requested by he other reviewer is very helpful. I have no further comments to make.