BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-045868
Article Type:	Original research
Date Submitted by the Author:	14-Oct-2020
Complete List of Authors:	GAO, Le; University of Hong Kong, Department of Pharmacology and Pharmacy Leung, Miriam T Y; University of Hong Kong, Department of Pharmacology and Pharmacy Li, Xue; University of Hong Kong, Department of Medicine; University of Hong Kong, Department of Pharmacology and Pharmacy Chui, Celine; University of Hong Kong, Department of Pharmacology and Pharmacy; University of Hong Kong, Department of Social Work and Social Administration Wong, Rosa Sze Man; University of Hong Kong, 4Department of Paediatrics and Adolescent Medicine Au Yeung, Shiu Lun; University of Hong Kong, School of Public Health Chan, Edward; University of Hong Kong Chan, Adrienne; University of Hong Kong, Department of Pharmacology and Pharmacy Chan, Esther; University of Hong Kong, Department of Pharmacology and Pharmacy Wong, HSW; University of Hong Kong, Department of Paediatrics & Adolescent Medicine Lee, Tatia; University of Hong Kong, Department of Psychology Rao, Nirmala; University of Hong Kong, Faculty of Education Wing, Yun-Kwok; The Chinese University of Hong Kong, Psychiatry Lum, Terry; University of Hong Kong, Department of Social Work and Social Administration Leung, Gabriel; University of Hong Kong, School of Public Health Ip, Patrick; University of Hong Kong, Department of Paediatrics and Adolescent Medicine Wong, Ian C. K.; University of Hong Kong, Pharmacology and Pharmacy; UCL, School of Pharmacy
Keywords:	STATISTICS & RESEARCH METHODS, PAEDIATRICS, EPIDEMIOLOGY, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

Le Gao[1], Miriam TY Leung[1], Xue Li[1,2], Celine SL Chui[1,3,4], Rosa S Wong[4], Shiu Lun Au Yeung[5], Edward WW Chan[1], Adrienne YL Chan[1], Esther W Chan[1], Wilfred HS Wong[4], Tatia MC Lee[6], Nirmala Rao[7], YK Wing[8], Terry YS Lum[3], Gabriel M Leung[5], Patrick Ip[4], Ian CK Wong[1,9+]

[1]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

[2]Department of Medicine, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[3]Department of Social Work and Social Administration, Faculty of Social Science, the University of Hong Kong, Hong Kong

[4]Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[5]School of Public Health, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[6]Department of Psychology, the University of Hong Kong, Hong Kong

[7]Faculty of Education, the University of Hong Kong, Hong Kong

[8]Department of Psychiatry, the Chinese University of Hong Kong, Hong Kong

[9]Research Department of Practice and Policy, UCL School of Pharmacy, London, United Kingdom

[+]Corresponding Author: Ian C K Wong, Department of Pharmacology and Pharmacy, 2/F Laboratory Block, 21 Sassoon Road, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Email: wongick@hku.hk, Tel: (852) 3917-9160, Fax: (852) 2817-0859.

Abstract

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

Method Participants in the "Children of 1997" Birth Cohort and the Chinese Early Development Instrument (CEDI) Cohort, who had provided written consent and Hong Kong Identity Card number (HKID) for record-linkage research, were separated into several batches. The HKIDs of each batch was 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. As no HKIDs can be returned upon request in CDARS, the unique combination of date of birth and sex will then be used for exact matching in each batch. Also, raw data collected at the establishment of the two cohorts was checked for the mismatched cases.

Results In total, 3,473 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 crude match rates were 99.76% and 93.05% in the two cohorts, and the match rates were confirmed to be 100% and 99.75% following checking the original records in the cohort.

Conclusions Using the date of birth and sex as identification variables, we linked the cohort data and hospital-based EHRs with high match rates. This method and the generated database will provide fundamentals for future multi-disciplinary research using CDARS.

Strengths and limitations of this study

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

Contribution PI, and ICKW conceptualised and designed the study. LG, MTYL, EWWC, and AYLC were equally involved in data collection, management. RSW, PI, SLAY, and GML were responsible for quality control of accuracy and integrity of data. EWWC analysed the data, and LG crosschecked the analysis. All the authors interpreted the data. LG and MTYL drafted the initial manuscript; All the authors critically reviewed the manuscript for important intellectual content. All authors contributed to the final draft and finally approved it to be published. All authors agreed to be accountable for all aspects of the work for any issue related to the accuracy or integrity of any part of the work. The corresponding author attests that all listed authors meet 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 statement 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.

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

Word count 2734

Introduction

In epidemiological studies, both cohort-based data and registry/hospital-based electronic health records (EHRs) are useful data sources, each of them has their own strengths and weaknesses. Cohort-based surveys usually focus on a specific topic of interest[1], with information such as health examination, biological indicators, socioeconomic information, lifestyle information like income, education, exercise, diet, or other qualitative data from questionnaires or interviews. However, they usually have limited years of follow-up with suboptimal follow-up rate[2]; they are labor-intensive for data collection and management[3], and may lack statistical power or suitable variables to address new research questions beyond the initial cohort establishment due to inadequate sample sizes. For clinical data management systems such as EHRs, they are real-time, recorded as part of daily clinical practice or population management, and usually cover a large population with diagnosis, prescription, laboratory test, and payment information etc that can facilitate the long-term follow-up cost-effectiveness[4 5]. However, EHRs rely on the 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 to other epidemiological studies like cohort studies[4].

Considering the strengths and limitations of different data sources, the opportunity to link data from different data collection methods and settings would expand the potential to address research questions of a broader scope. With the development of interdisciplinary research and big-data analytics, there is a trend of using record-linkage technologies to utilize the data from different settings. It is also very important to assess the validity and practicability before the linkage to make sure that it is useful for researches[6]. In many countries including Australia[7], the US[8], Scotland[9], New Zealand[10], China[11] etc, data-linkage has been practiced in medical research

and social research. To the best of our knowledge, only one other similar data-linking study was conducted in Hong Kong. It linked data from the social service databases and EHRs by getting the direct linkage from the Hong Kong Hospital Authority (HA)[12]. As the data was owned by the Government and it was a one-off linkage, it is not possible to maintin the databases as longtidual dataset to evaluate long-term outcomes of children. Therefore, in this study, we aim to identify a feasible way to link two previously established children cohorts data 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,[13] 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 8,300 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. 3,618 subjects participated in the Biobank clinical follow-up study for assessing body composition and providing biospecimens for biobanking from 2013-2018 where they provided consent for record linkage for future health-related studies. Another 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 a retention of >80%[14 15]. Participants in the two cohorts were asked for informed written consent of using their Hong Kong Identity Card number (HKID) for record-linkage and longitudinal follow-up for clinical research from their parents/guardians, or from the participant who was 18 years or older, and each of them provided their HKID voluntarily[15 16].

CDARS is an electronic database that includes EHRs since 1995 from all public hospitals and clinics in Hong Kong. It contains de-identified 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[17-20].

Record-linkage

Individuals in the two cohorts who provided HKID were included. Firstly, we used the combination of date of birth and sex to generate reference ID in each cohort database; and then we separated all the participants into several batches and ensured, within the same batch, each participant has a unique reference ID (Figure 1). Secondly, we used the HKID in each batch to retrieve their patient ID, sex and date of birth from CDARS. At this step, the CDARS would return the number of valid HKIDs uploaded and identify invalid HKIDs if any (Equation 1). Due to the protection of patient privacy, only the patient ID but not the HKID can be returned upon request in CDARS. Thus for records with valid HKIDs, 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. For those mismatched cases, we checked the raw data collected for the two cohorts (questionnaires in paper format) to exclude the possibilities 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 were 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.

We calculated the rate of each step using the following equations in Figure 3:

Reported outcomes

Demographic information from CDARS including age, sex, and all diagnosis information up to December 2019 (records from inpatient, outpatient, and emergency settings), especially neurodevelopment disorders was described for the final matched individuals. We used the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code to identify these diagnoses, ICD-9-CM code 314 for ADHD, 299.0 for autism, 296.2, 296.3, 296.82, 300.4, 309.0, 309.1, 311 for depression, 297 to 298 for psychosis, 295 for schizophrenia, 345 for epilepsy, 300, 293.84 for anxiety disorders, 303 to 304 for alcohol and substance use disorder, 301 for personality disorder, 278.0 for overweight and obesity and 250.01, 250.03, 250.11, 250.13, 250.21, 250.23, 250.31, 250.33, 250.41, 250.43, 250.51, 250.53, 250.61, 250.63,

250.71, 250.73, 250.81, 250.83, 250.91, 250.93 for type I diabetes mellitus. Microsoft Excel® and R v3.6.1 were used for data manipulation and analysis.

Ethics

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).

Results

In total, at the time of analyses, there were 3,473 HKIDs within 44 batches in the Birth Cohort submitted to the CDARS and all of these HKIDs are valid with successful data retrieval from the system. Among these 3,473 children included in the Birth Cohort, 95.85% have at least one attendance of the public hospitals and clinics up to the end of 2019, and were successfully matched from cohort data to CDARS data. For the 910 children separated in 5 batches in the CEDI cohort, 889 of them provided valid HKID, and 820 of them have 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.

Table 1 Data-linkage rate in each step

	"Children of 1997" Birth Cohort	**CEDI Cohort**
Submitted N	3473	910
Valid N (%)	3473 (100)	889 (97.69)
Retrieved N (%)	3329 (95.85)	820 (92.24)
Crude match N (%)	3321 (99.76)	763 (93.05)
Matched after checking N (%)	3329 (100)	818 (99.75)
Total link rate (%)	95.85	89.89

Table 2 Baseline information of the two cohorts

	"Children of 1997" Birth Cohort			CEDI Cohort		
	Total	**Female**	**Male**	**Total**	**Female**	**Male**
No. of final matched (%)	3329 (100)	1617 (48.57)	1712 (51.43)	818 (100)	366 (44.74)	452 (55.26)
Median age at 31st Dec 2019 (IQR)	22.67 (22.63 to 22.71)	22.67 (22.63 to 22.71)	22.67 (22.63 to 22.71)	13.62 (13.30 to 13.96)	13.64 (13.28 to 14.07)	13.62 (13.30 to 13.93)
No. of patients with psychiatric disorders (%)*						
ADHD	47 (1.41)	7 (0.43)	40 (2.34)	54 (6.60)	14 (3.83)	40 (8.85)
Autism	11 (0.33)	1 (0.06)	10 (0.58)	9 (1.10)	0 (0.00)	9 (1.99)
Depression	7 (0.21)	3 (0.19)	4 (0.23)	0 (0.00)	0 (0.00)	0 (0.00)
Psychosis	1 (0.03)	0 (0.00)	1 (0.06)	1 (0.12)	0 (0.00)	1 (0.22)
Schizophrenia	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Epilepsy	20 (0.60)	12 (0.74)	8 (0.47)	2 (0.24)	2 (0.55)	0 (0.00)
Anxiety disorder	7 (0.21)	3 (0.19)	4 (0.23)	3 (0.37)	2 (0.55)	1 (0.22)
No. of patients with alcohol and substance use disorder (%)*	1 (0.03)	1 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
No. of patients with personality disorder (%)*	2 (0.06)	1 (0.06)	1 (0.06)	1 (0.12)	0 (0.00)	1 (0.22)
No. of patients with overweight and obesity (%)*	23 (0.69)	8 (0.49)	15 (0.88)	10 (1.22)	0 (0.00)	10 (2.21)
No. of patients with type 1 diabetes (%)*	1 (0.03)	0 (0.00)	1 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)

Abbreviation: IQR, interquartile range; ADHD, attention deficit hyperactivity disorder. * summarised the events happened on or before one's 14th birthday.

After the data linkage, we summarised the baseline information using data from CDARS. In the Birth Cohort, 1617 individuals (48.75%) of final matched were female, and the percentage was 44.74% in the CEDI cohort. The median age of these finally matched individuals on 31st December 2019 was 22.67 in the Birth Cohort and 13.62 in the CEDI cohort. Considering the average age of those children in the CEDI cohort, we described the history of psychiatric disorders on or before 14 years old to make the information from these two cohorts more comparable. For psychiatric comorbidities diagnosed before 14 years old, ADHD (1.41%), epilepsy (0.60%), and autism (0.33%) were the top three frequent comorbidities. In the CEDI cohort, more children (6.60%) had the diagnosis of ADHD, but other psychiatry disorders were uncommon (Table 2).

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 (JPND)[21 22], Collaborative Initiative for Paediatric HIV Education and Research (CIPHER) Global Cohort Collaboration[23 24] and Biosocial Birth Cohort Research (BBCR) Network[25]. Meanwhile, many big data networks integrate EHRs for research, for example, the Neurological and mental health Global Epidemiology Network (NeuroGEN)[26 27] and the Asian Pharmacoepidemiology Network (AsPEN)[28 29]. These two kinds of data are both valuable for epidemiological research on different topics, with great potential to be used in policy research and social research too. Cohort studies can obtain more detailed and customised variables while EHRs can provide more data that are less subjected to attrition or response bias[30]. Therefore, making full use of

these two kinds of data will provide a larger research scope. 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[31]. However, there is still a lack of studies that utilize both cohort studies and EHRs in Hong Kong and examine the feasibility and implications of the linkage.

In this study, we used the 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 matched rates after checking, mainly because we included those without public hospital visits as well as those who provide invalid HKID in the denominator for calculation. Our link rates were comparable with a similar data linkage study in the United Kingdom[32], 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 the direct way to link the data of each individual by using their HKID collected from the cohort, the use of date of birth and sex to do exact matching is an easy and feasible way to avoid some potentially complex approval process. Also, 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 subjected to recall bias, so the accuracy of these factors is relatively high. Another advantage of this study is that we can use HKIDs which were collected from cohorts to retrieve data from CDARS and then do exact matching by using the date of birth and sex to maintain patient privacy. The use of HKID 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,

HKID 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, in particular 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 is applicable 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 overcome the obstacle by manually checking the physical copies of the questionnaires, which is labor-intensive and less practical for large cohort studies. Such error can be reduced by using electronic questionnaires to collect responses in future cohort studies, thus eliminating transcribing error. Another issue is that the CDARS data is collected by the HA from public hospitals, so that only individuals who had used service from public hospitals can be linked. Only around 5% of our cohort with valid HKIDs had not utilized 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 the 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 finally got almost all subjects matched (both >99%), and the resultant longtidual databases will allow researchers in Hong Kong to

conduct long-term study on neurodevelopment 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[33], 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 children cohorts 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 utilization. Data collected in these cohort studies are limited for the specific purpose, and when linking them with EHRs, we are able to obtain more comprehensive information for analysis. Take the CEDI as an example, the SWAN (Strength and Weakness of ADHD-symptoms and Normal-behavior) 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 the complementary data, such as the clinical diagnosis, prescription and admission records which are not available in the cohort data as well as the socio-economic information lacking in the hospital-based database, but can also be ascertained for life-long follow-up.

The linking method established in this study has been proven to be effective and to a large extent ensure individual privacy. There are some limitations from cohort studies or medical databases, but overall it provided a good basis for linking these types of data in the future to expand the use of richer data resources and answer more research questions.

Conclusion

Using batches of HKID to get EHRs and then doing exact matching by date of birth and sex as identifiable variables, we demonstrated the feasibility of record-linkage between cohort-based data and hospital-based EHRs with high data linkage rates. The record linkage methodology and linked database generated from this study will provide fundamentals for future multi-disciplinary research.

References

1. 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(12) doi: 10.3390/ijerph14121543 [published Online First: 2017/12/14]
2. 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(3):332-58. doi: 10.1177/0163278715625738 [published Online First: 2016/01/13]
3. The use of epidemiological tools in conflict-affected populations: open-access educational resources for policy-makers [cited 2020 June 5]. Available from: http://conflict.lshtm.ac.uk/page_51.htm accessed June 5 2020.
4. 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. doi: 10.1146/annurev-publhealth-032315-021353 [published Online First: 2015/12/17]
5. Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. *Cell* 2019;177(1):58-69. doi: 10.1016/j.cell.2019.02.039 [published Online First: 2019/03/23]
6. Rivera DR, Gokhale MN, Reynolds MW, et al. Linking electronic health data in pharmacoepidemiology: Appropriateness and feasibility. *Pharmacoepidemiol Drug Saf* 2020;29(1):18-29. doi: 10.1002/pds.4918 [published Online First: 2020/01/18]
7. 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. doi: 10.1017/S0950268819001171 [published Online First: 2019/08/01]
8. 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(1):399. doi: 10.1186/s12889-019-6725-1 [published Online First: 2019/04/13]
9. 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(4):e0213435. doi: 10.1371/journal.pone.0213435 [published Online First: 2019/04/11]
10. 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(5):e226-e34. doi: 10.1016/S2542-5196(19)30070-1 [published Online First: 2019/05/28]
11. Yu HT, Yang Q, Sun XX, et al. Association of birth defects with the mode of assisted reproductive technology in a Chinese data-linkage cohort. *Fertil Steril* 2018;109(5):849-56. doi: 10.1016/j.fertnstert.2018.01.012 [published Online First: 2018/05/21]
12. Lo CK, Ho FK, 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;202:291-99 e1. doi: 10.1016/j.jpeds.2018.06.033 [published Online First: 2018/07/22]
13. He B, Huang JV, Kwok MK, et al. The association of early-life exposure to air pollution with lung function at ~17.5years in the "Children of 1997" Hong Kong Chinese Birth Cohort. *Environ Int* 2019;123:444-50. doi: 10.1016/j.envint.2018.11.073 [published Online First: 2019/01/10]
14. Tso W, Rao N, Jiang F, et al. Sleep Duration and School Readiness of Chinese Preschool Children. *J Pediatr* 2016;169:266-71. doi: 10.1016/j.jpeds.2015.10.064 [published Online First: 2015/11/27]
15. 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 Childhood Research Quarterly* 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(9):e0222141. doi: 10.1371/journal.pone.0222141 [published Online First: 2019/09/11]
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. doi: 10.1136/bmj.j2350 [published Online First: 2017/06/02]
18. Man KKC, Coghill D, Chan EW, et al. Association of Risk of Suicide Attempts With Methylphenidate Treatment. *JAMA psychiatry* 2017;74(10):1048-55. doi: 10.1001/jamapsychiatry.2017.2183 [published Online First: 2017/07/27]
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(6):435-43. doi: 10.1016/S2352-4642(20)30100-0 [published Online First: 2020/05/26]
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(10):824-35. doi: 10.1016/S2215-0366(18)30293-1 [published Online First: 2018/09/18]
21. Adams HHH, Roshchupkin GV, DeCarli C, et al. Full exploitation of high dimensionality in brain imaging: The JPND working group statement and findings. *Alzheimers Dement (Amst)* 2019;11:286-90. doi: 10.1016/j.dadm.2019.02.003 [published Online First: 2019/04/13]
22. ABOUT JPND [cited 2020 July 28]. Available from: https://www.neurodegenerationresearch.eu/about/ accessed 28 July 2020.
23. Collaboration CGC. 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 doi: 10.1002/jia2.25044 [published Online First: 2018/02/28]
24. Collaborative Initiative for Paediatric HIV Education and Research (CIPHER) [cited 2020 July 28]. Available from: https://www.iasociety.org/CIPHER accessed 28 July 2020.

25. Biosocial Birth Cohort Research Network BBCR [cited 2020 July 28]. Available from: https://www.ucl.ac.uk/anthropology/research/biosocial-birth-cohort-research-network-bbcr accessed 28 July 2020.
26. Ilomaki 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 doi: 10.1007/s40263-020-00742-4 [published Online First: 2020/06/24]
27. NEUROLOGICAL AND MENTAL HEALTH GLOBAL EPIDEMIOLOGY NETWORK [cited 2020 July 28]. Available from: https://www.neurogen.hku.hk/ accessed 28 July 2020.
28. As PENc, Andersen M, Bergman U, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf* 2013;22(7):700-4. doi: 10.1002/pds.3439 [published Online First: 2013/05/09]
29. Asian Pharmacoepidemiology Network [cited 2020 July 29]. Available from: https://www.aspensig.asia/ accessed 29 July 2020.
30. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Ann Hum Biol* 2020;47(2):218-26. doi: 10.1080/03014460.2020.1742379 [published Online First: 2020/05/21]
31. About UK Biobank [cited 2020 September 9]. Available from: https://www.ukbiobank.ac.uk/about-biobank-uk/ accessed 9 September 2020.
32. Hockley C, Quigley MA, Hughes G, et al. Linking Millennium Cohort data to birth registration and hospital episode records. *Paediatr Perinat Epidemiol* 2008;22(1):99-109. doi: 10.1111/j.1365-3016.2007.00902.x [published Online First: 2008/01/05]
33. 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(3):289-96. doi: 10.1016/j.acap.2014.12.002 [published Online First: 2015/04/25]

Figure legends

Figure 1 Method to generate batches. Abbreviation: dob, date of birth; M, male; F, female.

Figure 2 Method to link data from cohort and CDARS in each batch. Abbreviation: dob, date of birth; dx, diagnosis information; rx, prescription information.

Figure 3 Method to calculate the rate of each step.

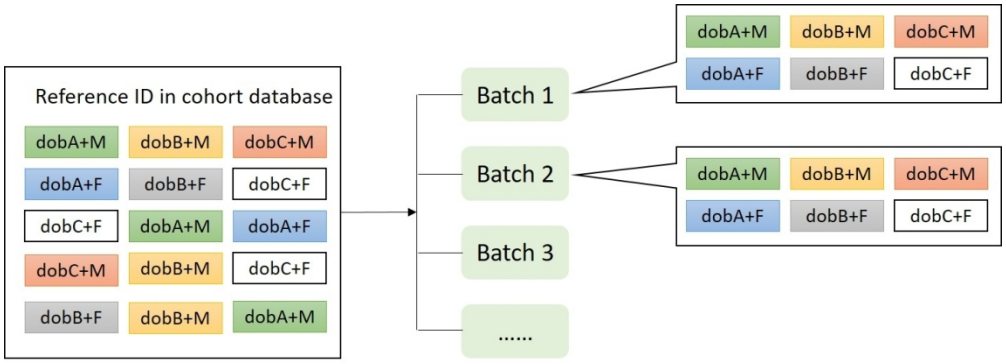

Figure 1 Method to generate batches. Abbreviation: dob, date of birth; M, male; F, female.

245x88mm (150 x 150 DPI)

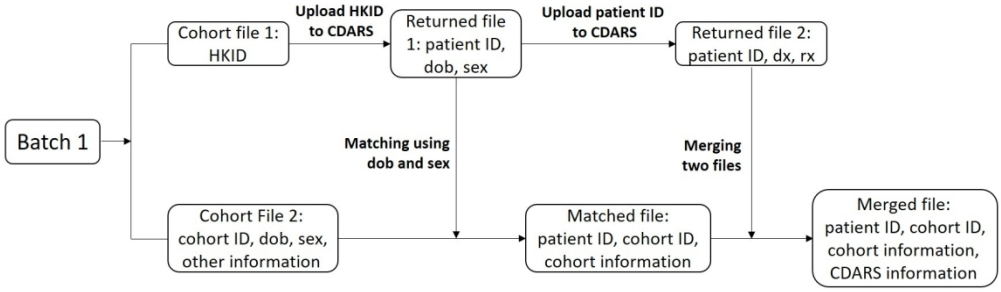

Figure 2 Method to link data from cohort and CDARS in each batch. Abbreviation: dob, date of birth; dx, diagnosis information; rx, prescription information.

295x89mm (150 x 150 DPI)

Equations:

1) $\text{valid Hong Kong Identity ID rate} = \frac{\text{no. of valid HKID}}{\text{no. of submitted HKID}} \times 100\%;$

2) $\text{retrieved rate} = \frac{\text{no. of retrieved records}}{\text{no. of valid HKID}} \times 100\%;$

3) $\text{crude match rate} = \frac{\text{no. of crude match records}}{\text{no. of retrieved records}} \times 100\%;$

4) $\text{match rate after checking} = \frac{\text{no. of matched records after checking}}{\text{no. of retrieved records}} \times 100\%;$

5) $\text{total link rate} = \frac{\text{no. of matched records after checking}}{\text{no. of submitted Hong Kong Identity ID}} \times 100\%.$

Figure 3 Method to calculate the rate of each step.

203x84mm (150 x 150 DPI)

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-045868.R1
Article Type:	Original research
Date Submitted by the Author:	29-Mar-2021
Complete List of Authors:	GAO, Le; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy Leung, Miriam T Y; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy Li, Xue; University of Hong Kong, Department of Medicine; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy Chui, Celine; University of Hong Kong, School of Nursing; University of Hong Kong, School of Public Health Wong, Rosa Sze Man; University of Hong Kong, Department of Paediatrics and Adolescent Medicine Au Yeung, Shiu Lun; University of Hong Kong, School of Public Health Chan, Edward; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy Chan, Adrienne; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy Chan, Esther; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy Wong, HSW; University of Hong Kong, Department of Paediatrics & Adolescent Medicine Lee, Tatia; University of Hong Kong, Department of Psychology Rao, Nirmala; University of Hong Kong, Faculty of Education Wing, Yun-Kwok; The Chinese University of Hong Kong, Department of Psychiatry Lum, Terry; University of Hong Kong, Department of Social Work and Social Administration Leung, Gabriel; University of Hong Kong, School of Public Health Ip, Patrick; University of Hong Kong, Department of Paediatrics and Adolescent Medicine Wong, Ian C. K.; University of Hong Kong, Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy; UCL, Research Department of Practice and Policy, UCL School of Pharmacy
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Paediatrics, Public health, Epidemiology
Keywords:	STATISTICS & RESEARCH METHODS, PAEDIATRICS, EPIDEMIOLOGY, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

Le Gao[1], Miriam TY Leung[1], Xue Li[1,2], Celine SL Chui[3,4], Rosa S Wong[5], Shiu Lun Au Yeung[4], Edward WW Chan[1], Adrienne YL Chan[1], Esther W Chan[1], Wilfred HS Wong[5], Tatia MC Lee[6], Nirmala Rao[7], YK Wing[8], Terry YS Lum[9], Gabriel M Leung[4], Patrick Ip[5], Ian CK Wong[1,10+]

[1]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

[2]Department of Medicine, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[3]School of Nursing, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[4]School of Public Health, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[5]Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong

[6]Department of Psychology, the University of Hong Kong, Hong Kong

[7]Faculty of Education, the University of Hong Kong, Hong Kong

[8]Department of Psychiatry, the Chinese University of Hong Kong, Hong Kong

[9]Department of Social Work and Social Administration, Faculty of Social Science, the University of Hong Kong, Hong Kong

[10]Research Department of Practice and Policy, UCL School of Pharmacy, London, United Kingdom

[+]Corresponding Author: Ian C K Wong, Department of Pharmacology and Pharmacy, 2/F Laboratory Block, 21 Sassoon Road, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Email: wongick@hku.hk, Tel: (852) 3917-9160, Fax: (852) 2817-0859.

Abstract

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

Method 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 HKIDs are returned in 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 SWAN survey and EHRs.

Results In total, 3,473 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 identification 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 multi-disciplinary 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 identification 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 identification 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 EHRs, including data from public settings only, can complicate the data linkage process and the use of linked data.

Contribution PI, and ICKW conceptualised and designed the study. LG, MTYL, EWWC, and AYLC were equally involved in EHRs data collection and management. RSW, 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, RSW, SLAY, AYLC, EWC, TMCL, NR, YKW, TYSL, GML, PI, ICKW interpreted the data. LG, MTYL and XL drafted the initial manuscript; XL, CSLC, 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 statement None declared.

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.

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

Acknowledgments: 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.

Word count 2867

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[1], 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[2]; they are labour-intensive for data collection and management[3], 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[4 5]. 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 to other epidemiological studies like cohort studies[4].

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 utilize 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[6]. In many countries including Australia[7], the US[8], Scotland[9], New Zealand[10], China[11], etc, data-linkage has been practiced 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)[12]. 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,[13] 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 8,300 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. 3,618 subjects participated in the Biobank clinical follow-up study for assessing body composition and provided biospecimens for biobanking from 2013-2018. 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%[14 15]. 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[15 16].

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[17-20].

Record-linkage process

Individuals in the two cohorts who provided HKID were included. We completed the record-linkage in 4 steps:

1) Firstly, 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) Secondly, 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 valid HKIDs 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 upon request in CDARS. Thus, for records with valid HKIDs, 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 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-off[21] 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.

Microsoft Excel® and R v3.6.1 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 3,473 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 3,473 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 5 batches in the CEDI cohort, 889 of them provided valid HKID, 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.

Table 1 Data-linkage rate in each step

	"Children of 1997" Birth Cohort	**CEDI Cohort**
Submitted N	3473	910
Valid N (%)	3473 (100)	889 (97.69)
Retrieved N (%)	3329 (95.85)	820 (92.24)
Crude match N (%)	3321 (99.76)	763 (93.05)
Matched after checking N (%)	3329 (100)	818 (99.75)
Total link rate (%)	95.85	89.89

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 cutoff (%)	11 (3.06)	25 (5.59)	36 (4.47)
No. of individuals with ADHD-I score over clinical cutoff (%)	18 (5.01)	27 (6.04)	45 (5.58)
No. of individuals with ADHD-HI score over clinical cutoff (%)	10 (2.79)	24 (5.37)	34 (4.22)
No. of individuals with ADHD-C score over borderline cutoff (%)	34 (9.47)	105 (23.49)	139 (17.25)
No. of individuals with ADHD-I score over borderline cutoff (%)	69 (19.22)	96 (21.48)	165 (20.47)
No. of individuals with ADHD-HI score over borderline cutoff (%)	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 cutoff (%)	0 (0.00)	16 (30.19)	16 (23.53)
No. of individuals with ADHD-I score over clinical cutoff (%)	2 (13.33)	16 (30.19)	18 (26.47)
No. of individuals with ADHD-HI score over clinical cutoff (%)	2 (13.33)	13 (24.53)	15 (22.06)
No. of individuals with ADHD-C score over borderline cutoff (%)	8 (53.33)	36 (67.92)	44 (64.71)
No. of individuals with ADHD-I score over borderline cutoff (%)	11 (73.33)	36 (67.92)	47 (69.12)
No. of individuals with ADHD-HI score over borderline cutoff (%)	8 (53.33)	31 (58.49)	39 (57.35)

Note: ADHD-C, ADHD–Combined; ADHD-I, ADHD–Inattentive; ADHD-HI, ADHD–Hyperactivity/Impulsivity.

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.

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 (JPND)[22 23], Collaborative Initiative for Paediatric HIV Education and Research (CIPHER) Global Cohort Collaboration[24 25] and Biosocial Birth Cohort Research (BBCR) Network[26]. Meanwhile, many big data networks integrate EHRs for research, for example, the Neurological and mental health Global Epidemiology Network (NeuroGEN)[27 28] and the Asian Pharmacoepidemiology Network (AsPEN)[29 30]. 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[31]. 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[32]. However, there is still a lack of studies that utilize both cohort studies and EHRs in Hong 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[33] 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 EHRs[34]. 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 matched 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[35], 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 HKID 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 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. 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 HKID 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, HKID 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 utilized 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[36], 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 utilization. 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 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 socio-economic 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 multi-disciplinary research in Hong Kong using EHRs.

References

1. 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(12) doi: 10.3390/ijerph14121543 [published Online First: 2017/12/14]
2. 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(3):332-58. doi: 10.1177/0163278715625738 [published Online First: 2016/01/13]
3. The use of epidemiological tools in conflict-affected populations: open-access educational resources for policy-makers [Available from: http://conflict.lshtm.ac.uk/page_51.htm accessed 5 June 2020.
4. 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. doi: 10.1146/annurev-publhealth-032315-021353 [published Online First: 2015/12/17]
5. Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. *Cell* 2019;177(1):58-69. doi: 10.1016/j.cell.2019.02.039 [published Online First: 2019/03/23]
6. Rivera DR, Gokhale MN, Reynolds MW, et al. Linking electronic health data in pharmacoepidemiology: Appropriateness and feasibility. *Pharmacoepidemiol Drug Saf* 2020;29(1):18-29. doi: 10.1002/pds.4918 [published Online First: 2020/01/18]
7. 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. doi: 10.1017/S0950268819001171 [published Online First: 2019/08/01]
8. 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(1):399. doi: 10.1186/s12889-019-6725-1 [published Online First: 2019/04/13]
9. 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(4):e0213435. doi: 10.1371/journal.pone.0213435 [published Online First: 2019/04/11]
10. 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(5):e226-e34. doi: 10.1016/S2542-5196(19)30070-1 [published Online First: 2019/05/28]
11. Yu HT, Yang Q, Sun XX, et al. Association of birth defects with the mode of assisted reproductive technology in a Chinese data-linkage cohort. *Fertil Steril* 2018;109(5):849-56. doi: 10.1016/j.fertnstert.2018.01.012 [published Online First: 2018/05/21]
12. Lo CK, Ho FK, 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;202:291-99 e1. doi: 10.1016/j.jpeds.2018.06.033 [published Online First: 2018/07/22]
13. He B, Huang JV, Kwok MK, et al. The association of early-life exposure to air pollution with lung function at ~17.5years in the "Children of 1997" Hong Kong Chinese Birth Cohort. *Environ Int* 2019;123:444-50. doi: 10.1016/j.envint.2018.11.073 [published Online First: 2019/01/10]
14. Tso W, Rao N, Jiang F, et al. Sleep Duration and School Readiness of Chinese Preschool Children. *J Pediatr* 2016;169:266-71. doi: 10.1016/j.jpeds.2015.10.064 [published Online First: 2015/11/27]

15. 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 Childhood Research Quarterly* 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(9):e0222141. doi: 10.1371/journal.pone.0222141 [published Online First: 2019/09/11]
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. doi: 10.1136/bmj.j2350 [published Online First: 2017/06/02]
18. Man KKC, Coghill D, Chan EW, et al. Association of Risk of Suicide Attempts With Methylphenidate Treatment. *JAMA psychiatry* 2017;74(10):1048-55. doi: 10.1001/jamapsychiatry.2017.2183 [published Online First: 2017/07/27]
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(6):435-43. doi: 10.1016/S2352-4642(20)30100-0 [published Online First: 2020/05/26]
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(10):824-35. doi: 10.1016/S2215-0366(18)30293-1 [published Online First: 2018/09/18]
21. Lai KY, Leung PW, Luk ES, et al. Validation of the Chinese strengths and weaknesses of ADHD-symptoms and normal-behaviors questionnaire in Hong Kong. *J Atten Disord* 2013;17(3):194-202. doi: 10.1177/1087054711430711 [published Online First: 2012/01/03]
22. Adams HHH, Roshchupkin GV, DeCarli C, et al. Full exploitation of high dimensionality in brain imaging: The JPND working group statement and findings. *Alzheimers Dement (Amst)* 2019;11:286-90. doi: 10.1016/j.dadm.2019.02.003 [published Online First: 2019/04/13]
23. ABOUT JPND [Available from: https://www.neurodegenerationresearch.eu/about/ accessed 28 July 2020.
24. Collaboration CGC. 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 doi: 10.1002/jia2.25044 [published Online First: 2018/02/28]
25. Collaborative Initiative for Paediatric HIV Education and Research (CIPHER) [Available from: https://www.iasociety.org/CIPHER accessed 28 July 2020.
26. Biosocial Birth Cohort Research Network BBCR [Available from: https://www.ucl.ac.uk/anthropology/research/biosocial-birth-cohort-research-network-bbcr accessed 28 July 2020.
27. Ilomaki 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 doi: 10.1007/s40263-020-00742-4 [published Online First: 2020/06/24]
28. NEUROLOGICAL AND MENTAL HEALTH GLOBAL EPIDEMIOLOGY NETWORK [Available from: https://www.neurogen.hku.hk/ accessed 28 July 2020.
29. As PENc, Andersen M, Bergman U, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf* 2013;22(7):700-4. doi: 10.1002/pds.3439 [published Online First: 2013/05/09]
30. Asian Pharmacoepidemiology Network [Available from: https://www.aspensig.asia/ accessed 29 July 2020.
31. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Ann Hum Biol* 2020;47(2):218-26. doi: 10.1080/03014460.2020.1742379 [published Online First: 2020/05/21]

32. About UK Biobank [Available from: https://www.ukbiobank.ac.uk/about-biobank-uk/ accessed 9 September 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(8):e030605. doi: 10.1136/bmjopen-2019-030605 [published Online First: 2019/08/07]
34. UK Biobank Study Protocol [Available from: https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf accessed 15 March 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(1):99-109. doi: 10.1111/j.1365-3016.2007.00902.x [published Online First: 2008/01/05]
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(3):289-96. doi: 10.1016/j.acap.2014.12.002 [published Online First: 2015/04/25]

Figure legends

Figure 1 Method to generate batches. Abbreviation: dob, date of birth; M, male; F, female.

Figure 2 Method to link data from cohort and CDARS in each batch. Abbreviation: dob, date of birth; EHRs, electronic health records; CDARS, Hong Kong Clinical Data Analysis and Reporting System.

Figure 3 Method to calculate the rate of each step.

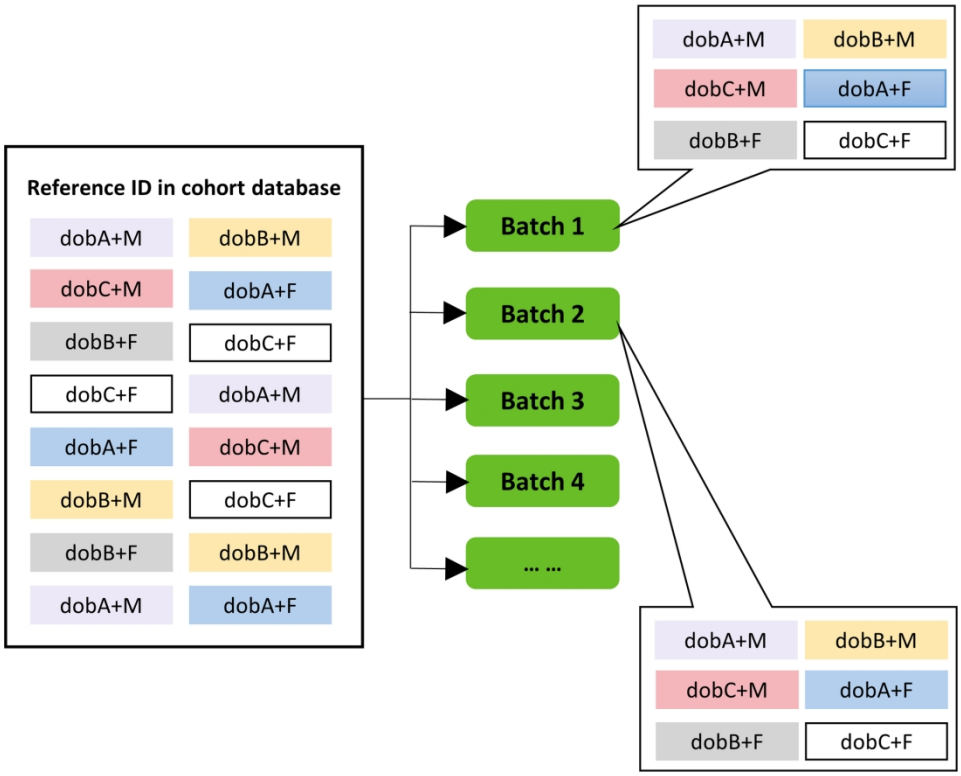

Figure 1 Method to generate batches.
Abbreviation: dob, date of birth; M, male; F, female.

89x89mm (600 x 600 DPI)

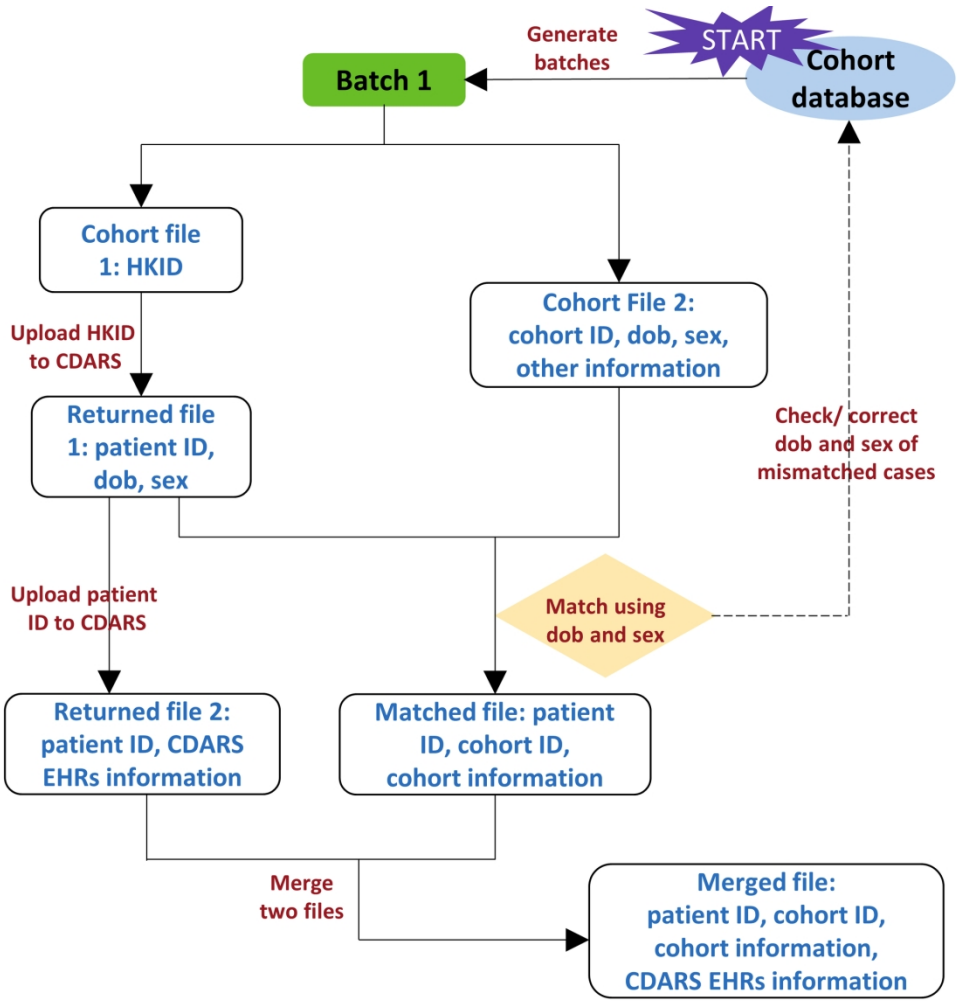

Figure 2 Method to link data from cohort and CDARS in each batch.
Abbreviation: dob, date of birth; EHRs, electronic health records; CDARS, Hong Kong Clinical Data Analysis and Reporting System.

89x89mm (600 x 600 DPI)

Equations:

$$1)\ \text{Valid Hong Kong Identity ID rate} = \frac{\text{No. of valid HKID}}{\text{No. of submitted HKID}} \times 100\%;$$

$$2)\ \text{Retrieved rate} = \frac{\text{No. of retrieved records}}{\text{No. of valid HKID}} \times 100\%;$$

$$3)\ \text{Crude match rate} = \frac{\text{No. of crude match records}}{\text{No. of retrieved records}} \times 100\%;$$

$$4)\ \text{Match rate after checking} = \frac{\text{No. of matched records after checking}}{\text{No. of retrieved records}} \times 100\%;$$

$$5)\ \text{Total link rate} = \frac{\text{No. of matched records after checking}}{\text{No. of submitted Hong Kong Identity ID}} \times 100\%.$$

Figure 3 Method to calculate the rate of each step.

83x48mm (600 x 600 DPI)