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

Le Guo1, Miriam TY Leung1, Xue Li1,2, Celine SL Chui1,3,4, Rona SWong1, Shin Lan Au Yeung1, Edward WWC Chan1, Adrienne YL Chan, Esther W Chan1, Wilfred HS Wong1, Tatia MC Lee1, Nimrana Rao1, YK Wing1, Tarry YS Lam1, Gabriel M Leung1, Patrick Ip1, Ian CK Wong1,6

1Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong; 2Department of Medicine, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong; 3Department of Social Work and Social Administration, Faculty of Social Science, the University of Hong Kong, Hong Kong; 4Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong; 5School of Public Health, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong; 6Department of Psychology, the University of Hong Kong, Hong Kong; 7Facility of Education, the University of Hong Kong, Hong Kong; 8Department of Psychiatry, the Chinese University of Hong Kong, Hong Kong; 9Research Department of Practice and Policy, UCL School of Pharmacy, London, United Kingdom

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

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-sources of data, we aim to identify a feasible way to link two cohorts with EHRs in Hong Kong (HK).

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 number (HKID) for record-linkage research, were separated into several batches and, within the same batch, each participant has a unique combination of date of birth and sex. Then HKID of each batch was uploaded to the Hong Kong Clinical Data Analysis and Reporting System (CDARS) to retrieve EHRs. As no HKID 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 variables, we linked the cohort data and hospital-based data and CDARS in each batch. Abbreviation: dob, date of birth; M, male; F, female.

Strengths and limitations

• 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 HK.
• 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 to be practical.
• 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.

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% were valid HKID respectively. The crude match rate was 99.76% and 93.05% in the two cohorts, and the match rate was confirmed to be 100% and 99.75% following checking the original records in the cohort. See table 1 for details of data-linkage success rate.

Conclusions Using the date of birth and sex as identifiable 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.