Systematic data quality assessment of general practitioner collected smoking information on smoking-associated disease risks

Jiayue Wang[1], A/Prof Freddy Sitas, D.Phil.[2], Jitendra Jonnagaddala, PhD[3]
1. School of Centre for Big Data Research in Health, UNSW
2. Centre for Primary Health Care and Equity, UNSW
3. School of Population Health, UNSW

Background

- Smoking is one of the leading causes of mortality and hospitalisation in Australia [1]. Monitoring the local impact of smoking-related effects is crucial.
- Recording smoking status and related information for patients aged 10 and above is recommended by the Royal Australian College of General Practitioners. Patients at risk also require various interventions depending on different conditions [2]. Thus, patients' smoking information records are essential in clinical visits and treatments.
- The quality of routine smoking data collection, including GP-based primary care data, is uncertain [3].
- The popularity of using GP-based data in research has risen over the years, such strength can be used to improve practices and policies [4].
- AU-ePBRN 2019 Linked Dataset was used to assess the data quality of smoking information on smoking-associated disease risks. This dataset is in the OMOP CDM structure.
- We hypothesised that smoking-related information is more readily collected among patients with respiratory or high-risk smoking-related conditions.

Methods

- This study aimed to assess the validity of collecting smoking-related information and address reporting biases.
- The AU-ePBRN 2019 Linked Dataset includes patient information from Southwest Sydney Local Health District and GP data of over 160,000 patients with clinical and administrative records. With an inclusion criterion of patients aged 10 years and older, the study includes 133,123 patients and 266,246 records.
- To achieve the study goal, three main objectives were involved:
1) Review data's plausibility, conformance and completeness using the OHDSI Data Quality Dashboard (DQD).
2) Investigate the disparity in smoking-status information collection across patients with and without smoking-related conditions using the OHDSI ATLAS and R, verify if the gaps of smoking information collected by GPs exist between patient groups with different conditions.
3) Build statistical models to assess the repeatability of results from existing studies about the association between smoking and some diseases and validate findings through replication.

- Medical conditions include asthma and COPD [5], Myocardial infarctions (MI) [6], a set of cardiovascular diseases (coronary artery disease, myocarditis, heart failure [7] and cardiac arrhythmia [8]), anxiety or fear [9] were involved since the evidenced association with smoking.
- Pregnancy conditions were also involved in the analysis, we expected a high prevalence of smoking status collection among pregnant patients.
- A selection of medical conditions unrelated to smoking was chosen for comparative analysis.
- The majority of cohorts involved in this study used definitions from the Phenotype Library.

Results

- 14.37% of the patients were collected for smoking status. 68.63% were never smokers, 16.70% were current smokers, and 14.67% quit smoking at the time of visit.
- According to the results of the DQD, 92% of the features across all the dimensions passed the data quality check, while plausibility, conformance and completeness checks for all the features had 92%, 95% and 91% pass rates, respectively. Source values representing the medical conditions failed the completeness check. The time interval between two consecutive visits was longer than one week for 99% of the records, which is acceptable for this GP-based setting. But there were 51 errors in gender recording and some implausible visit dates. (e.g. atrophy of prostate recorded for female patients)

Condition groups	Number of patients	Number and % of patients have smoking status
Asthma or COPD	14003	2553 (18.23%)
Myocardial infarctions	813	183 (22.51%)
Other heart conditions	12689	2265 (17.85%)
Pregnancy	2298	1483 (64.53%)
Mental health issue	9764	2548 (26.10%)
Selected smoking-unrelated conditions	52321	10689 (20.43%)
All patients	133123	19124 (14.37%)

- As expected, the smoking status collected from patients with pregnancy conditions was at a higher rate (64.53%) than other conditions. Besides the set of selected cardiovascular diseases, collection rates in all other involved medical conditions were higher than the average of all patients. From Z-tests and Chi-square test, the prevalence of smoking information collected differed in each condition group. Patients' age was not associated with smoking status collection.

- Associations between smoking and mental health issues, heart diseases (not including MI) and asthma or COPD were able to be reproduced using the ePBRN dataset. But the odds ratios of smoking to these diseases were lower than the results from other studies, especially for asthma or COPD. The association between smoking and MI did not present.

Condition	Result using ePBRN-2019	Result from other studies
Anxiety	Odds = 1.01	Odds = 1.03
COPD or asthma	OR = 1.0068	OR = 4.39
Heart disease	OR = 1.1123	OR = 1.72

Conclusions

OHDSI DQD as a data quality check tool was applied in this study, and pointed out some potential improvements that can be applied to data collection in GP settings. Furthermore, the confirmation of the link between patients' medical conditions and the collection of smoking status aligned with our expectations. However, by replicating previous study results, we possessed reservations regarding the overall quality of patients' tobacco smoking information in the ePBRN-2019 dataset. Also, the prevalence of smoking and the response rate to smoking information were much lower compared to other studies. Therefore, the smoking-related information within the study dataset did not meet the necessary criteria for its utilisation for research purposes.

References:
1.Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*. 2006 Nov 28;3(11):e442. https://doi.org/10.1371/journal.pmed.0030442
2.The Royal Australian College of General Practitioners. Guidelines for preventive activities in general practice. 9th edition. https://www.racgp.org.au/getattachment/1ad1a26f-9c8b-4e3c-b45b- 3237272b3a04/Guidelines-for-preventive-activities-in-general-practice.aspx Accessed 16th April 2023.
3.Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public health research & practice*. 2015 Sep 30. doi: http:// dx.doi.org/10.17061/phrp2541540.
4.Youens D, Moorin R, Harrison A, Varhol R, Robinson S, Brooks C, Boyd J. Using general practice clinical information system data for research: the case in Australia. *International Journal of Population Data Science*. 2020;5(1). doi: 10.23889/ijpds.v5i1.1099.
5.Kiljander T, Helin T, Venho K, Jaakkola A, Lehtimäki L. Prevalence of asthma–COPD overlap syndrome among primary care asthmatics with a smoking history: a cross- sectional study. *NPJ primary care respiratory medicine*. 2015 Jul 16;25(1):1-5. https://doi.org/10.1038/npjpcrm.2015.47
6.Gerber Y, Koren-Morag N, Myers V, Benyamini Y, Goldbourt U, Drory Y, Israel Study Group on First Acute Myocardial Infarction. Long-term predictors of smoking cessation in a cohort of myocardial infarction survivors: a longitudinal study. *European Journal of Cardiovascular Prevention & Rehabilitation*. 2011 Jun 1;18(3):533-41. https://doi.org/10.1177/1741826710389371
7.Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. *Nature Reviews Cardiology*. 2011 Jan;8(1):30-41. https://doi.org/10.1038/nrcardio.2010.165
8.D'Alessandro A, Boeckelman I, Hammwhöner M, Goette A. Nicotine, cigarette smoking and cardiac arrhythmia: an overview. *European journal of preventive cardiology*. 2012 Jun 1;19(3):297-305. https://doi.org/10.1177/1741826711411738
9.Fluharty M, Taylor AE, Grabski M, Munafò MR. The association of cigarette smoking with depression and anxiety: a systematic review. *Nicotine & Tobacco Research*. 2016 May 19;19(1):3-13. https://doi.org/10.1093/ntr/ntw140