Data-driven service model to profile healthcare needs and optimise the operation of community-based care: a multi-source data analysis using predictive artificial intelligence

Eman Leung[1], PhD, **Albert Lee**[1,2,3] *, FHKAM (Family Medicine), MD, **Hector Tsang**[2], PhD, **Martin CS Wong**[1,3], FHKAM (Family Medicine), MD

[1] *The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China*

[2] *Department of Rehabilitation Science, The Hong Kong Polytechnic University, Hong Kong SAR, China*

[3] *Centre for Health Education and Health Promotion, The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China*

* Corresponding author: alee@cuhk.edu.hk

This article was published on 13 Dec 2023 at www.hkmj.org.

Hong Kong Med J 2023;29:484–6

https://doi.org/10.12809/hkmj235154

As the needs of our ageing population grow in intensity and diversity, there is a need to achieve precision in public health via data-driven profiling of population-level preventive care, while optimising medical and social services to address those needs. These initiatives will maximise population health and minimise health care costs. Nevertheless, population-level precision public health research is rare; its application to drive service planning and deployment at the population level is even rarer.[1] Thus, with support from the Strategic Public Policy Research Funding Scheme managed by the Policy Innovation and Co-ordination Office of the Hong Kong SAR Government, we initiated a research programme to fill the gap in precision public health research and practice by triangulating data that represent population-level socioecology,[2] such as personal-level clinical and functional data, relational-level data for individual households, community-level data regarding socio-demographic characteristics and physical living environments, data describing organisations that meet population-level needs, and data reflecting the impacts of governmental policy. We sought to identify individuals who can receive the greatest benefit from primary, secondary, and tertiary preventive care. The resulting profiles could inform population-level planning and allocation of the three tiers of preventive care programmes.

Nevertheless, our research objectives were confronted with challenges related to the following contextual factors: (1) the inherent biases and quality of real-world data extracted from medical services' Electronic Health Records (EHRs) and social services' record systems; (2) the fragmentation among services (and their respective databases) which are required to address needs arising from specific aspects of population-level socioecology, including the distinct medical and social needs that our siloed medical and social services seek to address; and (3) the coronavirus disease 2019 (COVID-19) pandemic and the associated social and public health measures which emerged shortly after project initiation and have persisted throughout its life cycle. To overcome these challenges, we adopted a multi-source analytical approach,[3] whereby parallel and iterative analyses were performed across databases representing different socioecology aspects at the resident level. Specifically, an analytical profile developed in one database was applied to other databases with the goal of identifying research questions and facilitating the selection of corresponding features and analytics. The findings from multiple siloed databases could be triangulated to coherently address individual research objectives. In addition, where applicable, parameters extracted from siloed databases were integrated to model particular outcomes using our artificial intelligence (AI) algorithm, for which the input architecture was anthropomorphised[4] according to spheres described in the socioecological prevention framework of the Centers for Disease Control and Prevention. This approach enabled structuring of the hierarchically interrelated input layers.

In the following text, we describe our multi-source analytical approach and emerging findings from our research programme. Although the academic outputs of our research programme are in various stages of peer review, this description of a data-driven process to formulate research questions and develop sampling frames for examination across siloed databases in the construction of a population-level coherent care profile may serve as an alternative approach for other researchers to consider when they face similar contextual challenges in population-level precision public health research.

For example, using the study populations' EHRs (obtained via the Hospital Authority Data Collaboration Laboratory), we applied unsupervised and supervised machine learning algorithms in tandem to identify tertiary prevention needs and the service gaps that prevent those needs from being met in the study populations. Our analyses revealed that the highest rehospitalisation rates (>80%) and the shortest times between discharge and rehospitalisation occurred in sub-populations of patients who lacked specific ambulatory or postacute services. Nonetheless, these services were also available to patients who shared similar clinical and utilisation profiles but exhibited significantly lower rehospitalisation rates. Among the sub-populations with high rehospitalisation rates and low utilisation of rehospitalisation-mitigating post-discharge services, one had a typical profile (ie, population segment medoids) of patients aged 50 to 64 years with musculoskeletal pain–related disorders as primary diagnoses. These patients more frequently exhibited a history of multiple chronic illnesses and higher clinical complexity at index hospitalisation compared with other patients who had similar clinical and acute care utilisation profiles.

The profiling of sub-populations who fell through the service gaps and were rehospitalised at the highest rate enabled us to bring precision to tertiary prevention efforts and subsequently perform data-driven optimisation of population-level post-discharge service allocation, thereby minimising medical costs. Furthermore, the profile we constructed from EHRs could also be applied beyond medical settings to identify potential secondary prevention targets that may exacerbate the evolution of an underlying disease process, such that it interfered with quality of life among individuals who matched the EHR-based and machine-constructed profile, ultimately triggering health-seeking behaviour.

Thus, in a non-medical setting, we recruited residents of the study population aged 50 to 64 years who had musculoskeletal pain, according to community-based primary care clinicians. In addition to the residents' socio-demographic characteristics, behavioural health, and co-morbid chronic illness statuses, clinicians also assessed anthropometric measures and biomarkers of metabolic dysfunction that are often direct or indirect precursors to the most common forms of chronic illnesses. These factors were included as predictive features in a random forest model for selection and risk-scoring of potential secondary prevention targets that could mitigate the exacerbation of pain symptoms. The model also included features representing various aspects of the residents' living environments, which were separately parameterised and initially selected by our AI algorithm according to the following constraints: (1) they were sourced from multiple public domain datasets that belonged to governmental agencies such as the Census and Statistics Department, Housing Authority, Lands Department, Department of Health, and District Offices; (2) they were organised as layered input into a multi-headed hierarchical convolutional neural network, with an anthropomorphised architecture that captured the study population's internal and external built environments and socio-demographic profiles; and (3) they were selected according to the statistical importances of their unique and combined contributions to residential building-level aggregates of general health based on census data and COVID-19 case counts from the Department of Health.

Finally, after parameterisation and selection in accordance with their degrees of importance to the population's general health and COVID-19 susceptibility, features representing the built environments of the study district's residential buildings were processed as follows: (1) they were entered into a random forest model together with the aforementioned individual-level measures to compare their respective importances in the onset of pain interference; and (2) they were scored according to their individual and combined adverse health effects, then assigned to individual residential buildings in the study district for optimised allocation of local primary prevention programmes.

Our analyses revealed that, although features representing residents' socio-demographic characteristics and metabolic dysfunction had high importance with respect to the presence of pain interference in various residential quality of life domains, their feature importances were secondary to the importances of built-environment features, such as living area size, air quality, access to light, architecture conducive to social connectivity, and building age. In addition to scoring the risk of pain interference for individual residents, we scored the built environment of each building in public housing estates within the study district according to the likelihood that its residents would experience sufficient pain to interfere with their quality of life. This scoring approach can inform service planning in geospatially targeted secondary pain prevention programmes.

Patients with chronic obstructive pulmonary disease who exhibited high clinical complexity and multiple co-morbidities were another sub-population who typically exhibited high rehospitalisation rates and low utilisation of rehospitalisation-mitigating post-discharge services. This patient profile was used to guide the recruitment of study district residents outside of medical settings, enabling examination of the evolution of disease processes and hospitalisation trends among asymptomatic and symptomatic community residents. Together with the findings regarding musculoskeletal pain and health-related effects of the built environment, our work has provided the basis for a predictive AI platform that was commissioned by the Sham Shui Po District Office to support its social health surveillance and policy decision needs. Additionally, our work has been incorporated into an algorithm deployed at community diagnosis events hosted by the Sham Shui Po District Office and at events co-hosted by the Kwai Tsing Safe Community and Healthy City Association and the Kwai Tsing District Office.

Acknowledgement

The work described in the current editorial is made possible with the support of the Strategic Public Policy Research Funding Scheme (project No.: S2019.A4.015.19S). The authors thank Dr Jingjing Guan's analytical leadership, Mr Sam Ching's data science management, Ms Olivia Lam's data analytics and visualisation, Ms Yilin He's data wrangling, and Ms Hilliary Yee's data collection. The authors are also deeply grateful for the partnerships of Health In Action and People Service Centre, who have granted us permission to analyse the data under their custodianship under strict confidentiality agreements that safeguard the anonymity of their clients while driving improvements in their respective services.

Author contributions

Concept or design: E Leung, A Lee, H Tsang.
Acquisition of data: E Leung.
Analysis or interpretation of data: E Leung, A Lee.
Drafting of the manuscript: E Leung, A Lee.
Critical revision of the manuscript for important intellectual content: All authors.

All authors had full access to the data, contributed to the study, approved the final version for publication, and take responsibility for its accuracy and integrity.

Conflicts of interest

As editor and adviser of the journal, respectively, MCS Wong and E Leung were not involved in the peer review process. Other authors have disclosed no conflicts of interest.

Funding/support

This editorial is supported by the Strategic Public Policy Research Funding Scheme (project No.: S2019.A4.015.19S). The funder had no role in study design, data collection/analysis/interpretation or manuscript preparation.

References

1. Talias MA, Lamnisos D, Heraclides A. Data science and health economics in precision public health. Front Public Health 2022;10:960282.
2. Centers for Disease Control and Prevention and Health Resources and Services Administration. The social-ecological model: a framework for prevention. 2022. Available from: https://www.cdc.gov/violenceprevention/about/social-ecologicalmodel.html. Accessed 8 Dec 2023.
3. Noi E, Rudolph A, Dodge S. Assessing COVID-induced changes in spatiotemporal structure of mobility in the United States in 2020: a multi-source analytical framework. Int J Geogr Inf Sci 2022;36:585-616.
4. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 2020;14:627-60.