BMJ Open

Identifying non-traditional electronic datasets for population-level surveillance and prevention of cardiometabolic diseases: a scoping review protocol

Reid Rebinsky,[1] Laura N Anderson,[2] Jason D Morgenstern[2]

To cite: Rebinsky R, Anderson LN, Morgenstern JD. Identifying non-traditional electronic datasets for population-level surveillance and prevention of cardiometabolic diseases: a scoping review protocol. *BMJ Open* 2021;**11**:e053485. doi:10.1136/bmjopen-2021-053485

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-053485).

Received 14 May 2021
Accepted 29 July 2021

[1]Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada
[2]Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada

Correspondence to
Reid Rebinsky;
reid.rebinsky@medportal.ca

ABSTRACT

Introduction Cardiometabolic diseases, including cardiovascular disease, obesity and diabetes, are leading causes of death and disability worldwide. Modern advances in population-level disease surveillance are necessary and may inform novel opportunities for precision public health approaches to disease prevention. Electronic data sources, such as social media and consumer rewards points systems, have expanded dramatically in recent decades. These non-traditional datasets may enhance traditional clinical and public health datasets and inform cardiometabolic disease surveillance and population health interventions. However, the scope of non-traditional electronic datasets and their use for cardiometabolic disease surveillance and population health interventions has not been previously reviewed. The primary objective of this review is to describe the scope of non-traditional electronic datasets, and how they are being used for cardiometabolic disease surveillance and to inform interventions. The secondary objective is to describe the methods, such as machine learning and natural language processing, that have been applied to leverage these datasets.

Methods and analysis We will conduct a scoping review following recommended methodology. Search terms will be based on the three central concepts of non-traditional electronic datasets, cardiometabolic diseases and population health. We will search EMBASE, MEDLINE, CINAHL, Scopus, Web of Science and Cochrane Library peer-reviewed databases and will also conduct a grey literature search. Articles published from 2000 to present will be independently screened by two reviewers for inclusion at abstract and full-text stages, and conflicts will be resolved by a separate reviewer. We will report this data as per the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.

Ethics and dissemination No ethics approval is required for this protocol and scoping review, as data will be used only from published studies with appropriate ethics approval. Results will be disseminated in a peer-reviewed publication.

Strengths and limitations of this study

- ► This scoping review addresses a gap in understanding regarding the use of non-traditional electronic data sources for population-level surveillance and prevention of cardiometabolic diseases.
- ► Recommended scoping review guidelines and a rigorous search, including several databases and both peer-reviewed and grey literature, will be used for this study.
- ► This review will be restricted to studies published in English only, and results may not be generalisable globally.
- ► Given the broad research question which spans multiple areas of research, there is the potential that this review may not fully capture all instances of non-traditional electronic datasets and populational-level surveillance and intervention.

INTRODUCTION

Cardiometabolic diseases, defined here as cardiovascular disease (CVD), obesity, dyslipidaemia, metabolic syndrome and diabetes, are leading causes of morbidity and mortality.[1,2] CVD is the most prevalent non-communicable disease and is also the leading cause of death worldwide, resulting in 17.9 million deaths each year.[1,3] Similarly, obesity and diabetes continue to contribute to significant global disability and death.[1,4,5] Due to the burden of cardiometabolic diseases, large-scale population measures aiming to reduce this burden have been employed in many countries.[6,7] However, these cardiometabolic diseases are complex with numerous socioecological risk factors, including behavioural factors, such as diet and exercise, and social determinants of health.[8,9] Therefore, some populations and subgroups are disproportionately affected by cardiometabolic outcomes and may benefit from more targeted population health interventions.[10–12] Population surveillance can be used to better understand how people are differentially affected by cardiometabolic diseases and to inform prevention strategies accordingly.

Surveillance is a critical component of public health that requires population-level data and informs public health interventions.[13 14] Traditionally, data from national surveys, research studies or administrative data are used for this surveillance and evaluation; however, these data have limitations. The response rates of national surveys, such as the National Health and Nutrition Survey, have been declining in recent years, thereby increasing the risk of non-response bias.[15 16] Research studies can be limited by small sample sizes and lack of generalisability, while administrative and electronic health record datasets can be larger, but often still do not capture the entire population.[17] Furthermore, the primary purpose of these data is not always to inform disease surveillance, so key risk factors may be missed.[17] With our society becoming increasingly electronic, complex data from other sources are routinely captured and artificial intelligence methods capable of accommodating these data are advancing.[18 19] Thus, non-traditional data sources need to be better understood.

Non-traditional electronic sources of data include websites, web browsers, apps and other electronic applications and programmes. Unlike traditional data derived from electronic health records, national surveys and research studies, non-traditional electronic sources of data have primary purposes unrelated to encounters with the healthcare system or research but can provide public health-relevant data as a secondary function.

Emerging use of these non-traditional electronic sources of data, such as consumer data generated by airlines and news media, has been shown to support public health predictions for infectious disease.[20 21] Additionally, a previous scoping review that investigated the use of consumer-generated data for public health surveillance found that platforms like Twitter and restaurant review websites can be used to monitor the spread of foodborne diseases.[22] They found that the advantages of these datasets included being available more quickly and containing more information when compared with traditional data.[22] Social media platforms, such as Twitter, continue to be investigated for their applicability in disease surveillance, including cardiometabolic disease surveillance.[23] Some of these non-traditional electronic datasets that have been used for public health prediction and surveillance come from electronic platforms that can also be used to implement population-level prevention strategies. A randomised controlled trial demonstrated that passively collected data from social media could be used to personalise health promotion messages for improved efficacy.[24] It has also been demonstrated that during pandemics, when the public acquires abundant disease-related information via social media, consuming social media can influence disease prevention behaviours.[25] Overall, these types of non-traditional data are applicable to disease surveillance and prevention.

While previous studies have evaluated cardiometabolic surveillance using traditional datasets, such as national survey, research and administrative data,[26 27] there is no published synthesis of the scope of non-traditional electronic datasets and how they have been used for cardiometabolic disease surveillance and prevention. Therefore, the primary objective of this study is to review and map the literature on non-traditional electronic datasets and how they have been applied for surveillance and prevention of cardiometabolic conditions. Our secondary objective is to describe the methods that have been used to make use of these non-traditional datasets.

METHODS AND ANALYSIS

Scoping review

The method for this study will be a scoping review. A scoping review is a type of systematic literature review that investigates an area of research that has yet to be extensively reviewed; it provides a comprehensive overview of the breadth of literature available and identifies where further research may be required.[28] The methodology will follow the recommended scoping review framework.[28 29] This framework describes six steps which can be applied to determine the range of research that is currently available, identify any research gaps and summarise research findings for use by decision-makers. The protocol for this study was also guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews where applicable.[30]

Identifying the research question

As per the methodological framework for scoping reviews, an iterative process was followed to develop our research questions. This included seeking out relevant literature, identifying gaps noted in previous scoping reviews, and discussion with a team including population health researchers and a public health physician.

1. What are the non-traditional electronic sources of data (i.e., not traditional health records or research data) that have been used for studies on cardiometabolic diseases?
2. How have non-traditional electronic data sources been used for cardiometabolic disease surveillance or to inform interventions?
3. What are the methods that have been applied to leverage these non-traditional electronic datasets?

Identifying relevant studies

To comprehensively identify studies relevant in answering the above-mentioned research questions, a broad search was applied across several databases for peer-reviewed literature and grey literature.

The databases EMBASE, MEDLINE, CINAHL, Scopus, Web of Science and Cochrane Library were searched for peer-reviewed literature on 27 January 2021. The search strategy was developed focusing on three key topics: novel data sources, cardiometabolic diseases and public health (table 1). Related systematic and scoping reviews helped inform the choice of terms for the search strategy.[31 32] The

Table 1 Sample database search strategy for Medline

1	**exp social media/**
2	exp wearable electronic devices/
3	exp fitness trackers/
4	exp remote sensing technology
5	exp satellite imagery
6	exp big data/
7	data mining/
8	exp internet/
9	exp search engine/
10	exp smartphone/
11	social media.ti,ab,kw.
12	wearable*.ti,ab,kw.
13	Fitbit.ti,ab,kw.
14	phone record*.ti,ab,kw.
15	consumer purchas*.ti,ab,kw.
16	geospatial*.ti,ab,kw.
17	meteorologic*.ti,ab,kw.
18	air qualit*.ti,ab,kw.
19	healthmap.ti,ab,kw.
20	bluedot.ti,ab,kw.
21	web-based.ti,ab,kw.
22	consumer health information.ti,ab,kw.
23	google*.ti,ab,kw.
24	search engine*.ti,ab,kw.
25	consumer rewards program*.ti,ab,kw
26	air miles.ti,ab,kw.
27	pc optimum.ti,ab,kw.
28	data mining.ti,ab,kw.
29	smartphone*.ti,ab,kw.
30	myfitnesspal.ti,ab,kw.
31	app.ti,ab,kw.
32	flight data.ti,ab,kw.
33	apple watch.ti,ab,kw.
34	big data.ti,ab,kw.
35	exp cardiovascular diseases/
36	exp hypertension/
37	exp obesity/
38	exp body mass index/
39	exp stroke/
40	exp myocardial infarction/
41	exp angina pectoris/
42	exp diabetes mellitus/
43	cardiovascular.mp.
44	hypertension.mp.
45	obesity.mp.
46	body mass index.mp.
47	stroke.mp.
48	myocardial infarction.mp.
49	angina.mp.
50	diabetes.mp.
51	exp social marketing/
52	exp public health surveillance/
53	exp behavioral risk factor surveillance system/
54	exp internet-based intervention/
55	disease prevent*.ti,ab,kw.
56	public health.ti,ab,kw.
57	population health.ti,ab,kw.
58	community medicine.ti,ab,kw.
59	health promot*.ti,ab,kw.
60	health protect*.ti,ab,kw.
61	disease surveillance.ti,ab,kw.
62	infoveillance.ti,ab,kw.
63	infodemiology.ti,ab,kw.
64	computational epidemiology.ti,ab,kw.
65	community health plan*.ti,ab,kw.
66	behavio* change system*.ti,ab,kw.
67	behavio* change support system*.ti,ab,kw.
68	digital health behavio*.ti,ab,kw.
69	preventive intervention*.ti,ab,kw.
70	preventative intervention*.ti,ab,kw.
71	health promotion/
72	preventive medicine/
73	preventive health services/
74	community medicine/
75	public health/
76	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34
77	35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50
78	51 or 52 or 53 or 54 or 55 or 56 or 57 or 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 75
79	76 and 77 and 78 (**2079**)

health research librarians at McMaster University also assisted in optimising the search query to each database. We also conducted a grey literature search in the general Google and Google Scholar databases and reviewed the first 50 hits from each database for inclusion. Our search was limited to articles published since 2000, as our goal is to explore novel, non-traditional datasets and well-documented sources of these types of data, including twitter and google trends,[23 33] were launched in the early 2000s. This systematic search yielded 15 492 articles prior to removing duplicates.

Inclusion criteria

- Uses a non-traditional electronic data source.
  - We will seek out datasets with new features that have not been historically used to inform public health surveillance or intervention strategies, but that can provide relevant public health-related information. Examples of the type of data that will be included are data from websites (eg, search engine data), social media, wearables, consumer rewards systems, etc. Other novel datasets of this nature will be included.
- Pertains to a cardiometabolic disease.
  - Cardiometabolic diseases include hypertension, stroke, ischaemic heart disease, diabetes mellitus, dyslipidaemia, metabolic syndrome and obesity or any other continuous body mass index (BMI) outcomes.
- Includes population or public health surveillance, prediction or informs intervention.
  - Surveillance will include use of data to assess or predict cardiometabolic diseases in groups of people or populations.
  - Interventions will include any programmes, policies or changes implemented by any entities which affect cardiometabolic diseases in groups of people or populations.
- Published in English.
- Published after 2000.

Exclusion criteria

- Studies that include only traditional health data, including data generated from research studies (eg, surveys, physical measures), health records or administrative data.
  - Studies using traditional health data to inform interventions being delivered on electronic platforms, such as wearables, websites and social media will still be excluded.
- Irrelevant to population or public health surveillance, prediction or intervention for cardiometabolic diseases.
  - Studies solely reporting on individuals' opinions of cardiometabolic diseases without any application for cardiometabolic disease surveillance, prediction or intervention will be excluded.
  - Personalised medicine with the potential for upscaling to use at a population level will be included.

Study selection

The results of the searches were imported into Covidence systematic review software to remove any duplicates and host the two-stage selection process. Two independent reviewers will first screen the articles' titles and abstracts to determine if they should be included or excluded based on the criteria provided and conflicts will be resolved by a third reviewer. Any uncertainty about the relevance of an article will not lead to its exclusion at this stage, and instead, the article will be categorised as inconclusive. The second stage of review will include a full-text review of all papers identified as having potential for inclusion or those marked as inconclusive. Two independent reviewers will complete this process. A third reviewer will use the inclusion and exclusion criteria to arbitrate a discussion and resolve any disagreements at this final stage. Any review articles or commentaries identified at this stage will be removed and their reference lists will be searched for other articles that should be included. Currently, we are in the screening stage of the review.

Charting the data

The data charting forms will be created in Covidence. As explained by Levac *et al*, the process of data charting is iterative, and updates will be made as new, relevant data surfaces.[29] Some information we will extract includes article title, first author, year, study design, study location, type of dataset used, dataset availability, cardiometabolic disease, methods applied to dataset, use of machine learning, outcomes, study limitations and method of validation.

Collating, Summarising and reporting the results

The results will be appropriately mapped to provide an overview of the breadth of literature currently available. A narrative summary of data extracted from all included articles will depict how this information relates to our research questions. Any gaps we identify based on this summary will also be noted. We will aim to involve stakeholders and the public so that the diverse data sources we will analyse in this review can help better inform multilevel prevention strategies.

ETHICS AND DISSEMINATION

The role of big data in cardiometabolic diseases is growing. Prior studies have described the use of traditional datasets, such as electronic health records, in population health.[34–36] However, to our knowledge, this will be the first study exploring the use of non-traditional electronic datasets to survey cardiometabolic diseases and inform interventions for these diseases. We hope that this information will aid in precision public health efforts to reduce the cardiometabolic disease burden. Findings of this review will be relevant to researchers, policy-makers and public health officials. The gaps we identify in the literature may also help guide future research.

No ethics approval is required as the proposed scoping review will only include data from previously published articles with appropriate ethical approval. We plan to publish the scoping review in a peer-reviewed journal.

Twitter Reid Rebinsky @RRebinsky

Contributors All authors (RR, LNA and JDM) contributed to the development of the initial research questions and study design, including the search strategy and data

extraction template. RR drafted the initial manuscript and LNA and JDM critically revised the manuscript. All authors approved the final version of the manuscript for publication and are accountable for the accuracy and integrity of the work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

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

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs

Reid Rebinsky http://orcid.org/0000-0002-5386-6284
Laura N Anderson http://orcid.org/0000-0002-6106-5073

REFERENCES

1 Abbafati C, Machado DB, Cislaghi B, *et al*. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *Lancet* 2020;396:1204–22.
2 Roth GA, Abate D, Abate KH, *et al*. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1736–88.
3 World Health Organization. Cardiovascular disease fact sheet: World Health Organization, 2017. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
4 Groce NE. Global disability: an emerging issue. *Lancet Glob Health* 2018;6:e724–5.
5 Bellanger TM, Bray GA. Obesity related morbidity and mortality. *J La State Med Soc* 2005;157 Spec No 1:S42–9.
6 Miranda JJ, Barrientos-Gutiérrez T, Corvalan C, *et al*. Understanding the rise of cardiometabolic diseases in low- and middle-income countries. *Nat Med* 2019;25:1667–79.
7 Benson G, Sidebottom AC, Sillah A, *et al*. Population-Level changes in lifestyle risk factors for cardiovascular disease in the heart of new Ulm project. *Prev Med Rep* 2019;13:332–40.
8 Martínez-García M, Salinas-Ortega M, Estrada-Arriaga I, *et al*. A systematic approach to analyze the social determinants of cardiovascular disease. *PLoS One* 2018;13:e0190960.
9 Elagizi A, Kachur S, Carbone S, *et al*. A review of obesity, physical activity, and cardiovascular disease. *Curr Obes Rep* 2020;9:571–81.
10 Dolley S. Big data's role in precision public health. *Front Public Health* 2018;6:68.
11 Miranda JJ, Carrillo-Larco RM, Ferreccio C, *et al*. Trends in cardiometabolic risk factors in the Americas between 1980 and 2014: a pooled analysis of population-based surveys. *Lancet Glob Health* 2020;8:e123–33.
12 Havranek EP, Mujahid MS, Barr DA, Mujahid Mahasin S, Barr Donald A, *et al*. Social determinants of risk and outcomes for cardiovascular disease. *Circulation* 2015;132:873–98.
13 Nsubuga P, White ME, Thacker SB. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. In: Jamison DT, Breman JG, Measham AR, eds. *Disease control priorities in developing countries*. Washington (DC), New York: The International Bank for Reconstruction and Development/The World Bank, Oxford University Press. Copyright © 2006, The International Bank for Reconstruction and Development/The World Bank Group, 2006.
14 Gilbert R, Cliffe SJ. Public health surveillance. *Public Health Intelligence* 2016:91–110.
15 Fakhouri THI, Martin CB, Chen T-C, *et al*. An investigation of nonresponse bias and survey location variability in the 2017-2018 National health and nutrition examination survey. *Vital Health Stat 2* 2020;2:1–36.
16 Peytchev A. Consequences of survey nonresponse. *Ann Am Acad Pol Soc Sci* 2013;645:88–111.
17 Eggleston EM, Weitzman ER. Innovative uses of electronic health records and social media for public health surveillance. *Curr Diab Rep* 2014;14:468.
18 Hu H, Galea S, Rosella L, *et al*. Big data and population health: focusing on the health impacts of the social, physical, and economic environment. *Epidemiology* 2017;28:759–62.
19 Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep* 2014;16:441.
20 Bogoch II, Brady OJ, Kraemer MUG, German M, *et al*. Anticipating the International spread of Zika virus from Brazil. *Lancet* 2016;387:335–6.
21 Kamel Boulos MN, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *Int J Health Geogr* 2020;19:8.
22 Oldroyd RA, Morris MA, Birkin M. Identifying methods for monitoring foodborne illness: review of existing public health surveillance techniques. *JMIR Public Health Surveill* 2018;4:e57.
23 Sinnenberg L, DiSilvestro CL, Mancheno C, *et al*. Twitter as a potential data source for cardiovascular disease research. *JAMA Cardiol* 2016;1:1032–6.
24 Yom-Tov E, Shembekar J, Barclay S, *et al*. The effectiveness of public health advertisements to promote health: a randomized-controlled trial on 794,000 participants. *NPJ Digit Med* 2018;1:24.
25 Oh S-H, Lee SY, Han C. The effects of social media use on preventive behaviors during infectious disease outbreaks: the mediating role of Self-relevant emotions and public risk perception. *Health Commun* 2021;36:1–10.
26 Matlock DD, Groeneveld PW, Sidney S, *et al*. Geographic variation in cardiovascular procedure use among Medicare fee-for-service vs Medicare advantage beneficiaries. *JAMA* 2013;310:155–61.
27 Reynolds K, Go AS, Leong TK, *et al*. Trends in incidence of hospitalized acute myocardial infarction in the cardiovascular research network (CVRN). *Am J Med* 2017;130:317–27.
28 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8:19–32.
29 Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci* 2010;5:69.
30 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018;169:467–73.
31 Morgenstern JD, Buajitti E, O'Neill M, *et al*. Predicting population health with machine learning: a scoping review. *BMJ Open* 2020;10:e037860.
32 Krittanawong C, Virk HUH, Bangalore S, *et al*. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* 2020;10:16057.
33 Verma M, Kishore K, Kumar M, *et al*. Google search trends predicting disease outbreaks: an analysis from India. *Healthc Inform Res* 2018;24:300–8.
34 Horth RZ, Wagstaff S, Jeppson T, *et al*. Use of electronic health records from a statewide health information exchange to support public health surveillance of diabetes and hypertension. *BMC Public Health* 2019;19:1106.
35 Kruse CS, Stein A, Thomas H, *et al*. The use of electronic health records to support population health: a systematic review of the literature. *J Med Syst* 2018;42:214.
36 Ng K, Steinhubl SR, deFilippi C, *et al*. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circ Cardiovasc Qual Outcomes* 2016;9:649–58.