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)	Model-based Methods for Case Definitions from Administrative Health Data: Application to Rheumatoid Arthritis
AUTHORS	Kroeker, Kristine; Widdifield, Jessica; Muthukumarana, Saman; Jiang, Depeng; Lix, Lisa

VERSION 1 - REVIEW

REVIEWER	Alys Havard Centre for Big Data Research in Health UNSW Sydney Australia
REVIEW RETURNED	07-Mar-2017

GENERAL COMMENTS	INTRODUCTION 1. Unfortunately I was left wondering about the value of the proposed model-based approach to selecting case definitions. It is stated that inferential methods are preferred over the traditional descriptive approach 'because they can provide valuable empirical evidence about the case definition characteristics associated with validity estimates'. As I understood it, however, this model-based approach requires that the descriptive approach be undertaken first (quite extensively to generate enough case definitions to support a multivariable model). If one can simply select the case definition with the highest sensitivity/specificity (or whichever measure of validity is prioritised by the user), I do not see what is added by further modelling all the case definitions examined descriptively. Also, if a model-based approach is preferred, would it not be more efficient to simply build a model directly using the raw data, with potential case definition characteristics as the independent variables, and values on the gold standard as the dependent variable? OBJECTIVES 2. It should be clarified that objective a) relates to the validity of case definitions for RA only 3. The purpose of objective b) is not clear to me. First, I think it could be clarified that it is the findings of the models that are compared. Second, it would be helpful if it was explained how the models differed in their design and what readers could expect to learn from this comparison. METHODS I found the methods hard to follow, and I think more detail and clarification of concepts is required. I think the authors should bear in mind that they are demonstrating the application of this approach, so that other users may apply it to their own data. Specifically: 4. I think it would be clearer to refer to 'characteristics' as 'criteria', where case definitions comprise certain criteria 5. It would be helpful if the unit of analysis was described explicitly in each of the different analyses. For example, if I understood correctly,

	case definitions were the unit of analysis in the inferential analyses, while individuals were the unit of analysis when examining potential collinearity (or perhaps not, does collinearity here indicate that most case definitions that contained one criterion also contained another?). 6. How were characteristics/criteria not present in certain case definitions managed in unadjusted models and the multivariable models? Presumably the criterion was set to missing if it was not present in the case definition, but a missing category is not mentioned in any of the results. 7. It is explained that combinations of case definition characteristics that were collinear were not included in the same model. How were they managed then, were case definitions that included both excluded from the analyses? 8. The rows in Table 2 seem to represent sensitivity and specificity estimates stratified by levels within each characteristic, which I think means they represent an average of sensitivity and specificity across individuals. I don't see how it is possible to calculate sensitivity and specificity per person. I wondered at the appropriateness of some of the methodological decisions made, and would like to see them justified in the manuscript. Specifically: 9. Using an alpha of 0.01 when the sample size was only 91 or 148. It is mentioned that this was to reduce the probability of Type 1 error, but why was this considered a greater risk in this study compared to others that use an alpha of 0.05? 10. On a related note, I had expected to see collinearity between more of the characteristics/criteria (eg RA-related medications including steroids and RA-related medication excluding steroids) and I wonder if this collinearity was not detected because the alpha cut-off was too strict. DISCUSSION 11. It is acknowledged that the proposed method cannot be applied to data from validation studies that have tested only 'a few' case definitions, with the current study based on 148 case definitions. The manuscript should be clearer about how many case definitions are required to support this approach. For example, the RA studies cited in the introduction in which more than 40 case definitions were tested- would the model-based approach be appropriate in those cases? 12. As with my first point, the Discussion does not make the value of this approach any clearer. Perhaps a discussion of whether the findings of the model-based approach and the original descriptive approach would lead to a different decision about which case definition to pursue would be helpful. FIGURES The confidence intervals in the figures are represented by additional point estimates, which I found confusing. Initially I interpreted this as 3 estimates for each level of the characteristic. It would also be helpful if the x-axis was labelled to indicate that points on the left hand side represent validity estimates that are lower than the reference, and vice versa for the right hand side.

REVIEWER	Quan, Hude University of Calgary, Canada
REVIEW RETURNED	10-Mar-2017

GENERAL COMMENTS	Major comments: Study objective in the abstract should be revised. This study used models with beta distribution to analyze the influence of case definition characteristics on the sensitivity and specificity. The model did not allow for selecting the case definitions as the case definition selection is still a process of balance between sensitivity and specificity depending on the cost of false positive and false negative. The sensitivity and specificity were calculated using the different combinations of cases definition characteristics in the same dataset. The sensitivity from different definitions are correlated with each other. How the correlation was taken into account in the models with Beta distribution? The use of model with Beta distribution to model the Youden's index can be problematic as the Youden's index can take any value ranged from -1 to 1. The use of Youden's index for case definition evaluation is unintuitive as the author discussed the second paragraph of discussion (Page 14). Could you explain why you want to incorporate the Youden's index in your analysis? The analysis on the Youden's index seems redundant as the author also used the bivariate model to analyze the sensitivity and specificity. Also another important measures of case definition the author did not mentioned in the analysis is the positive predictive value (PPV). Pls clarify how you used the Spearman's correlation coefficient to assess the relationship between the cases definition characters. How do you ranked the variable such as physician diagnosis variable with value of 0, 1+, 2+ and 3+? Minor comments: On Page 5, the author provided additional 87 cases definitions not reported in publication. Brief introduction about the strategy about case definitions development should be provided. In the adjusted model (Figure 1 to Figure 3), the level of "never" for "diagnosis observation time" coincide with the levels of "0 physician diagnoses" for physician diagnoses. This should be clarified the Figure legend. Line 24 Page 5: the sentence of "the validation of administrative data was conducted using the medical records for …." Should be revised. The study did not validate the administrative data. It used the medical record as the gold standard to validate the case definition developed in administrative data.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer #1

Comment #1: Unfortunately I was left wondering about the value of the proposed model-based approach to selecting case definitions. It is stated that inferential methods are preferred over the traditional descriptive approach 'because they can provide valuable empirical evidence about the case definition characteristics associated with validity estimates'. As I understood it, however, this model-based approach requires that the descriptive approach be undertaken first (quite extensively to generate enough case definitions to support a multivariable model). If one can simply select the case definition with the highest sensitivity/specificity (or whichever measure of validity is prioritised by the user), I do not see what is added by further modelling all the case definitions examined descriptively. Also, if a model-based approach is preferred, would it not be more efficient to simply build a model directly using the raw data, with potential case definition characteristics as the independent variables, and values on the gold standard as the dependent variable?
Response: Thank you for your comments. Modeling of diagnostic validity estimates can provide empirical evidence about (a) the joint association of case definition criteria with two or more measures of accuracy, such as sensitivity and specify, or (b) the association of case definition criteria with a single measure of accuracy, such as Youden's index. Descriptive analyses are important to ensure correct application of the model-based method, as we have conducted and now describe on page 7:

"Descriptive analyses of the case definition attributes and the estimates of sensitivity, specificity, and Youden's index were conducted using frequencies, percentages, and means to inform the model fitting process."

As well, we have added additional justification for our model-based approach in the Discussion section on page 16:

"The case definition with the highest sensitivity or specificity estimates may not be significantly more accurate than other case definitions. A model-based approach provides empirical evidence about the case definition criteria that are associated with significant increases/decreases in validity estimates."

Individual-level data can be modelled if they are available. However, our model-based methodology can be applied to estimates of sensitivity, specificity, positive predictive value, or negative predictive value as found in published studies. We have noted this strength of our approach in the Discussion section on page 17.

"These methods enable modeling of case definitions from published validation studies and when the individual-level administrative health data are not available.

OBJECTIVES
Comment #2: It should be clarified that objective a) relates to the validity of case definitions for RA only
Response: Thank you for the suggestion. We have included the following in the Introduction section on page 5:

"The model-based method is demonstrated for a rheumatoid arthritis (RA) validation study."

Comment #3: The purpose of objective b) is not clear to me. First, I think it could be clarified that it is the findings of the models that are compared. Second, it would be helpful if it was explained how the

models differed in their design and what readers could expect to learn from this comparison.
Response: Thank you for your suggestions. The objectives have been revised as follows:

“The objectives were to: (a) test the administrative health data criteria associated with the validity of case definitions, and (b) compare competing models applied to case definition validity estimates.”

The model description has been revised on page 7 of the manuscript as follows:

“For the univariate models, sensitivity, specificity, or Youden’s index were the outcome variables and the case definition criteria were covariates. The bivariate mixed-effects model jointly modelled sensitivity and specificity as the outcome variables and the case definition criteria were covariates. In the bivariate model, estimates of sensitivity and specificity were treated as repeated measures to account for their dependence.”

METHODS
Comment #4: I found the methods hard to follow, and I think more detail and clarification of concepts is required. I think the authors should bear in mind that they are demonstrating the application of this approach, so that other users may apply it to their own data. Specifically:
I think it would be clearer to refer to ‘characteristics’ as ‘criteria’, where case definitions comprise certain criteria
Response: Thank you for your suggestion. The term "case definition characteristics” has changed to “case definition criteria” throughout the manuscript.

Comment #5: It would be helpful if the unit of analysis was described explicitly in each of the different analyses. For example, if I understood correctly, case definitions were the unit of analysis in the inferential analyses, while individuals were the unit of analysis when examining potential collinearity (or perhaps not, does collinearity here indicate that most case definitions that contained one criterion also contained another?).
Response: The unit of is the case definition. This has been clarified in the Study Variables section in the Methods on page 6 and Statistical Analysis section in the Methods on page 7.

“The study dataset included each case definition as an observation. The case definition criteria and estimates of sensitivity and specificity were included as variables.”

“Spearman correlation coefficients were used to identify potential collinearity (defined as a correlation of 0.70 or greater [13]) amongst the case definition criteria.”

Comment #6: How were characteristics/criteria not present in certain case definitions managed in unadjusted models and the multivariable models? Presumably the criterion was set to missing if it was not present in the case definition, but a missing category is not mentioned in any of the results.
Response: The only criteria that was missing (i.e., could not be defined for all case definitions) were medication-related criteria. This has been clarified in the Study Variables section in the Methods on page 6 with the following:

“The RA-related medication criteria were set to missing for the case definitions applied to the 20+ age group, because medication data were not available for this age group.”

Comment #7: It is explained that combinations of case definition characteristics that were collinear were not included in the same model. How were they managed then, were case definitions that included both excluded from the analyses?
Response: Thank you for your comment. The following has been added to the Descriptive Analyses section in the Results on page 9:

“These combinations of case definition criteria were not included in the same model; rather, one criterion from each pair was used in a model at a time.”

Comment #8: The rows in Table 2 seem to represent sensitivity and specificity estimates stratified by levels within each characteristic, which I think means they represent an average of sensitivity and specificity across individuals. I don’t see how it is possible to calculate sensitivity and specificity per person.
Response: Clarification has been added to the Statistical Analyses section in the Methods on page 7 and interpretation of Table 1 has been added to the Descriptive Analyses in the Results on Page 9.

“Descriptive analyses of the case definition attributes and estimates of sensitivity, specificity, and Youden’s index were conducted using frequencies, percentages, and means to inform the model fitting process.”

“Compared to the case definitions for the 65+ years age group, the case definitions for the 20+ age group, had slightly lower average estimates of sensitivity (20+ years: 90.9 and 65+ years: 91.3), specificity (20+ years: 82.2 and 65+ years: 86.1), and Youden’s index (20+ years: 73.1 and 65+ years: 77.4).”

Comment #9: I wondered at the appropriateness of some of the methodological decisions made, and would like to see them justified in the manuscript. Specifically:
Using an alpha of 0.01 when the sample size was only 91 or 148. It is mentioned that this was to reduce the probability of Type 1 error, but why was this considered a greater risk in this study compared to others that use an alpha of 0.05?
Response: Thank you for your comment. We conducted a Bonferroni correction calculation to determine a more stringent criteria because multiple significance tests were conducted. The following has been added to the Statistical Analyses section in the Methods on page 8:

“A nominal α = 0.01, based on the Bonferroni correction, was used to evaluate statistical significance in the multivariable model to limit the overall probability of a Type I error [17].”

Comment #10: On a related note, I had expected to see collinearity between more of the characteristics/criteria (eg RA-related medications including steroids and RA-related medication excluding steroids) and I wonder if this collinearity was not detected because the alpha cut-off was too strict.
Response: Thank you for your comment. The majority of the correlations between case definition characteristics were less than 0.25. The covariates of RA-related medications including steroids and RA-medications excluding steroids were not strongly corrected (-0.23; p=0.0285). The following has been included in the Statistical Analyses section in the Methods on page 7 and the Descriptive Analyses section in the Results on page 9 in the manuscript:

“Spearman correlation coefficients were used to identify potential collinearity (defined as a correlation of 0.70 or greater [13]) amongst the case definition criteria.”

“The following case definition criteria were highly correlated (data not shown): exclusion criteria A and B ($r = 0.89$; $p<0.0001$), exclusion criteria A and ≥ 60 days of separation between physician claims ($r = 0.83$; $p<0.0001$), and exclusion criteria B and ≥ 60 days of separation between physician claims ($r = 0.74$; $p<0.0001$). These combinations of case definition criteria were not included in the same model; rather, one criterion from each pair was used in a model at a time.”

DISCUSSION

Comment #11: It is acknowledged that the proposed method cannot be applied to data from validation studies that have tested only 'a few' case definitions, with the current study based on 148 case definitions. The manuscript should be clearer about how many case definitions are required to support this approach. For example, the RA studies cited in the introduction in which more than 40 case definitions were tested- would the model-based approach be appropriate in those cases?
Response: We appreciate this comment and have added the following sentence to the Discussion on page 17:

"Green (1991) suggested a minimum of 50 observations plus eight observations for every parameter estimated from a multiple regression model. Based on this, our model-based approach would require a minimum of 50 case definitions, and preferably more, in order to be implemented [30]."

Comment #12: As with my first point, the Discussion does not make the value of this approach any clearer. Perhaps a discussion of whether the findings of the model-based approach and the original descriptive approach would lead to a different decision about which case definition to pursue would be helpful.
Response: Thank you for your comment. The following has been added to the Discussion on page 16:

"The recommended case definition based on the univariate sensitivity and specificity models is simpler than the recommended case definition from Widdifield et al. (2013); however, both case definitions produce similar diagnostic accuracy estimates. The main difference between the two recommended case definitions is that Widdifield et al. recommended using one diagnosis in hospital discharge records to ascertain cases while our model-based approach did not support this."

FIGURES
Comment #13: The confidence intervals in the figures are represented by additional point estimates, which I found confusing. Initially I interpreted this as 3 estimates for each level of the characteristic. It would also be helpful if the x-axis was labelled to indicate that points on the left hand side represent validity estimates that are lower than the reference, and vice versa for the right hand side.
Response: Thank you for your suggestion. The figures have been updated by changing the x-axis titles to "Logit estimates with 99% confidence intervals" and by representing the confidence intervals by lines with no symbol.

Reviewer # 2

Comment #1: Study objective in the abstract should be revised. This study used models with beta distribution to analyze the influence of case definition characteristics on the sensitivity and specificity. The model did not allow for selecting the case definitions as the case definition selection is still a process of balance between sensitivity and specificity depending on the cost of false positive and false negative.
Response: Thanks for your suggestion. The objective in the abstract on page 2 has been updated as follows:

"This research proposes a model-based method to facilitate the selection of disease case definitions from validation studies for administrative health data."

Comment #2: The sensitivity and specificity were calculated using the different combinations of cases definition characteristics in the same dataset. The sensitivity from different definitions are correlated

with each other. How the correlation was taken into account in the models with Beta distribution?
Response: Null univariate models with a random intercept were initially fit to the data to ensure the correlation between case definitions did not need to be modelled. Univariate fixed-effects models were used since the intra-class correlation was very low.

Comment #3: The use of model with Beta distribution to model the Youden's index can be problematic as the Youden's index can take any value ranged from -1 to 1. The use of Youden's index for case definition evaluation is unintuitive as the author discussed the second paragraph of discussion (Page 14). Could you explain why you want to incorporate the Youden's index in your analysis? The analysis on the Youden's index seems redundant as the author also used the bivariate model to analyze the sensitivity and specificity. Also another important measures of case definition the author did not mentioned in the analysis is the positive predictive value (PPV).
Response: Thank you for your comments. A limitation has been added in the Discussion section on page 17:

"A beta distribution may not always be an appropriate choice for Youden's index, because this index can, in theory, range from -1 to +1. However, in practice, values of Youden's index less than zero are rare."

The following has been added to the Discussion on page 15:

"However, Youden's index provides a simple summary measure that places equal value on sensitivity and specificity [21–23]."

We also acknowledge the researcher's ability to apply our methods to other conditions and to model other accuracy measures.

"Finally, the methods used in this study can be applied to other chronic diseases and other diagnostic validity measures such as positive predictive value and/or negative predictive value."

Comment #4: Pls clarify how you used the Spearman's correlation coefficient to assess the relationship between the cases definition characters. How do you ranked the variable such as physician diagnosis variable with value of 0, 1+, 2+ and 3+?
Response: The following has been added to the Statistical Analyses section in the Methods on page 7:

"All criteria were treated as ordinal measures. Spearman correlation coefficients were used to identify potential collinearity (defined as a correlation of 0.70 or greater [13]) amongst the case definition criteria."

Minor comments:
Comment #5: On Page 5, the author provided additional 87 cases definitions not reported in publication. Brief introduction about the strategy about case definitions development should be provided.
Response: Thank you for your comment. When we were identifying a validation study to demonstrate the model-based methods, Dr. Widdifield was contacted to learn more about her rheumatoid arthritis validation study. She provided us with an additional 87 case definitions that were conducted with the original study. We elected to model both the published and unpublished case definitions as noted in the manuscript.

Comment #6: In the adjusted model (Figure 1 to Figure 3), the level of "never" for "diagnosis observation time" coincide with the levels of "0 physician diagnoses" for physician diagnoses. This

should be clarified the Figure legend.
Response: The following information has been added to page 6 of the manuscript. We felt this was a more suitable location for the information than in the figure legend:

"A case definition with no physician diagnoses corresponds with having no physician diagnosis observation time."

Comment #7: Line 24 Page 5: the sentence of "the validation of administrative data was conducted using the medical records for …." Should be revised. The study did not validate the administrative data. It used the medical record as the gold standard to validate the case definition developed in administrative data.
Response: Thank you for your suggestion. We have revised the Methods section as follows:

"Case definitions for administrative health data were developed using medical records for 450 patients from 18 rheumatology clinics as the gold standard."

VERSION 2 – REVIEW

REVIEWER	Alys Havard UNSW Sydney Australia
REVIEW RETURNED	28-Apr-2017

GENERAL COMMENTS	I would like to congratulate the authors for taking on board the feedback provided. I believe the manuscript is very much improved as a result. I am satisfied that the reviewers' comments have been addressed with the exception of three important points: 1. In the first comment in my previous review, I expressed that I did not understand the value of this model-based approach over the traditional descriptive approach. In their response letter, and in the text added to the Discussion, the authors have now made this clear. It would be helpful if this was made clear earlier in the manuscript ie in the Introduction, so the reader understands the value upfront and is compelled to read the rest of the study. I believe all that is required in the Introduction is a statement about case definitions with the highest Sn and Sp (ascertained using descriptive methods) not necessarily being the best choice. 2. On a related note, I don't know how the following conclusion was drawn, as the sensitivity and specificity estimates for the resulting case definitions were not presented: 'The recommended case definition based on the univariate Sn and Sp models is simpler than the recommended case definition from Widdifield; however both case definitions produce similar diagnostic accuracy estimates'. I wonder at how sensitivity and specificity for the resulting case definitions would even be calculated without the raw data, but I would think knowledge of this would be important to researchers planning to apply a new case definition developed through this model building approach. 3. I am still left unclear of what readers could expect to learn from the comparison of competing models (objective 2). Although the differing findings of each are noted, there is limited interpretation of the different findings eg why they would produce different findings, which might be most applicable in this case, and what a reader is supposed to do when different modelling approaches produce different results. This is touched on in the abstract: 'The choice between univariate and bivariate models depends on the goals of the validation study and the number of case definitions', but this is

	not expanded on elsewhere in the manuscript.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer #1

Comment #1: In the first comment in my previous review, I expressed that I did not understand the value of this model-based approach over the traditional descriptive approach. In their response letter, and in the text added to the Discussion, the authors have now made this clear. It would be helpful if this was made clear earlier in the manuscript ie in the Introduction, so the reader understands the value upfront and is compelled to read the rest of the study. I believe all that is required in the Introduction is a statement about case definitions with the highest Sn and Sp (ascertained using descriptive methods) not necessarily being the best choice.
Response: Thank you for your suggestion. The following sentence has been added to the Introduction on page 4:

"However, the case definition with the highest diagnostic validity estimate may not be more accurate than the case definition with the next highest diagnostic validity estimate, due to sampling error in the estimates."

Comment #2: On a related note, I don't know how the following conclusion was drawn, as the sensitivity and specificity estimates for the resulting case definitions were not presented: 'The recommended case definition based on the univariate Sn and Sp models is simpler than the recommended case definition from Widdifield; however both case definitions produce similar diagnostic accuracy estimates'. I wonder at how sensitivity and specificity for the resulting case definitions would even be calculated without the raw data, but I would think knowledge of this would be important to researchers planning to apply a new case definition developed through this model building approach.
Response: Thank you for your comment. The following has been added to the Discussion on page 16:

"Our recommendation derived from a model-based approach might lead to subsequent re-analysis of the original validation data, to produce estimates of sensitivity and specificity for the model-supported case definition."

Comment #3: I am still left unclear of what readers could expect to learn from the comparison of competing models (objective 2). Although the differing findings of each are noted, there is limited interpretation of the different findings eg why they would produce different findings, which might be most applicable in this case, and what a reader is supposed to do when different modelling approaches produce different results. This is touched on in the abstract: 'The choice between univariate and bivariate models depends on the goals of the validation study and the number of case definitions', but this is not expanded on elsewhere in the manuscript.
Response: Thank you for your comment. The following recommendations for model selection is provided on page 15:

"All of the models resulted in similar performance in our numeric example, but this may not always be the case. Selection of one model over competing alternatives depends on the study goals and the number of case definitions. Overall however, the bivariate model is recommended when the number of case definitions is large and sensitivity and specificity are moderately or highly correlated. The

univariate model applied to Youden's index is recommended when the researcher places equal weight on maximizing sensitivity and specificity [12,19–21]. However, Youden's index can result in the same estimate for different combinations of sensitivity and specificity. Thus, univariate models applied separately to sensitivity and specificity are recommended when the researcher does not place equal weight on these validity measures [22]."

VERSION 3 – REVIEW

REVIEWER	Alys Havard Centre for Big Data Research in Health UNSW Sydney Australia
REVIEW RETURNED	30-May-2017

GENERAL COMMENTS	Thank you for addressing my concerns