

# BMJ Open

## Linking cohort-based data with electronic health records: a proof-of-concept methodological study in Hong Kong

Le Gao ,<sup>1</sup> Miriam T Y Leung,<sup>1,2</sup> Xue Li,<sup>1,2,3</sup> Celine S L Chui,<sup>2,4,5</sup> Rosa S M Wong,<sup>1,6,7</sup> Shiu Lun Au Yeung,<sup>5</sup> Edward W W Chan,<sup>1,2</sup> Adrienne Y L Chan,<sup>1,2,8</sup> Esther W Chan,<sup>1,2</sup> Wilfred H S Wong,<sup>6</sup> Tatia M C Lee,<sup>9</sup> Nirmala Rao,<sup>10</sup> Yun Kwok Wing,<sup>11</sup> Terry Y S Lum ,<sup>7</sup> Gabriel M Leung,<sup>2,5</sup> Patrick Ip,<sup>6</sup> Ian C K Wong<sup>1,2,12</sup>

**To cite:** Gao L, Leung MTY, Li X, et al. Linking cohort-based data with electronic health records: a proof-of-concept methodological study in Hong Kong. *BMJ Open* 2021;11:e045868. doi:10.1136/bmjopen-2020-045868

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-045868>).

Received 14 October 2020  
Accepted 25 May 2021

### ABSTRACT

**Objectives** Data linkage of cohort-based data and electronic health records (EHRs) has been practised in many countries, but in Hong Kong there is still a lack of such research. To expand the use of multisource data, we aimed to identify a feasible way of linking two cohorts with EHRs in Hong Kong.

**Methods** Participants in the 'Children of 1997' birth cohort and the Chinese Early Development Instrument (CEDI) cohort were separated into several batches. The Hong Kong Identity Card Numbers (HKIDs) of each batch were then uploaded to the Hong Kong Clinical Data Analysis and Reporting System (CDARS) to retrieve EHRs. Within the same batch, each participant has a unique combination of date of birth and sex which can then be used for exact matching, as no HKID will be returned from CDARS. Raw data collected for the two cohorts were checked for the mismatched cases. After the matching, we conducted a simple descriptive analysis of attention deficit hyperactivity disorder (ADHD) information collected in the CEDI cohort via the Strengths and Weaknesses of ADHD Symptoms and Normal Behaviour Scale (SWAN) and EHRs. **Results** In total, 3473 and 910 HKIDs in the birth cohort and CEDI cohort were separated into 44 and 5 batches, respectively, and then submitted to the CDARS, with 100% and 97% being valid HKIDs respectively. The match rates were confirmed to be 100% and 99.75% after checking the cohort data. 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.

**Conclusions** Using date of birth and sex as identifiable variables, we were able to link the cohort data and EHRs with high match rates. This method will assist in the generation of databases for future multidisciplinary research using both cohort data and EHRs.

### Strengths and limitations of this study

- Our study links cohort data with a regionwide electronic healthcare database that covers more than 90% of inpatient services and more than 80% of outpatient services in Hong Kong.
- The use of date of birth and sex as identifiable variables for exact matching is easy and feasible and is highly accurate as it is not likely to be affected by recall bias.
- Privacy is well protected in the process of data linkage through the separate management of different documents.
- The use of date of birth and sex as identifiable variables is less efficient when linking data which needs to be split into many batches.
- Inherent problems within the different data sources, such as erroneous data entries in the cohort data and electronic health records including data from public settings only, can complicate the data linkage process and the use of linked data.



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

**Correspondence to**  
Professor Ian C K Wong;  
wongick@hku.hk

### INTRODUCTION

In epidemiological studies, both cohort-based data and registry/hospital-based

electronic health records (EHRs) are useful data sources, each of them having strengths and weaknesses. Cohort-based surveys usually focus on a specific topic of interest,<sup>1</sup> such as health examination, biological indicators, socioeconomic information, lifestyle information including income, education, exercise and diet, or other qualitative data from questionnaires or interviews. However, they usually have limited years of follow-up with suboptimal follow-up rate<sup>2</sup>; they are labour-intensive for data collection and management,<sup>3</sup> and may lack statistical power or suitable variables to address new research questions beyond the initial cohort establishment due to inadequate sample sizes. Clinical data management systems such as EHRs are

real-time, and recorded as part of daily clinical practice or population management, and usually cover a large population. They include information on diagnosis, prescriptions, laboratory tests and payment information etc, that can facilitate the cost effectiveness of long-term follow-up.<sup>4,5</sup> However, EHRs rely on information routinely collected in clinical settings. Some fundamental risk factors including social, behavioural and environmental factors, and patient-reported outcomes are not well documented in EHRs compared with other epidemiological studies like cohort studies.<sup>4</sup>

Considering the strengths and limitations of different data sources, the opportunity to link data using different data collection methods and across different settings would potentially enable a wider range of research questions to be addressed. With the development of interdisciplinary research and big data analytics, there is a trend of using record-linkage technologies to use the data from different settings. It is also very important to assess the validity and practicability of the record-linkage beforehand to make sure that it is useful for researchers.<sup>6</sup> In many countries including Australia,<sup>7</sup> the USA,<sup>8</sup> Scotland,<sup>9</sup> New Zealand,<sup>10</sup> China<sup>11</sup> etc, data linkage has been practised in medical and social research.

To the best of our knowledge, only one other similar data linkage study has been conducted in Hong Kong. It linked data from the social service databases and EHRs by obtaining the direct linkage from the Hong Kong Hospital Authority (HA).<sup>12</sup> As the data was owned by the Government and it was a one-off linkage, it is not possible to maintain the databases as a longitudinal dataset to evaluate long-term outcomes of children. Therefore, in this study, we aim to identify a feasible way to link data from two previously established cohorts of children and EHRs, to provide methodological fundamentals for the life trajectory and long-term assessment of various health conditions in Hong Kong.

## METHOD

### Data source

We performed the record-linkage of two cohort studies with the Clinical Data Analysis and Reporting System (CDARS), an electronic database used by the public healthcare system in Hong Kong. The 'Children of 1997' birth cohort,<sup>13</sup> established by the School of Public Health at the University of Hong Kong (HKU) and the Department of Health, is one of Asia's largest birth cohorts. The study successfully recruited over 8300 babies born in 1997. Since 2007, direct contact with subjects has been re-established and postal surveys have been regularly conducted in the entire cohort. 3618 subjects participated in the Biobank clinical follow-up study for assessing body composition and provided biospecimens for biobanking from 2013 to 2018. They also consented to record-linkage for future health-related studies. The second cohort is the Chinese Early Development Instrument (CEDI) cohort, which was established in 2011 by

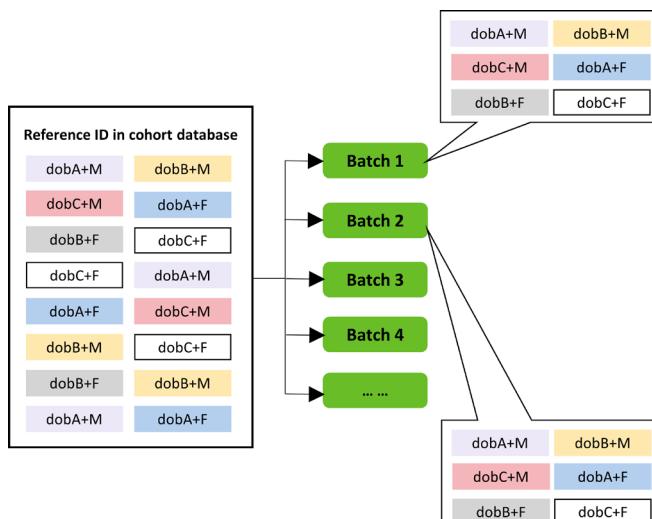
the Department of Paediatrics & Adolescent Medicine at HKU to study the impact of socioeconomic disparity on child health and development. Stratified samples of K3 children from high-income and low-income districts were successfully recruited in 2011/12 (K3, 5–6 years, N=567). These children were followed up in 2014/15 (grade 3, 8–9 years, N=519, N=832 with chain-referral) and 2018/19 (grade 7, 12–13 years, ongoing, expected N=583 with chain referral), respectively, with retention of >80%.<sup>14,15</sup> Parents/guardians of participants in the two cohorts, or participants 18 years or older, were asked to provide informed written consent agreeing to the use of their Hong Kong Identity Card Number (HKID) for record-linkage and longitudinal follow-up for clinical research. Each of them provided their HKID voluntarily.<sup>15,16</sup>

CDARS is an electronic database that includes EHRs since 1995 from all public hospitals and clinics in Hong Kong. It contains anonymised inpatient, outpatient (ambulatory care) and emergency department admissions records to protect patient confidentiality. Information including diagnosis, hospital admissions and discharges, payment method and prescription and dispensing information are recorded in CDARS. Data from CDARS has been validated and used in many previous epidemiological studies on children's neurodevelopment disorders.<sup>17–20</sup>

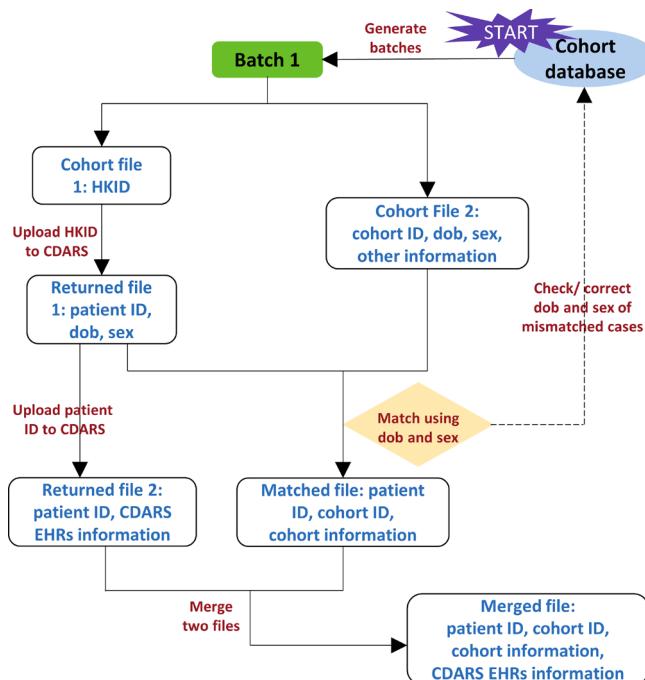
### Record-linkage process

Individuals in the two cohorts who provided HKID were included. We completed the record-linkage in four steps:

1. First, we used the combination of date of birth and sex to generate a reference ID in each cohort database; we then separated all the participants into several batches and ensured, within the same batch, each participant had a unique reference ID (figure 1).
2. Second, we used the HKID in each batch to retrieve their patient ID, sex and date of birth from CDARS. At this stage, the CDARS should return the number of



**Figure 1** Method to generate batches. Dob, date of birth.



**Figure 2** Method to link data from cohort and CDARS in each batch. CDARS, Hong Kong Clinical Data Analysis and Reporting System; Dob, date of birth; EHRs, electronic health records; HKID, Hong Kong Identity Card Number.

- valid HKID uploaded and identify invalid HKIDs if any (equation 1).
3. Due to the protection of patient privacy, only the patient ID, but not the HKID can be returned on request in CDARS. Thus, for records with valid HKID, we used unique combinations of date of birth and sex retrieved from CDARS (equation 2) for further matching in each batch with the information from the cohort database (equation 3) to shorten the matching time.
  4. For those mismatched cases, we checked the raw data collected for the two cohorts (questionnaires in paper format) to exclude the possibility of data entry errors and ensure the highest match rate (equation 4).

To protect data security and patient privacy, we separated the management of cohort ID, HKID and patient ID. The data retrieval process and record-linkage flow are illustrated in figure 2. EC had access to the cohort data including cohort ID (not HKID), generated the matching batches. ML, the only person who had access to both HKID and cohort ID, then uploaded HKID and retrieved patient ID from CDARS data by batches, but was not included in the data management and analysis. LG did the batch splitting independently for quality control as well as the remaining analysis.

### Reported outcomes

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

### Equations:

1. Valid HKID rate =  $\frac{\text{No. of valid HKID}}{\text{No. of submitted HKID}} \times 100\%$ ;
2. Retrieved rate =  $\frac{\text{No. of retrieved records}}{\text{No. of valid HKID}} \times 100\%$ ;
3. Crude match rate =  $\frac{\text{No. of crude matched records}}{\text{No. of retrieved records}} \times 100\%$ ;
4. Match rate after checking =  $\frac{\text{No. of matched records after checking}}{\text{No. of retrieved records}} \times 100\%$ ;
5. Total link rate =  $\frac{\text{No. of matched records after checking}}{\text{No. of submitted HKID}} \times 100\%$ .

**Figure 3** Method to calculate the rate of each step. HKID, Hong Kong Identity Card Number.

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 Behaviour Scale (SWAN) were conducted in the primary school phase (March 2014–December 2015) and the secondary school phase (June 2018–September 2019). We used both clinical cut-off and alternative (borderline) cut-off<sup>21</sup> 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 code of 314 for the ADHD diagnosis, and the British National Formulary chapter 4.4 for the ADHD medication prescription.

Microsoft Excel and R V.3.6.1 were used for data manipulation and analysis.

### Patient and public involvement

This is a methodological study to assess the feasibility of a data linkage method. Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

## RESULTS

In total, at the time of analyses, there were 3473 HKIDs within 44 batches in the birth cohort submitted to the CDARS and all of these HKIDs were valid with successful data retrieval from the system. Of the 3473 children included in the birth cohort, 95.85% had at least one public hospital/clinic attendance up to the end of 2019, and were successfully matched from cohort data to CDARS data. For the 910 children separated into five batches in the CEDI cohort, 889 of them provided valid HKIDs, and 820 of them had records in CDARS. The crude match rate was 93.05%, and the match rate was increased to 99.75% after checking the raw data about the date of birth and sex records in the CEDI cohort. The rate of each match step is shown in table 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

**Table 1** Data linkage rate in each step

	'Children of 1997' birth cohort	CEDI cohort
No. of submitted HKID	3473	910
No. of valid HKID (%)	3473 (100)	889 (97.69)
No. of retrieved records (%)	3329 (95.85)	820 (92.24)
No. of crude matched records (%)	3321 (99.76)	763 (93.05)
No. of matched records after checking (%)	3329 (100)	818 (99.75)
Total link rate (%)	95.85	89.89

CEDI, Chinese Early Development Instrument; HKID, Hong Kong Identity Card Number.

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.

## DISCUSSION

In recent years, with the increasing use of electronic mobile devices, investigation and follow-up in cohort studies have become easier to implement, so a large number of cohort studies were set up and related networks were formed to collaborate, such as the EU Joint Programme-Neurodegenerative Disease Research,<sup>22–23</sup> Collaborative Initiative for Paediatric HIV Education and Research Global Cohort Collaboration<sup>24–25</sup> and Biosocial Birth Cohort Research Network.<sup>26</sup> Meanwhile, many big data networks integrate EHRs for research, for example, the Neurological and mental health Global Epidemiology Network<sup>27–28</sup> and the Asian Pharmacoepidemiology Network.<sup>29–30</sup> These two kinds of data are both valuable for epidemiological research on different topics, with the potential to be used in both policy and social research too. Cohort studies can obtain more detailed and customised variables while EHRs can provide more data that are less subject to attrition or response bias.<sup>31</sup> Therefore, making full use of these two kinds of data will increase the scope for research. There are already good practices for linking cohort studies to EHRs in other countries, for example, the UK Biobank has been linked to different kinds of EHRs.<sup>32</sup> However, there is still a lack of studies that use both cohort studies and EHRs in Hong

**Table 2** Summary of ADHD information in CEDI cohort

	Female	Male	Total
Cohort SWAN information			
No. of individuals answering the survey (%)	359 (44.54)	447 (55.46)	806 (100)
No. of individuals with ADHD-C score over clinical cut-off (%)	11 (3.06)	25 (5.59)	36 (4.47)
No. of individuals with ADHD-I score over clinical cut-off (%)	18 (5.01)	27 (6.04)	45 (5.58)
No. of individuals with ADHD-HI score over clinical cut-off (%)	10 (2.79)	24 (5.37)	34 (4.22)
No. of individuals with ADHD-C score over borderline cut-off (%)	34 (9.47)	105 (23.49)	139 (17.25)
No. of individuals with ADHD-I score over borderline cut-off (%)	69 (19.22)	96 (21.48)	165 (20.47)
No. of individuals with ADHD-HI score over borderline cut-off (%)	52 (14.48)	72 (16.11)	124 (15.38)
CDARS EHRs information			
No. of final matched (%)	366 (44.74)	452 (55.26)	818 (100)
No. of individuals with ADHD diagnosis (%)	14 (3.83)	40 (8.85)	54 (6.60)
No. of individuals with ADHD medication (%)	13 (3.55)	47 (10.40)	60 (7.33)
No. of individuals with ADHD diagnosis or medication (%)	15 (4.10)	53 (11.73)	68 (8.31)
In individuals with ADHD diagnosis or medication			
No. of individuals (%)	15 (22.06)	53 (77.94)	68 (100)
No. of individuals with ADHD-C score over clinical cut-off (%)	0 (0.00)	16 (30.19)	16 (23.53)
No. of individuals with ADHD-I score over clinical cut-off (%)	2 (13.33)	16 (30.19)	18 (26.47)
No. of individuals with ADHD-HI score over clinical cut-off (%)	2 (13.33)	13 (24.53)	15 (22.06)
No. of individuals with ADHD-C score over borderline cut-off (%)	8 (53.33)	36 (67.92)	44 (64.71)
No. of individuals with ADHD-I score over borderline cut-off (%)	11 (73.33)	36 (67.92)	47 (69.12)
No. of individuals with ADHD-HI score over borderline cut-off (%)	8 (53.33)	31 (58.49)	39 (57.35)

ADHD, attention deficit hyperactivity disorder; ADHD-C, ADHD-Combined; ADHD-HI, ADHD-Hyperactivity/Impulsivity; ADHD-I, ADHD-Inattentive; CDARS, Clinical Data Analysis and Reporting System; CEDI, Chinese Early Development Instrument; EHRs, electronic health records; SWAN, Strengths and Weaknesses of ADHD Symptoms and Normal Behaviour Scale.

Kong and examine the feasibility and implications of the linkage.

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 al*<sup>33</sup> 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 email addresses) were used for the follow-up and the linkage with EHRs.<sup>34</sup> In this study, we used date of birth and sex to identify and match the individuals' data across different data sources. The matching rate after checking the original cohort data was 100% for the 'Children of 1997' birth cohort and 99.75% for the CEDI cohort. The total link rates of the two cohorts of 95.85% and 89.89% were lower than the match rates after checking, mainly because we included those without public hospital visits as well as those who provided an invalid HKID in the denominator for calculation. Our link rates were comparable with a similar data linkage study in the United Kingdom,<sup>35</sup> where out of the 90% who gave consent for data linkage, 99% of the Millennium Cohort were linked with birth registration data and 83% linked with hospital record data.

Although we do not have a direct way of linking the data of each individual using their HKIDs collected from the cohort, the use of date of birth and sex to conduct exact matching is an easy and feasible way of avoiding some potentially complex approval processes. The identifiable 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 HKIDs with birth registry data with accurate information on date of birth and sex, which can be used as the unique identifier within each batch. Another advantage of this study is that we can use HKIDs which were collected from cohorts to retrieve data from CDARS followed by exact matching using the date of birth and sex to maintain patient privacy. The use of HKIDs allows us to obtain data from CDARS, but at the same time, CDARS will not return data with HKID, which makes the privacy of non-consented patients well protected. Also, in our study, HKIDs and other cohort information were stored in separate files and kept by different researchers, which further strengthened the protection of privacy.

The first limitation of this study is that we need to split all individuals into several batches so that the individuals in each group have a unique combination of date of birth and sex. There were 44 batches in the 'Children of 1997' birth cohort. Therefore, this method is less efficient when linking data with large sample sizes, for example, millions of individuals, especially in cohorts with relatively concentrated dates of birth because it is time-consuming to split the data into thousands of batches, and then upload them by batch and load the data from CDARS. However,

for a general cohort study, the sample size may not be so large and the dates of birth not too concentrated, so this method can be applied to link cohort studies and EHRs in Hong Kong. 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. Another issue is that the CDARS data are collected by the HA from public hospitals, so that only individuals who had utilised public hospital services can be linked. Only around 5% of our cohort with valid HKIDs had not used public hospitals and were not linked. Similarly, the lower than expected prevalence of the diseases reported may be due to the inclusion of people who do not frequently go to public hospitals, leading to underestimation of the prevalence. In future studies on disease epidemiology, we can consider using the number of individuals who frequently visit the public hospital as the denominator to eliminate such bias.

We linked two cohorts with the EHRs and were able to achieve almost all matching of subjects (both >99%). The resultant longitudinal databases will allow researchers in Hong Kong to conduct long-term studies on neurodevelopmental disorders such as ADHD and Autism Spectrum Disorder. Although many countries have developed longitudinal cohorts (databases or registries) to systematically collect data on patients with ADHD,<sup>36</sup> Hong Kong lacks a comparable cohort and an evidence-based policy to tackle the challenges of treating patients with ADHD locally. Establishing an ADHD cohort with record-linkage from multiple datasets is essential to investigate the long-term impact of ADHD and inform policymakers on effective management and support of patients through their life trajectory. Based on the established cohorts of children in Hong Kong developed by the research teams for various proposes, this study developed a record-linkage model to link project-based data and routine clinical data and assess the impact of ADHD on health outcomes, education attainment and social service utilisation. Data collected in these cohort studies are for specific purposes, and when linking them with EHRs, we are able to obtain more comprehensive information for analysis. Take the CEDI cohort as an example, the SWAN questionnaire was used to identify the ADHD symptoms, and socio-economic information was also available. After linking the cohort data with hospital-based data, not only can we use complementary data, such as the clinical diagnosis, prescription and admission records which are not available in the cohort data but also the socioeconomic information lacking in the hospital-based database, for life-long follow-up.

The linking method established in this study has proved to be effective and, to a large extent, ensures the privacy



of individuals. There are some limitations from cohort studies or medical databases, but overall it will provide a good basis for linking these types of data in the future allowing us to expand the use of richer data resources and to be able to answer further research questions.

## CONCLUSION

This study has demonstrated the feasibility of record-linkage between cohort-based data and hospital-based EHRs with high data linkage rates in Hong Kong using batches of HKID to obtain EHRs and exact matching using date of birth and sex as identifiable variables. The record-linkage methodology and linked database generated from this study will enable future multidisciplinary research in Hong Kong using EHRs.

## Author affiliations

<sup>1</sup>Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong

<sup>2</sup>Laboratory of Data Discovery for Health (D<sup>2</sup>H), Hong Kong Science and Technology Park, Hong Kong, Hong Kong

<sup>3</sup>Department of Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong

<sup>4</sup>School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong

<sup>5</sup>School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong

<sup>6</sup>Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong

<sup>7</sup>Department of Social Work and Social Administration, Faculty of Social Science, The University of Hong Kong, Hong Kong, Hong Kong

<sup>8</sup>Groningen Research Institute of Pharmacy, Unit of PharmacoTherapy, -Epidemiology and -Economics, University of Groningen, Groningen, The Netherlands

<sup>9</sup>Department of Psychology, The University of Hong Kong, Hong Kong, Hong Kong

<sup>10</sup>Faculty of Education, The University of Hong Kong, Hong Kong, Hong Kong

<sup>11</sup>Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong, Hong Kong

<sup>12</sup>Research Department of Practice and Policy, UCL School of Pharmacy, University College London, London, UK

**Twitter** Terry Y S Lum @TerryLum

**Acknowledgements** We would like to thank the Hong Kong Hospital Authority for access to the data from CDARS for research purposes. We also thank Dr Liz Jamieson for proofreading the manuscript.

**Contributors** PI and ICKW conceptualised and designed the study. LG, MTYL, EWWC and AYLC were equally involved in EHRs data collection and management. RSMW, WHSW, PI, SLAY and GML were responsible for quality control of accuracy and integrity of the cohort data. EWWC analysed the data, and LG cross-checked the analysis. LG, MTYL, XL, CSLC, RSMW, SLAY, AYLC, EWC, TMCL, NR, YKW, TYSL, GML, PI and ICKW interpreted the data. LG, MTYL and XL drafted the initial manuscript; XL, CSLC, SLAY, EWWC, AYLC, EWC, TMCL, NR, YKW, TYSL, GML, PI and ICKW critically reviewed the manuscript for important intellectual content. All authors contributed to and approved the final draft. All authors agree to be accountable for all aspects of the work and any issues related to the accuracy or integrity of any part of the work. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

**Funding** This study was supported by Hong Kong Research Grants Council Collaborative Research Fund (No. C7009-19GF).

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not required.

**Ethics approval** The study protocols were approved by the Institutional Review Board of the University of Hong Kong/ Hospital Authority Hong Kong West Cluster (Reference No. UW 13-056 for the CEDI cohort and Reference No. UW13-367 and UW15-412 for 'Children of 1997' birth cohort, Reference No. UW 19-517 for this project). Parents/ guardians of participants or participants 18 years or older, were asked to provide informed written consent agreeing to take part.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on reasonable request. Data from the study can be requested from the corresponding author.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

## ORCID iDs

Le Gao <http://orcid.org/0000-0002-9349-6599>

Terry Y S Lum <http://orcid.org/0000-0003-1196-5345>

## REFERENCES

- March S. Individual data linkage of survey data with claims data in Germany-An overview based on a cohort study. *Int J Environ Res Public Health* 2017;14. doi:10.3390/ijerph14121543. [Epub ahead of print: 09 12 2017].
- Funkhouser E, Vellala K, Baltuck C, et al. Survey methods to optimize response rate in the National dental practice-based research network. *Eval Health Prof* 2017;40:332–58.
- The use of epidemiological tools in conflict-affected populations: open-access educational resources for policy-makers. Available: [http://conflict.lshtm.ac.uk/page\\_51.htm](http://conflict.lshtm.ac.uk/page_51.htm) [Accessed 5 Jun 2020].
- Casey JA, Schwartz BS, Stewart WF, et al. Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health* 2016;37:61–81.
- Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. *Cell* 2019;177:58–69.
- Rivera DR, Gokhale MN, Reynolds MW, et al. Linking electronic health data in pharmacoepidemiology: appropriateness and feasibility. *Pharmacoepidemiol Drug Saf* 2020;29:18–29.
- McHugh L, Andrews RM, Leckning B, et al. Baseline incidence of adverse birth outcomes and infant influenza and pertussis hospitalisations prior to the introduction of influenza and pertussis vaccination in pregnancy: a data linkage study of 78 382 mother-infant pairs, Northern Territory, Australia, 1994–2015. *Epidemiol Infect* 2019;147:e233.
- Lohr AM, Ingram M, Carvajal SC, et al. Protocol for LINKS (linking individual needs to community and clinical services): a prospective matched observational study of a community health worker–community clinical linkage intervention on the U.S.–Mexico border. *BMC Public Health* 2019;19:399.
- Griffiths LJ, Cortina-Borja M, Tingay K, et al. Are active children and young people at increased risk of injuries resulting in hospital admission or accident and emergency department attendance? analysis of linked cohort and electronic Hospital records in Wales and Scotland. *PLoS One* 2019;14:e0213435.
- Donovan GH, Michael YL, Gatziolis D, et al. Association between exposure to the natural environment, rurality, and attention-deficit hyperactivity disorder in children in New Zealand: a linkage study. *Lancet Planet Health* 2019;3:e226–34.
- Yu H-T, Yang Q, Sun X-X, et al. Association of birth defects with the mode of assisted reproductive technology in a Chinese data-linkage cohort. *Fertil Steril* 2018;109:849–56.
- Lo CK-M, Ho FK-W, Chan KL, et al. Linking healthcare and social service databases to study the epidemiology of child maltreatment and associated health problems: Hong Kong's experience. *J Pediatr* 2018;202:291–9.
- Schooling CM, Hui LL, Ho LM, et al. Cohort profile: 'children of 1997': a Hong Kong Chinese birth cohort. *Int J Epidemiol* 2012;41:611–20.
- Tso W, Rao N, Jiang F, et al. Sleep duration and school readiness of Chinese preschool children. *J Pediatr* 2016;169:266–71.
- Ip P, Rao N, Bacon-Shone J, et al. Socioeconomic gradients in school readiness of Chinese preschool children: the mediating role

- of family processes and kindergarten quality. *Early Child Res Q* 2016;35:111–23.
- 16 Liu J, Au Yeung SL, He B, et al. The effect of birth weight on body composition: evidence from a birth cohort and a Mendelian randomization study. *PLoS One* 2019;14:e0222141.
- 17 Man KKC, Chan EW, Ip P, et al. Prenatal antidepressant use and risk of attention-deficit/hyperactivity disorder in offspring: population based cohort study. *BMJ* 2017;357:j2350.
- 18 Man KKC, Coghill D, Chan EW, et al. Association of risk of suicide attempts with methylphenidate treatment. *JAMA Psychiatry* 2017;74:1048–55.
- 19 Man KKC, Lau WCY, Coghill D, et al. Association between methylphenidate treatment and risk of seizure: a population-based, self-controlled case-series study. *Lancet Child Adolesc Health* 2020;4:435–43.
- 20 Raman SR, Man KKC, Bahmanyar S, et al. Trends in attention-deficit hyperactivity disorder medication use: a retrospective observational study using population-based databases. *Lancet Psychiatry* 2018;5:824–35.
- 21 Lai KYC, Leung PWL, Luk ESL, et al. Validation of the Chinese strengths and weaknesses of ADHD-symptoms and normal-behaviors questionnaire in Hong Kong. *J Atten Disord* 2013;17:194–202.
- 22 Adams HHH, Roschupkin GV, DeCarli C, et al. Full exploitation of high dimensionality in brain imaging: the JPND working group statement and findings. *Alzheimers Dement* 2019;11:286–90.
- 23 About JPND. Available: <https://www.neurodegenerationresearch.eu/about/> [Accessed 28 July 2020].
- 24 CIPHER Global Cohort Collaboration. Inequality in outcomes for adolescents living with perinatally acquired HIV in sub-Saharan Africa: a Collaborative Initiative for Paediatric HIV Education and Research (CIPHER) Cohort Collaboration analysis. *J Int AIDS Soc* 2018;21 Suppl 1.
- 25 Collaborative Initiative for Paediatric HIV Education and Research (CIPHER). Available: <https://www.iasociety.org/CIPHER> [Accessed 28 July 2020].
- 26 Biosocial Birth Cohort Research Network BBCR. Available: <https://www.ucl.ac.uk/anthropology/research/biosocial-birth-cohort-research-network-bbccr> [Accessed 28 July 2020].
- 27 Ilomäki J, Bell JS, Chan AYL, et al. Application of healthcare 'Big Data' in CNS drug research: the example of the Neurological and mental health Global Epidemiology Network (NeuroGEN). *CNS Drugs* 2020;34:897–913.
- 28 Neurological and Mental Health Global Epidemiology Network. Available: <https://www.neurogen.hku.hk/> [Accessed 28 Jul 2020].
- 29 Andersen M, Bergman U, et al, AsPEN collaborators. The Asian pharmacoepidemiology network (aspen): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf* 2013;22:700–4.
- 30 Asian Pharmacoepidemiology Network. Available: <https://www.aspensig.asia/> [Accessed 29 Jul 2020].
- 31 Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Ann Hum Biol* 2020;47:218–26.
- 32 About UK Biobank. Available: <https://www.ukbiobank.ac.uk/about-biobank-uk/> [Accessed 9 Sep 2020].
- 33 Peacock A, Chiu V, Leung J, et al. Protocol for the Data-Linkage Alcohol Cohort Study (DACS): investigating mortality, morbidity and offending among people with an alcohol-related problem using linked administrative data. *BMJ Open* 2019;9:e030605.
- 34 UK Biobank study protocol. Available: <https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf> [Accessed 15 Mar 2021].
- 35 Hockley C, Quigley MA, Hughes G, et al. Linking Millennium Cohort data to birth registration and hospital episode records. *Paediatr Perinat Epidemiol* 2008;22:99–109.
- 36 Geltman PL, Fried LE, Arsenault LN, et al. A planned care approach and patient registry to improve adherence to clinical guidelines for the diagnosis and management of attention-deficit/hyperactivity disorder. *Acad Pediatr* 2015;15:289–96.