Commentary

ENVIRONMENTAL EPIDEMIOLOGY

OPEN

Centralizing environmental datasets to support (inter)national chronic disease research

Values, challenges, and recommendations

Jeffrey R. Brook[a,b*], Dany Doiron[c], Eleanor Setton[d], Jeroen Lakerveld[e,f]

Background: Whereas environmental data are increasingly available, it is often not clear how or if datasets are available for health research. Exposure metrics are typically developed for specific research initiatives using disparate exposure assessment methods and no mechanisms are put in place for centralizing, archiving, or distributing environmental datasets. In parallel, potentially vast amounts of environmental data are emerging due to new technologies such as high resolution imagery and machine learning.
Objectives: The Canadian Urban Environmental Health Research Consortium (CANUE) and the Geoscience and Health Cohort Consortium (GECCO) provide a proof of concept that centralizing and disseminating environmental data for health research is valuable and can accelerate discovery. In this essay, we argue that more efficient use of exposure data for environmental epidemiological research over the next decade requires progress in four key areas: metadata and data access portals, linkage with health databases, harmonization of exposure measures and models over large areas, and leveraging "big data" streams for exposure characterization and evaluation of temporal changes.
Discussion: Optimizing the use of existing environmental data and exploiting emerging data streams can provide unprecedented research opportunities in environmental epidemiology through a better characterization of individuals' exposures and the ability to study the intersecting impacts of multiple environmental features or urban attributes across different populations around the world. Proper documentation, linkage, and dissemination of new and emerging exposure data leads to a better awareness of data availability, a reduction of duplication of effort and increases research output.

Keywords: Environmental data; Environmental exposures; Data linkage; Harmonization; Data integration; Big data; Metadata

Introduction

Environmental exposures and urban form are increasingly acknowledged to be important contributors to the development of noncommunicable diseases. Exposure to ambient air pollution has been recently recognized as a leading cause of global disease burden.[1] Environmental attributes such as greenness, walkability, land use, noise, climate, and food environments are other established risk factors for chronic health conditions, as shown in different contexts around the world. Elucidating the relationships between such attributes and health outcomes and how they interact requires disentangling numerous correlated exposures characterized by small relative risks. Major challenges for environmental epidemiology in the coming decade are to channel information on environmental datasets available for research, ensure linkage to health datasets, standardize and sufficiently resolve environmental data across different populations, as well as make use of new data streams to characterize environmental exposures.

Research stakeholders, such as funding agencies in Europe[2] and North America,[3,4] are urging the research community to make a shift toward open science, data sharing, and collaborations as a means to advance scientific innovation and discovery. Regulatory agencies such as the European Commission and the Environmental Protection Agency (EPA) are also pushing for easier access to spatial and environmental data used in regulatory science decision-making.[5,6] In parallel, potentially vast amounts of environmental data are emerging due to new technologies such as high-resolution imagery and machine learning. Such new data streams are offering unprecedented possibilities for environmental epidemiology by generating environmental datasets of high spatial and temporal resolution over larger and larger areas.[7]

To optimize the utility of existing and emerging environmental data for health research, structures and mechanisms should be put in place to ensure that data are findable, accessible, interoperable, and reusable (FAIR).[8] In this essay, we argue that a more efficient use of exposure data for environmental epidemiological research over the next decade requires progress in four key areas: (1) establishing and promoting publicly accessible exposure metadata and data access portals, (2) facilitating and streamlining linkage with health databases, (3) adopting harmonized

[a]Department of Chemical Engineering and Applied Chemistry, Southern Ontario Centre for Atmospheric Aerosol Research, University of Toronto, Toronto, Ontario, Canada; [b]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [c]Respiratory Epidemiology and Clinical Research Unit, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada; [d]Geography Department, University of Victoria, Victoria, British Columbia, Canada; [e]Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, VU University Amsterdam, Amsterdam, the Netherland; and [f]Upstream Team, www.upstreamteam.nl, Amsterdam UMC, VU University Amsterdam, Amsterdam, the Netherlands.

This work was supported by grants from the Canadian Institutes of Health Research (CIHR), the Netherlands Organisation for Scientific Research (NWO), the Netherlands Organisation for Health Research and Development (ZonMw, 91118017), and Amsterdam UMC.

*Corresponding Author. Department of Chemical Engineering and Applied Chemistry, Southern Ontario Centre for Atmospheric Aerosol Research, University of Toronto, 155 College Street, Toronto, Ontario M5T 1P8. E-mail: jeff.brook@utoronto.ca (J.R. Brook).

Environmental Epidemiology (2021) 5:e129

Received: 1 October 2020; Accepted 23 December 2020

Published online 25 January 2021

DOI: 10.1097/EE9.0000000000000129

approaches to measuring and modeling environmental exposures over large areas, and (4) exploiting "big data" streams for exposure characterization and evaluation of temporal changes. A number of past and existing initiatives provide background and can support future developments in these areas.[9–13]

Improving access to (meta)data and linkages with health data

Environmental epidemiologists often develop exposure metrics for specific research initiatives using disparate exposure assessment methods. Once developed, environmental exposure datasets are linked to health datasets which reside with individual researchers, and no mechanisms are typically put in place for centralizing, archiving, or widely sharing them. Whereas environmental data are increasingly available, it is often not clear how or if datasets are available for health research and metadata standards are lacking. The seemingly simple task of locating existing environmental exposure data available for research and understanding them (e.g., variable definitions, measurement methods, geolocation options) is in fact one of the most basic challenges faced by environmental health investigators.

Furthermore, considerable health data residing in the medical community are not applied in environmental epidemiology; either for lack of geographic identifiers at sufficient spatial resolution (e.g., home address or postal code) that are required to enable linkage with spatial environmental data, or due to the inability to send these identifiers to third parties for linkage due to privacy and confidentiality considerations. Given many health databases and cohorts *do* collect geographic identifiers, there are good and useful guidelines for doing so in a secure way within secure data facilities to protect privacy of individuals in administrative health databases or enrolled in observational cohorts. Since the majority of medical/health researchers are not equipped to generate their own state-of-the-art environmental data, efforts are needed to facilitate secure environmental and health data linkages.

Centralizing, documenting, linking, and disseminating environmental exposure datasets requires considerable resources and coordination. Organizations such as the Canadian Urban Environmental Health Research Consortium (CANUE)[14] and the Geoscience and Health Cohort Consortium (GECCO)[15,16] in the Netherlands are helping to fill these needs. Both infrastructures are academically funded and aim to collate and generate spatial measures of environmental exposures and urban form across Canada (CANUE) and the Netherlands (GECCO) in an effort to advance environmental health research. Environmental data housed in the CANUE and GECCO infrastructures are indexed to postal codes or small geographic areas such as those used in national Census, disseminated in simple, analysis ready formats via publicly accessible web portals and linked to health databases for broader distribution. Clear and detailed metadata of the available measures are provided, as well as technical information on procedures, operationalisations, and standards used to develop the data.[14,16] To date, hundreds of research projects have been facilitated by data distributed via CANUE and GECCO. These projects, in turn, have furthered the evidence base and served as entry points for policy makers.[17–23] In their respective countries, CANUE and GECCO are increasingly being recognized as a key source of environmental exposure data and facilitators of health and exposure data linkages through strong partnerships with administrative health data custodians and cohort studies.

Standardizing new data for surveillance and epidemiological analyses

There is growing recognition among health researchers of the advantages of harmonizing and pooling health databases.[24–27] These include increased statistical power to explore rare outcomes, small effects and interactions between risk factors, including gene-environment interactions, minimization of bias due to consistency in confounder adjustment and missing data, a better assessment of the robustness and generalizability of findings, and larger exposure ranges. Normalized difference vegetation index (NDVI), which estimates "greenness" or vegetation exposure from satellite imagery covering the entire planet is good example of a standardized metric used in epidemiological investigations around the world.[28] The field of air pollution epidemiology has also spearheaded the use of standardized exposure data for cross-cohort and multinational collaborations. For example, the European Study of Cohorts for Air Pollution Effects (ESCAPE)[12] and Effects of Low-Level Air Pollution: A Study in Europe (ELAPSE) projects[11] have leveraged standardized approaches to measuring and modelling air pollution concentrations, and linked estimates to cohorts from across Europe to quantify and reduce the uncertainty of air pollutants' health impacts.[29–33] Globally standardized air pollution estimates combined with mortality rates and effect estimates from epidemiological studies have also allowed estimating the global burden of disease associated with air pollution exposure[1,34,35] and has consequently helped drive public and policy awareness of the scale of impact of air pollution on human health. Nonetheless, relatively few environmental or urban exposures have been widely harmonized thus far and challenges remain. For example, few gold standards exist for environmental exposure assessment and the transferability of locally developed models is often limited. The importance of local context might also preclude the development of globally standardized metrics for certain environmental data (e.g., food environments, housing, walkability). Still, challenges to data standardization do not discount the potential benefits of developing and sharing approaches at some level of commonality. A balance might therefore be reached by developing less detailed but more consistent measures with a broader geographic coverage *as well as* more detailed measures that are better adapted to local context but cover smaller areas. For the latter, research continues to be needed to understand the geographic differences between the measures and how they may relate to health. Finally, new initiatives to expand the contents of global environmental datasets and to increase coverage in areas lacking data (e.g., low- and middle-income countries) can be expected to spark innovation in addressing these challenges or at the least better characterize when, where, why, and how context matters, helping environmental epidemiologists interpret findings and exploit geographic differences.

Exploiting new data streams for exposure characterization

New data streams such as high-resolution satellite and street-level imagery combined with machine-learning techniques are providing, for the first time, local data for much of the urbanized world.[36] For example, daily global satellite imagery is now available at 0.5 to 3 m spatial resolutions.[37,38] Street-level imagery is also becoming ubiquitous, via proprietary sources such as Google Street View and openly via crowdsourcing efforts like Open Street Cam. Using these images, computer programs can be trained to identify urban features, which can be turned into geospatial data and used to estimate urban exposures appropriate for environmental health research. Machine-learning techniques and algorithms applied to satellite and street view images have been used to estimate air pollution,[39] greenness,[40] walkability,[41] urban heat island intensity,[42] and to predict spatial distribution of social and environmental health inequities.[43] Ever-increasing coverage and resolution of these new technologies provide opportunities for building locally relevant but globally comparable environmental datasets across large geographical areas and can help bring data of equal quality to regions of the world where resources for environmental monitoring and surveillance infrastructure are limited.

Recommendations

Optimizing the use of existing environmental data and exploiting emerging data streams can provide unprecedented research opportunities in environmental epidemiology through a better characterization of individuals' exposures and the ability to study the intersecting impacts of multiple environmental features or urban attributes across different populations around the world. Key recommendations for a more efficient use of exposure data for environmental epidemiological research over the next decade are provided below.

First, national and international efforts should be directed toward collating and cataloguing existing and emerging datasets of area-level environmental exposures in central, publicly accessible web portals. Use of such web portals should be promoted in the research community and expansion of open data portals beyond national boundaries should be prioritized. Second, controlled vocabularies and compatible metadata standards should be developed and implemented for environmental exposure datasets. The use of compatible metadata standards across data platforms would facilitate multiplatform browsing and eventual data integration. Third, automated processes for indexing of spatial datasets to commonly used linkage fields such as points (e.g., addresses or postal codes) or small area census boundaries should be developed and implemented for existing and future spatial data streams. Fourth, systems and procedures to facilitate routine linkage of exposure files with health databases should be established. This requires substantial collaboration with health data custodians, potentially starting with existing international multicohort consortia, and with particular focus on addressing challenges presented by ethics, consent, and data confidentiality requirements. Fifth, once linked, health data holders should make both health and exposure data available via regular data access channels. Providing access to analysis-ready data will accelerate the research and discovery process. Sixth, and when possible, use of standardized measurement devices and modelling techniques should be prioritized for environmental exposure assessment to improve consistency of variables across studies. This includes exploring the potential of making use of historical exposure data covering large areas (national, continental) to generate compatible exposures. Seventh, international collaborations should be put in place to exploit opportunities offered by new technologies such as imagery and deep learning to scale up environmental exposures, with emphasis on the potential for exposure estimation in areas where lack of resources prevent environmental exposure monitoring and assessment. Finally, buy-in and ongoing support from funding agencies is needed to ensure sustainability and innovation in these areas.

Conclusion

The CANUE and GECCO consortia provide proof of concept that centralizing and widely distributing environmental data for health research is valuable and can accelerate discovery in environmental epidemiology. While considerable investments are required for personnel (i.e., coordination, data scientists, and GIS specialists), data storage, and web development, these infrastructures have shown that proper documentation, linkage and dissemination of new, and emerging exposure data leads to a better awareness of data availability, a reduction of duplication of effort, and increases research output. Leveraging standardized exposures can also lead to larger sample sizes and the possibility of expanding research projects across different populations. We urge groups in other countries to set up open environmental data infrastructures in order to help catalyze novel research and collaborations on the environmental determinants of chronic diseases. The current COVID-19 pandemic has also revealed the relevance of health and environmental data linkages in infectious disease epidemiology.[23,44,45]

Ultimately, the environmental epidemiology and exposure sciences communities should work toward a global open data infrastructure capable of advancing knowledge on the health impacts of environmental exposures and informing policy for healthy city planning and hence, more-broadly, sustainable development. National and international science funding councils should allocate funds to support such initiatives in order to meet current and future data challenges and help advance the field of environmental health research.

Conflict of interest statement

The authors declare that they have no conflicts of interest with regard to the content of this report.

Acknowledgments

The Canadian Urban Environmental Health Research Consortium (CANUE) is financially supported by the Canadian Institutes for Health Research (CIHR). GECCO is financially supported by the Netherlands Organisation for Scientific Research (NWO), the Netherlands Organisation for Health Research and Development (ZonMw, 91118017), and Amsterdam UMC.

References

1. Cohen AJ, Brauer M, Burnett R, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet. 2017;389:1907–1918.
2. European Commission. *Open Science (Open Access)*. 2020. Available at: http://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access. Accessed 2 December 2020.
3. Government of Canada. *Third Biennial Plan to the Open Government Partnership*. 2019. Available at: https://open.canada.ca/en/content/third-biennial-plan-open-government-partnership. Accessed 2 December 2020.
4. National Institutes of Health. *NIH Data Management and Sharing Activities Related to Public Access and Open Science*. 2020. Available at: https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/. Accessed 2 December 2020.
5. Environmental Protection Agency. *Strengthening Transparency in Regulatory Science*. 2018, Federal Register. 18768–18774.
6. European Commission. *Infrastructure for spatial information in Europe: INSPIRE Directive*. 2020. Available at: https://inspire.ec.europa.eu/inspire-directive/2. Accessed 2 December 2020.
7. Jia P. Spatial lifecourse epidemiology. Lancet Planet Health. 2019;3:e57–e59.
8. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
9. Abramic A, Kotsev A, Cetl V, et al. A spatial data infrastructure for environmental noise data in Europe. Int J Environ Res Public Health. 2017;14:726.
10. Kotsev A, Peeters O, Smits P, et al. Building bridges: experiences and lessons learned from the implementation of INSPIRE and e-reporting of air quality data in Europe. Earth Sci Inform. 2015;8:353–365.
11. de Hoogh K, Gulliver J, Donkelaar AV, et al. Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ Res. 2016;151:1–10.
12. Beelen R, Hoek G, Vienneau D, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – the ESCAPE project. Atmos Environ. 2013;72:10–23.
13. McGeehin Michael A, Qualters Judith R, Niskar Amanda S. National environmental public health tracking program: bridging the information gap. Environ Health Perspect. 2004;112:1409–1413.
14. Brook JR, Setton EM, Seed E, Shooshtari M, Doiron D; CANUE—The Canadian Urban Environmental Health Research Consortium. The Canadian Urban Environmental Health Research Consortium—a protocol for building a national environmental exposure data platform for integrated analyses of urban form and health. BMC Public Health. 2018;18:114.

15. Timmermans EJ, Lakerveld J, Beulens JWJ, et al. Cohort profile: the Geoscience and Health Cohort Consortium (GECCO) in the Netherlands. BMJ Open. 2018;8:e021597.
16. Lakerveld J, Wagtendonk A, Vaartjes I, Karssenberg D; GECCO Consortium. Deep phenotyping meets big data: the Geoscience and hEalth Cohort COnsortium (GECCO) data to enable exposome studies in The Netherlands. Int J Health Geogr. 2020;19:49.
17. Li L, Carrino L, Reinhard E, et al. Aircraft noise control policy and mental health: a natural experiment based on the Longitudinal Aging Study Amsterdam (LASA) [published online ahead of print November 4, 2020]. J Epidemiol Commun Health. 2020:jech-2020-214264. doi:10.1136/jech-2020-214264.
18. Generaal E, Hoogendijk EO, Stam M, et al. Neighbourhood characteristics and prevalence and severity of depression: pooled analysis of eight Dutch cohort studies. Br J Psychiatry. 2019;215:468–475.
19. Crouse DL, Pinault L, Balram A, et al. Complex relationships between greenness, air pollution, and mortality in a population-based Canadian cohort. Environ Int. 2019;128:292–300.
20. Pappin AJ, Christidis T, Pinault LL, et al. Examining the shape of the association between low levels of fine particulate matter and mortality across three cycles of the Canadian census health and environment cohort. Environ Health Perspect. 2019;127:107008.
21. To T, Zhu J, Stieb D, et al. Early life exposure to air pollution and incidence of childhood asthma, allergic rhinitis and eczema. Eur Respir J. 2020;55:1900913.
22. Doiron D, Setton EM, Shairsingh K, et al. Healthy built environment: Spatial patterns and relationships of multiple exposures and deprivation in Toronto, Montreal and Vancouver. Environ Int. 2020;143:106003.
23. Stieb DM, Evans GJ, To TM, Brook JR, Burnett RT. An ecological analysis of long-term exposure to PM2.5 and incidence of COVID-19 in Canadian health regions. Environ Res. 2020;191:110052.
24. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. Eur J Epidemiol. 2009;24:727–731.
25. Khoury MJ, Lam TK, Ioannidis JP, et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomarkers Prev. 2013;22:508–516.
26. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: consolidating data harmonization–how to obtain quality and applicability? Am J Epidemiol. 2011;174:261–264.
27. Fortier I, Doiron D, Little J, et al; International Harmonization Initiative. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol. 2011;40:1314–1328.
28. James P, Banay RF, Hart JE, Laden F. A review of the health benefits of greenness. Curr Epidemiol Rep. 2015;2:131–142.
29. Cai Y, Zijlema WL, Doiron D, et al. Ambient air pollution, traffic noise and adult asthma prevalence: a BioSHaRE approach. Eur Respir J. 2017;49:1502127.
30. Doiron D, de Hoogh K, Probst-Hensch N, et al. Residential air pollution and associations with wheeze and shortness of breath in adults: a combined analysis of cross-sectional data from two large European cohorts. Environ Health Perspect. 2017;125:097025.
31. Hvidtfeldt UA, Severi G, Andersen ZJ, et al. Long-term low-level ambient air pollution exposure and risk of lung cancer - a pooled analysis of 7 European cohorts. Environ Int. 2020;146:106249.
32. Cai Y, Hansell AL, Blangiardo M, et al; BioSHaRE. Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts. Eur Heart J. 2017;38:2290–2296.
33. Beelen R, Raaschou-Nielsen O, Stafoggia M, et al. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. Lancet. 2014;383:785–795.
34. Brauer M, Amann M, Burnett RT, et al. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. Environ Sci Technol. 2012;46:652–660.
35. van Donkelaar A, Martin RV, Brauer M, Boys BL. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environ Health Perspect. 2015;123: 135–143.
36. Weichenthal S, Hatzopoulou M, Brauer M. A picture tells a thousand…exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environ Int. 2019;122:3–10.
37. Maxar Technologies. *Maxar*. 2020. Available at: https://www.maxar.com/. Accessed 2 December 2020.
38. Planet Labs Inc. *Planet*. 2020. Available at: https://www.planet.com/. Accessed 2 December 2020.
39. Apte JS, Messier KP, Gani S, et al. High-resolution air pollution mapping with Google street view cars: exploiting big data. Environ Sci Technol. 2017;51:6999–7008.
40. Li X, Zhang C, Li W, et al. Assessing street-level urban greenery using Google Street View and a modified green view index. Urban Forest Urban Green. 2015;14:675–685.
41. Yin L, Wang Z. Measuring visual enclosure for street walkability: using machine learning algorithms and Google Street View imagery. Appl Geogr. 2016;76:147–153.
42. Chakraborty T, Lee X. A simplified urban-extent algorithm to characterize surface urban heat islands on a global scale and examine vegetation control on their spatiotemporal variability. Int J Appl Earth Obs Geoinf. 2019;74:269–280.
43. Suel E, Polak JW, Bennett JE, Ezzati M. Measuring social, environmental and health inequalities using deep learning and street imagery. Sci Rep. 2019;9:6229.
44. Wu X, Braun D, Schwartz J, et al. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. Sci Adv. 2020;6:eaba5692.
45. Liang D, Shi L, Zhao J, et al. Urban air pollution May enhance COVID-19 case-fatality and mortality rates in the United States. Innovation (N Y). 2020;1:100047.