PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Linking cohort-based data with electronic health records: a proof-of-concept methodological study in Hong Kong
AUTHORS	Gao, Le; Leung, Miriam T Y; Li, Xue; Chui, Celine; Wong, Rosa Sze Man; Au Yeung, Shiu Lun; Chan, Edward; Chan, Adrienne; Chan, Esther; Wong, HSW; Lee, Tatia; Rao, Nirmala; Wing, Yun-Kwok; Lum, Terry; Leung, Gabriel; Ip, Patrick; Wong, Ian C. K.

VERSION 1 – REVIEW

REVIEWER	Noyd, David Duke University Medical Center, Pediatrics
REVIEW RETURNED	28-Dec-2020

GENERAL COMMENTS	The authors present a detailed approach to link data from cohort studies and electronic health records in Hong Kong. This is certainly an important area of research for health systems and countries to recognize how EHR data can enhance pre-existing cohort studies. The creative approach of a batching technique to use sex and date of birth to link cohort and patient identification numbers is particularly novel. Although it took a couple of times to follow the process, Figure 2 was helpful to understand the process. While larger populations may be tedious to employ this method, as the authors acknowledge, there are certainly a number of cohort and population-based studies that could benefit from this approach. Overall, this manuscript would be of interest for readers to describe and emphasize the methodology to link unique data sources in a local context. The main suggested revisions are with regards to the “Reported Outcomes” section, which are not included in the abstract and can be a distraction. The main objective of the article was to describe and ensure appropriate data quality for the linkage of data from cohort studies and the EHR. The ICD-9-CM codes could certainly be an example of clinical data gleaned from the EHR; however, a focus on either mental health disorders (perhaps just ADHD, as this is elaborated upon in the discussion) or endocrine (overweight/obesity, Type 1 Diabetes Mellitus) would be sufficient and clearly state that this is an example case of a potential application. At least a sentence in the abstract to highlight this as a case example would also clarify the inclusion of this in the manuscript. The other concern for publication is the language standards, as there are some grammatical errors and the overall flow of the manuscript would benefit from and English-language review. Some specific examples, the word “data” is plural and should state “these data are”, etc. at the bottom of page 14, “longitudinal” is misspelled. There are other examples of word choice that could be changed to improve readability and flow.

	Specific Responses from Review Checklist: Is the abstract accurate, balanced, and complete? As above, it would help to include a sentence on ADHD as a case example of information added by the EHR. Are the outcomes clearly defined? The use of equations to illustrate the utility of the described methods to ensure sufficient data quality are clear; however, the “Reported Outcomes” section is confusing and deters from the overall message of the manuscript. Do the results address the research question or objective? Again, Table 2 is a bit confusing. Perhaps, just focusing on psychiatric disorders would be helpful. Alternatively, it may be a good opportunity to emphasize data elements that are captured in the initial cohort and those that are derived from the EHR, as this would further strengthen the importance of this work. In the discussion, it is mentioned that the CEDI cohort includes SWAN data for the ADHD information as well as SES information. The clinical diagnosis from the EHR is included in Table 2; however, the discussion also mentions prescription and admission records. If the ADHD population was used as a case example, the authors could highlight prescriptions for stimulant medications, for example. Is the standard of written English acceptable for publication? As above, there are a number of grammatical errors and word choice that detract from the readability of the manuscript, which need to be addressed to meet publication standards.

REVIEWER	Shi, Xu University of Michigan
REVIEW RETURNED	25-Jan-2021

GENERAL COMMENTS	Comments on \Linking cohort-based data with electronic health records: a proof-of-concept methodological study in Hong Kong" This paper demonstrated a deterministic record linkage method for matching between cohort-based data and EHR data while protecting patient privacy by creating batches of patients according to date of birth (DOB) and sex. The analysis is thorough and the methodology is well developed. The study objective is clearly de ned. The study design is appropriate. The outcomes are clearly de ned, and de nition of match rate is clear. The references are up-to-date. I particularly appre-ciate the thorough discussion on the study limitations. I have a few comments below. The novelty and signi cance of the proposed record linkage method is insu ciently described. Record linkage has a long history with a large number of existing methodologies in the liter-ature. However, there is no review of existing methods in the paper, and it is unclear what the methodological contribution is.

	Unlike the discussion section, the method section is not very clearly written and the paper could bene t a lot from reorganization. In particular, the rationale behind creating batches was not described in the Method section but much later in the Discussion section. It would be helpful to evaluate the computation e ciency of the proposed method by recording the computation time of variations of the proposed method. Line 34-54 of Page 8: just a clari cation question { does information up to Dec 2019 from EHR cover both the ICD-9 and ICD-10 eras? If so, why is ICD-10 not mentioned in identifying diseases diagnoses? Line 27-46 of page 9: if I understand correctly, within each batch, only HKID was used to link the cohort data to EHR data. Could the authors elaborate on why the potential error in DOB and sex may have led to a reduced match rate? If the errors in DOB and sex impact how batches were created, did you need to re-de ne batches after manual correction of DOB and sex records? Line 45-54 of page 14: what are other reasons for underestimation of disease prevalence? For example, evaluation on the quality of EHR data and data quality control may help identify errors. Also, would natural language processing of clinical narrative help improve identi cation of psychiatric disorders and other diseases? Manual validation to obtain gold standard labels (beyond checking the paper-based question-naires) can allow for a more precise evaluation with the true match rate.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. David Noyd, Duke University Medical Center Comments to the Author:
The authors present a detailed approach to link data from cohort studies and electronic health records in Hong Kong. This is certainly an important area of research for health systems and countries to recognize how EHR data can enhance pre-existing cohort studies. The creative approach of a batching technique to use sex and date of birth to link cohort and patient identification numbers is particularly novel. Although it took a couple of times to follow the process, Figure 2 was helpful to understand the process. While larger populations may be tedious to employ this method, as the authors acknowledge, there are certainly a number of cohort and population-based studies that could

benefit from this approach. Overall, this manuscript would be of interest for readers to describe and emphasize the methodology to link unique data sources in a local context.
The main suggested revisions are with regards to the "Reported Outcomes" section, which are not included in the abstract and can be a distraction. The main objective of the article was to describe and ensure appropriate data quality for the linkage of data from cohort studies and the EHR. The ICD-9-CM codes could certainly be an example of clinical data gleaned from the EHR; however, a focus on either mental health disorders (perhaps just ADHD, as this is elaborated upon in the discussion) or endocrine (overweight/obesity, Type 1 Diabetes Mellitus) would be sufficient and clearly state that this is an example case of a potential application. At least a sentence in the abstract to highlight this as a case example would also clarify the inclusion of this in the manuscript.
Author reply: Thank you for your comment. We agree that it is a methodological paper, and in table 2, we need to only focus on one or two specific diseases to show the value of the data linkage. So we have removed most of the current outcomes and only kept ADHD as an example. We also summarised the SWAN results in the CEDI cohort and have a preliminary comparison with the EHRs of ADHD. This will clarify the main part of our research, which is the data linkage. The ADHD section is only used as an example to prove that our data linkage is conducive to the development of future studies in many aspects.
Method: **Page 8, line 39**: "To evaluate the success of our data linkage method, validated HKID rate, CDARS retrieved rate, crude match rate, match rate after checking and total link rate were calculated using the equations in Figure 3. In addition, after the data linkage, we took attention deficit hyperactivity disorder (ADHD) as an example and conducted a simple descriptive analysis in the CEDI cohort to compare the survey results and EHRs in CDARS. In the CEDI cohort, two surveys using the Strengths and Weaknesses of ADHD Symptoms and Normal-Behaviors (SWAN) questionnaire were conducted in the primary school phase (March 2014 - Dec 2015) and the secondary school phase (June 2018 - September 2019). We used both clinical cut-off and alternative (borderline) cut-off21 to identify individuals who scored above the threshold in three domains. Also, EHRs of ADHD in these matched participants were summarised using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code of 314 for the ADHD diagnosis, and the drug name of methylphenidate, atomoxetine, and modafinil for the ADHD medication prescription."
Result Page 12, line 1: "The information of ADHD in the CEDI cohort is summarised in Table 2. In 806 individuals who answered at least one survey, 4.47%, 5.58% and 4.22% of these individuals had an ADHD–Combined score, ADHD–Inattentive score, and ADHD–Hyperactivity/Impulsivity score over the clinical cut-off. After the data linkage, we found 54 individuals had at least one diagnosis of ADHD, and 60 individuals had the prescription record of ADHD medication. Then we compared the ADHD information from the cohort survey and the EHRs. Of the 68 individuals who had a history of ADHD diagnosis or medication treatment, less than 30% of them had scores in three domains above the clinical cut-off and more than half of them had scores above the borderline cut-off."
We also add the description of ADHD outcome in the abstract. **Page 2, line 30**: "After the matching, we conducted a simple descriptive analysis of attention deficit hyperactivity disorder (ADHD) information collected in the CEDI cohort SWAN survey and EHRs." and **line 44** "From our illustration using the ADHD information in the CEDI cohort, 36 (4.47%) individuals had ADHD–Combined score over the clinical cut-off in the SWAN survey, and 68 (8.31%) individuals had ADHD records in EHRs."
The other concern for publication is the language standards, as there are some grammatical errors and the overall flow of the manuscript would benefit from and English-language review. Some specific examples, the word "data" is plural and should state "these data are", etc. at the bottom of page 14, "longitudinal" is misspelled. There are other examples of word choice that could be changed to improve readability and flow.
Author reply: Thank you for your comment. We have had a native English speaker proofread the revised version.
Specific Responses from Review Checklist:
Is the abstract accurate, balanced, and complete? As above, it would help to include a sentence on ADHD as a case example of information added by the EHR.
Author reply: Thank you for your comment. We added a content about using ADHD as an example in the method and result sections of the abstract.
Are the outcomes clearly defined? The use of equations to illustrate the utility of the described methods to ensure sufficient data quality are clear; however, the "Reported Outcomes" section is confusing and deters from the overall message of the manuscript.
Author reply: Thank you for your comment. As mentioned above, we changed our reported outcomes to ADHD only and added some results from the CEDI cohort survey.

Do the results address the research question or objective? Again, Table 2 is a bit confusing. Perhaps, just focusing on psychiatric disorders would be helpful. Alternatively, it may be a good opportunity to emphasize data elements that are captured in the initial cohort and those that are derived from the EHR, as this would further strengthen the importance of this work. In the discussion, it is mentioned that the CEDI cohort includes SWAN data for the ADHD information as well as SES information. The clinical diagnosis from the EHR is included in Table 2; however, the discussion also mentions prescription and admission records. If the ADHD population was used as a case example, the authors could highlight prescriptions for stimulant medications, for example.
Author reply: Thank you for your comment, with which we agree. As there are SWAN results in the CEDI cohort, we focused on the SWAN results and comparison with diagnosis and prescriptions data from the EHR.
Is the standard of written English acceptable for publication? As above, there are a number of grammatical errors and word choice that detract from the readability of the manuscript, which need to be addressed to meet publication standards.
Author reply: Thank you for your comment. The revised version has been proofread by a native English speaker.

Reviewer: 2

This paper demonstrated a deterministic record linkage method for matching between cohort-based data and EHR data while protecting patient privacy by creating batches of patients according to date of birth (DOB) and sex. The analysis is thorough and the methodology is well developed. The study objective is clearly defined. The study design is appropriate. The outcomes are clearly defined, and definition of match rate is clear. The references are up-to-date. I particularly appreciate the thorough discussion on the study limitations. I have a few comments below.

- The novelty and significance of the proposed record linkage method is insufficiently described. Record linkage has a long history with a large number of existing methodologies in the literature. However, there is no review of existing methods in the paper, and it is unclear what the methodological contribution is.
  Author reply: Thank you for your comment. In the previous version, we only focused on the data linkage practice in HK as we thought it depends in part on variables contained in the data in different locales and how it is requested. But we do agree with your comment, and we cited two examples to introduce the current development of record-linkage methodology.
  Page 13, line 11: "Due to the different information contained in each database and the data request method, there are various ways to link to different databases in different parts of the world. For example, Peacock et al33 used the name, address, date of birth and gender as the Master Linkage Key to link the cohort data with other health records; in the UK Biobank, NHS number together with other identifiers (name, date of birth, address, general practice, phone numbers and e-mail addresses) were used for the follow-up and the linkage with EHRs34.".
  In HK, the current studies on data linkage obtained the linkage from the Hong Kong Hospital Authority, which is a one-off linkage. Therefore, this study would like to identify a feasible way of linking cohort data with EHRs in Hong Kong so that more record-linkage studies can be done in Hong Kong using our methods.

- Unlike the discussion section, the method section is not very clearly written and the paper could benefit a lot from reorganization. In particular, the rationale behind creating batches was not described in the Method section but much later in the Discussion section.
  Author reply: Thank you for your comment. We have re-written the reported outcome part in the method section to make it clearer. For the batch generation, we described each step of the linkage in the record-linkage part in the method section **Page 7, line 40**: "Individuals in the two cohorts who provided HKID were included. We completed the record-linkage in 4 steps:"

together with two figures to show the logic of the batch generating and the data linkage. We have numbered each step for greater clarity.

- It would be helpful to evaluate the computation efficiency of the proposed method by recording the computation time of variations of the proposed method.
  Author reply: Thank you for your comment. Given the current sample size in the study, generating batches with the unique combination of date of birth and sex is time-efficient – the task and cross-check can be completed in one hour. The time-consuming part is waiting for data retrieval from CDARS (which may not be applicable to other international datasets) and checking for any mismatched records, and then verifying the original record in the cohort. The time spent on cross-checking and verification will depend on the quality of EHR data and original cohort data, the experience of the record-linkage researcher, and the analytic scenarios that are encountered. Hence the overall time required is difficult to estimate and might vary according to linkage scenarios.
- Line 34-54 of Page 8: just a clarification question - does information up to Dec 2019 from EHR cover both the ICD-9 and ICD-10 eras? If so, why is ICD-10 not mentioned in identifying diseases diagnoses?
  Author reply: Thank you for your comment. In CDARS, all the diagnoses are coded by the International Classification of Diseases, 9th Revision (ICD-9) [*Ref: Validity of major osteoporotic fracture diagnosis codes in the Clinical Data Analysis and Reporting System in Hong Kong*]. ICD-10 is only used for the coding of inpatient records, some ICD-9 undefined rare diseases and cause of death. We changed the example to ADHD only, and for this disease, all the related local studies used ICD-9 code [*Ref: Prenatal antidepressant use and risk of attention-deficit/hyperactivity disorder in offspring: population based cohort study*]. We chose to use ICD-9 codes to make full use of the diagnosis data in CDARS from all settings (inpatient, emergency and outpatient). To complement the ICD-9 codes-based diagnosis, we also included the prescription information of ADHD medication as an indicator of ADHD diagnosis [*Ref: Trends in attention-deficit hyperactivity disorder medication use: a retrospective observational study using population-based databases; Maternal Gestational Diabetes Mellitus, Type 1 Diabetes, and Type 2 Diabetes During Pregnancy and Risk of ADHD in Offspring*].
- Line 27-46 of page 9: if I understand correctly, within each batch, only HKID was used to link the cohort data to EHR data. Could the authors elaborate on why the potential error in DOB and sex may have led to a reduced match rate? If the errors in DOB and sex impact how batches were created, did you need to re-define batches after manual correction of DOB and sex records?
  Author reply: Thank you for your comment. As we mentioned in the method, when we upload the HKID to CDARS, only the patient id (de-identified patient number) will be returned, thus in each batch, we need to use some other identifiable information to do matching, so DOB and sex were used in this study. And for the cohort data, they need to enter the information on hardcopy into the electronic database manually where errors may be introduced, leading to the mismatches. We also mentioned it in the discussion part, **Page 14, line 38**: "<u>One of the obstacles identified in our study was erroneous data entries that arose from the transcription of written responses of the paper questionnaire to the electronic database. We overcame the obstacle by manually checking the physical copies of the questionnaires, which is labour-intensive and therefore not so practical for large cohort studies. Such transcribing errors can be eliminated or reduced by using electronic questionnaires to collect responses in future cohort studies</u>". It would also be better to re-define the batches. But we can also remove the erroneous ones, and identify those individuals with erroneous DOB or sex separately.
- Line 45-54 of page 14: what are other reasons for underestimation of disease prevalence? For example, evaluation on the quality of EHR data and data quality control may help identify errors. Also, would natural language processing of clinical narrative help improve identification of psychiatric disorders and other diseases?
  Author reply: Thank you for your comments. We also added the prescription of ADHD medication to make sure that we can identify as many ADHD patients as possible. The other reasons for underestimation include patients used private service instead of HA, moved out of HK etc. These are common issues with other EHR.
  Due to current privacy law in Hong Kong freetext is not available from our current CDARS dataset so NLP is not applicable.

- Manual validation to obtain gold standard labels (beyond checking the paper-based questionnaires) can allow for a more precise evaluation with the true match rate.
  Author reply: Thank you for your comment. We agree manual validation can potentially obtain gold standard labels. However due to privacy protection, when we upload the HKID to CDARS, only the patient reference-id (psudo patient ID number) will be returned, so that in each batch, we need to use some other identifiable information to do the matching. As we cannot get the HKID of each patient in CDARS, we are unable to manually check the matching. But CDARS has already linked HKID with birth registry with accurate information on dob and sex – this is the gold standard labels, we believe. Thus, we mentioned in the discussion, **Page 13, line 51**: "The identification variables for the exact matching, date of birth and sex, are fixed demographics, which are easy to collect in various types of studies and not subject to recall bias, so the accuracy of these factors is relatively high. Also, CDARS has already linked HKID with birth registry data with accurate information on data of birth and sex, which can be used as the unique identifier within each batch.". Therefore, we think the combination of DOB and sex is a good way to match within each batch.

VERSION 2 – REVIEW

REVIEWER	Shi, Xu University of Michigan
REVIEW RETURNED	31-Mar-2021

GENERAL COMMENTS	Thank you for your thoughtful and thorough responses to my prior comments. I believe this manuscript is a valuable contribution to the literature.