Cancer Causes & Control (2019) 30:671–675
https://doi.org/10.1007/s10552-019-01182-2

COMMENTARY

Artificial intelligence (AI) and cancer prevention: the potential application of AI in cancer control programming needs to be explored in population laboratories such as COMPASS

Scott T. Leatherdale[1] · **Joon Lee**[1,2]

Received: 4 September 2018 / Accepted: 10 May 2019 / Published online: 15 May 2019

Abstract
Understanding the risk factors that initiate cancer is essential for reducing the future cancer burden. Much of our current cancer control insight is from cohort studies and newer large-scale population laboratories designed to advance the science around precision oncology. Despite their promise for improving diagnosis and treatment outcomes, their current reductionist focus will likely have little impact shifting the cancer burden. However, it is possible that these big data assets can be adapted to have more impact on the future cancer burden through more focus on primary prevention efforts that incorporate artificial intelligence (AI) and machine learning (ML). ML automatically learns patterns and can devise complex models and algorithms that lend themselves to prediction in big data, revealing new unexpected relationships and pathways in a reliable and replicable fashion that otherwise would remain hidden given the complexities of big data. While AI has made big strides in several domains, the potential application in cancer prevention is lacking. As such, this commentary suggests that it may be time to consider the potential of AI within our existing cancer control population laboratories, and provides justification for why some small targeted investments to explore their impact on modelling existing real-time cancer prevention data may be a strategic cancer control opportunity.

Keywords Artificial intelligence · Machine learning · Prevention · Population laboratory · Precision oncology

The number of new cancer cases in Canada is expected to rise 40% in the next 15 years (~ 277,000 new case per year) [1]. Understanding the risk factors that initiate or accelerate cancer (or other chronic diseases that share common risk factors) is essential for reducing the future cancer burden. Current cancer control insight in Canada is being garnered from clinical trails, such as the Canadian Cancer Clinical Trials Network [2] or cohort studies such as the Canadian Study of Diet, Lifestyle and Health [3], and population-based surveillance systems such as the Canadian Community Health Survey (CCHS) or the Canadian Tobacco Use Monitoring Survey (CTUMS). As our knowledge base and technology have advanced in recent decades, new larger-scale and sophisticated cohort studies have been established that act as population laboratories, such as the Canadian Partnership for Tomorrow Project (CPTP) [4] or the COMPASS study (COMPASS) [5, 6]. While these Canadian-based population laboratories are becoming gold standards for chronic disease and cancer control research nationally, they add unique value to different pillars of the cancer control spectrum.

For instance, the CPTP addresses the pillars of etiology, detection, and treatment. Through the incorporation of biological specimens linked to population surveillance data on a variety of behavioural and environmental risk factors among adults, the CPTP has stimulated much interest around the opportunity for advancing the science and practice around personalized medicine (more specifically *precision oncology*) [4]. The CPTP is poised to have a profound impact on understanding disease risk and onset, and improving individual-based cancer treatment outcomes in the future. Despite the potential promise that precision oncology derived from the CPTP will have on diagnosis and treatment outcomes, the reductionist focus does little to shift the population

✉ Scott T. Leatherdale
sleather@uwaterloo.ca

[1] School of Public Health and Health Systems, University of Waterloo, 200 University Avenue, Waterloo, ON N2L 3G1, Canada

[2] Departments of Community Health Sciences & Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, Canada

prevalence of the large-effect modifiable risk factors that actually drive meaningful changes in the cancer burden (i.e., the prevention of tobacco use, obesity, poor diet, alcohol use, and physical inactivity that generate the most disease [7]). Conversely, COMPASS addresses the pillar of primary prevention. Recently COMPASS has stimulated much interest pertaining to the opportunity for advancing the science and practice around natural experiment learning systems in cancer prevention [8]. This is accomplished through the longitudinal measurement of cancer-relevant behavioural and risk factor data linked to ongoing changes in cancer control-related program and policy data locally, provincially and nationally among youth populations [5, 6]. The linked longitudinal data can then be used to evaluate the impact the ongoing program or policy changes have on the risk behaviours of youth in the cohort. COMPASS is also uniquely positioned to impact the cancer burden by working directly with local, provincial and federal stakeholders to share and use timely COMPASS evidence to guide their primary prevention program and policy efforts that are focused on shifting large-effect risk factors (*precision prevention*). Through the future incorporation of biological specimens in subsamples of youth participating in COMPASS, it could also quickly be transformed to also act as a precision oncology platform. These data could also be linked to other available big data resources (air pollution data, wearable technology data) increasing the value but also the complexity of the data. Similarly, the incorporation of relevant cancer control program and policy data (resources available in the Canadian Partnership Against Cancer's Prevention Policies Directory [9]) layered onto the CPTP [as was originally intended in the Ontario arm of this initiative] [10], would allow the CPTP be transformed to also act as a precision prevention platform. Regardless, Canada is in the advantageous position of having two of the largest and most robust cancer control research platforms located 'in house'; now we just need to determine how to quickly and effectively adapt and use these big data resources to have timely and meaningful real-world impact on the cancer burden. This is where artificial intelligence (AI), may provide some benefit to cancer control researchers by exploring a multitude of predictive models to learn how interventions, or unique combinations of interventions shape the way individuals or sub-groups of individuals behave.

AI, a form of intelligence demonstrated by machines, is an enormously popular research topic today in both academia and industry. The current surge in the interest in AI has been fuelled by the recent advances in big data analytics, mobile technology, natural language processing, and the 'Internet of Things'. Within the domain of AI, machine learning (ML) has played a significant role particularly in recent years. ML uses statistical and computational techniques to give computers the ability to progressively improve performance on a specific task without being explicitly programmed. So although ML is just one approach in AI, most successful AI techniques today are ML based, particularly deep learning. The reason for this has much to do with the emergence of big data and advances in how big data can be processed and analyzed; ML methods require large quantities of data in order to be successful. The primary strength of ML lies in that it automatically learns the patterns in a data set of past examples for a specific task (image recognition, language translation); it turns out that simply giving a machine a list of rules and heuristics to follow is often not robust enough for many complex real-world applications (all the complexities associated with the big data available in precision oncology and precision prevention platforms). ML can be used to devise complex models and algorithms that lend themselves to prediction (predictive analytics), where consistent results can be identified in big data. This can be especially exciting when new unexpected relationships and pathways are discovered in a reliable and replicable fashion that otherwise would remain hidden given the complexities of the data. However, one concern is that if there are systematic biases in the data collected, the results may just be a reliable replication of that bias. Because ML learns patterns present in data, it is true that any biases embedded in data can be translated into algorithmic biases in ML models. This is a broader data quality/collection issue that is not specific to ML. Other quantitative fields such as epidemiology or biostatistics are also limited by data quality, and any biases stemming from the unique characteristics of the target population or data collection protocol can lead to biased results. Hence, emphasis should be placed on careful design of data collection and evaluation of degree of bias and data quality in general. For instance, one way this is addressed in COMPASS is by collecting data from students from a broad swath of geographic locations and socioeconomic contexts so that student demographics vary widely. Ideally, one should not blindly trust ML model output without rigorous, comprehensive validation for a wide variety of cases. Additional methodological challenges would exist when incorporating AI or ML methods to data systems such as CPTP or COMPASS, including reporting standards, missing data, and the need for validation data sets, and these methodological considerations are described in more detail elsewhere [11, 12].

While AI has made big strides in several non-health domains (self-driving cars, digital assistants such as Siri and Alexa, AlphaGo), it has been applied in a more limited way in health, healthcare, or health policy more generally. While recent publications discussed the importance of ML [13] and big data [14–16] in medicine, relatively fewer articles have focused on how data technologies can improve public health [17], and even those few articles tend to be limited to population health surveillance [18, 19]. The connection between AI and cancer control programming has been almost

non-existent, except for some recent evidence where ML was used for colorectal cancer screening [20, 21] or monitoring media coverage of electronic nicotine delivery systems (e-cigarettes) [22]. The use of ML to evaluate the impact of primary prevention cancer control efforts is currently lacking but full of potential. In Canada, as elsewhere, there are literally hundreds of prevention-related interventions (cannabis legalization, excise tobacco taxes, marketing of vaping products, behavioural modification programs) implemented in an ongoing basis across a variety of different domains (economics, education, agriculture, clinical practice) that are not evaluated. This represents numerous lost opportunities for generating timely practice-based evidence [8] to inform future cancer control action. While some authors suggest this may in part be due to researchers having limited access to high-quality population data sets [23], this is not the case with the Canadian-based CPTP and COMPASS population laboratories that make their data widely available to researchers via online data request portals (CPTP https://portal.partnershipfortomorrow.ca/; COMPASS https://uwaterloo.ca/compass-system/information-researchers/data-usage-application). Moreover, if these data systems were linked to other available data resources where feasible (linking data from cohort participants to other data systems collected via apps such as Fitbit, My Fitness Pal or MyPlate Calorie Tracker), the potential for adding to the evidence base can increase dramatically. As such, the limitation may in fact be financial resources or the capacity to perform the necessary analytical models.

In Canada, there are currently limited resources to support cancer control research despite the burden the disease imposes on the population. In 2015, $480 million was invested in cancer research [24]; 4% on prevention (~$19 million), 38% on small-risk research paradigms (including personalized medicine ~$182 million) and 30% spent on treatment (~$144 million)]. It is clear that we are underspending on cancer control, especially in our efforts associated with prevention. Moreover, the Canadian Cancer Research Alliance does not even report on our cancer control investments in research on systems or models anymore [24]; the investments remain negligible. As such, the timing to test the potential application of AI within our existing big data focused on cancer control may be ideal, especially as Canada is positioning itself to be a world leader in AI and investing resources in a corresponding fashion. For example, several key world leaders in AI research are located in Canada, most notably, Drs. Geoffrey Hinton (University of Toronto & Google) and Yoshua Bengio (Université de Montréal) who have pioneered deep learning which currently is the most successful ML technique. Moreover, in 2016 Canada invested over $1.3 billion to create three world-class AI institutes: Vector Institute in Toronto (led by Dr. Hinton), Montreal Institute for Learning Algorithms (MILA, led by Dr. Bengio), and the Alberta Machine Intelligence Institute (Amii) in Edmonton [25]. Considering large nation-wide influxes in cancer prevention funding are unlikely in the near future, even some small targeted investments to explore the impact AI and ML can have on modelling existing real-time cancer prevention data may be a strategic opportunity. For instance, the Canadian Institutes of Health Research (CIHR) recently announced the Equitable AI Initiative designed to build the capacity necessary in the population and public health research community equip the next generation of researchers with the necessary tools to incorporate AI and Ml into public health efforts [26]. It may be advantageous for CIHR to work with CPTP and COMPASS leadership to develop these systems as the validation data sets for the emerging Equitable AI training platform to use moving forward.

For instance, COMPASS provides a robust but complex source of data focused on cancer prevention. In brief, COMPASS is a prospective cohort (2012–2021) collecting hierarchical longitudinal data from 100,000's of students in grades 9–12 (ages 13–18) and the schools they attend in Ontario, British Columbia, Alberta, Quebec, and Nunavut [6]. Each year, COMPASS measures numerous cancer-related behavioural and outcome data from students (tobacco use, e-cigarettes, physical activity, diet, alcohol use, cannabis use, depression, anxiety) and then links those student-level data to measures of real-world cancer control related programs, policies and built environment resources occurring within the schools, communities, provinces/territory, and even nationally that those students are situated in. Every year, hundreds of different real-world program, policy, and/or built environment changes (*interventions*) take place in the various contextual environments surrounding students in the cohort that can be evaluated as natural experiments using the quasi-experimental design embedded in the system [8]. These COMPASS data can be used to explore and mine the literally thousands of different permutations of simple to complex hierarchical program-to-policy-to-environment interactions that impact youth, or sub-populations of youth, to ultimately understand what interventions (or mix of interventions) work for whom and under what conditions. For example, with the recent legalization of cannabis in Canada in 2018, it would be insightful to use the ML approach of decision trees to explore how the legalization on cannabis impacts patterns of cannabis use among students and sub-populations of students, while also exploring how different school cannabis policies or the location of cannabis dispensaries impact trajectories of use. While it is possible for researchers to develop such models to evaluate the impact that these different interventions or mixes of interventions may have on youth, we also know this would be incredibly resource intensive and slow, and many of the potential combinations would never be explored. Using AI methods

to rapidly and robustly examine these interactions in real-time may dramatically speed up the process of building our understanding in a domain of science where currently very little is known about what works or doesn't work. Even if AI were only used to mine COMPASS data using researcher-driven heuristics to discover opportunities where particular mixes of interventions are most effective, or where the particular sequential order of intervention implementation is most effective, may help to identify promising practices for researchers to explore and examine with more rigour and critical insight moving forward.

Research has recently demonstrated that ML methods when applied to longitudinal health data can provide accurate and more parsimonious results when compared to a traditional statistical modelling approach [27]. Applying the conceptual thinking behind precision oncology (providing the right treatment to the right person at the right time) to the domain of precision prevention (targeting the right program or policy to the right population or context at the right time) has tremendous potential if AI could be correctly applied to complex sources of data such as those available in COMPASS. Clearly, similar benefits may be possible by testing the potential utility of AI for advancing precision oncology within the CPTP as well, but given the dearth of resources currently invested in primary prevention in Canada, the available COMPASS data may provide a strategic opportunity for testing the potential of ML in cancer control in Canada.

For some time now it has been clear that cancer control could benefit from more action-oriented research that has timely real-world impact on the large-effect modifiable risk factors for the cancer burden [7]. However, our current research mechanisms, especially those focused on prevention, are too slow. Some evidence suggests there is roughly a 17-year gap between the time when research funding priorities are set and there being sufficient research evidence available to change prevention practice [28]. Given that only 4% of cancer control investment in Canada is spent on prevention research [24], in most relevant domains of interest there is no evidence available despite the immediate evidence needs often identified by cancer control decision makers. It now seems feasible in Canada, given the investments in AI in domains outside of cancer control, to test if the principles of AI can successfully be incorporated into existing cancer prevention data systems such as COMPASS. This seems like the sort of 'low risk high reward' opportunity that presently should be explored with the cancer control relevant prevention data Canada is fortunate to be privy to.

Acknowledgments Dr. Leatherdale is a Chair in Applied Public Health Research funded by the Public Health Agency of Canada (PHAC) in partnership with the Canadian Institutes of Health Research (CIHR).

References

1. Canadian Cancer Society's Advisory Committee on Cancer Statistics (2015) Canadian cancer statistics 2015. Canadian Cancer Society, Toronto
2. Canadian Cancer Clinical Trials Network. https://3ctn.ca/. Accessed 10 Mar 2019
3. Rohan TE, Soskolne CL, Carroll KK, Kreiger N (2007) The Canadian study of diet, lifestyle, and health: design and characteristics of a new cohort study of cancer risk. Cancer Detect Prev 31:12–17
4. Dummer TJB, Awadalla P, Boileau C, Craig C, Fortier I, Goel V, Hicks JMT, Jacquemont S, Knoppers BM, Le N, McDonald T, McLaughlin J, Mes-Masson AM, Nuyt AM, Palmer LJ, Parker L, Purdue M, Robson PJ, Spinelli JJ, Thompson D, Vena J, Zawati M, with the CPTP Regional Cohort Consortium (2018) The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. CMAJ. https://doi.org/10.1503/cmaj.170292
5. Leatherdale ST, Brown KS, Carson V, Childs RA, Dubin JA, Elliott SJ, Faulkner G, Hammond D, Manske S, Sabiston CM, Laxer RE, Bredin C, Thompson-Haile A (2014) The COMPASS study: a longitudinal hierarchical research platform for evaluating natural experiments related to changes in school-level programs, policies and built environment resources. BMC Public Health. https://doi.org/10.1186/1471-2458-14-331
6. The COMPASS Study (2018) https://uwaterloo.ca/compass-system/compass-system-projects/compass-study. Accessed 16 Aug 2018
7. Colditz GA, Wolin KY, Gehlert S (2012) Applying what we know to accelerate cancer prevention. Sci Transl Med. https://doi.org/10.1126/scitranslmed.3003218
8. Leatherdale ST (2018) Natural experiment methodology for research: a review of how different methods can support real-world research. Int J Soc Res Methodol. https://doi.org/10.1080/13645579.2018.1488449
9. Canadian Partnership Against Cancer Prevention Policies Directory. https://www.partnershipagainstcancer.ca/tools/prevention-policies-directory/. Accessed 15 Mar 2019
10. Ashbury F, Kirsh V, Kreiger N, Leatherdale ST, McLaughlin J (2006) An invitation to develop Ontario's cancer research platform: report of the "Ontario Cancer Cohort Workshop". Chronic Disease Can 27:94–97
11. Dinov ID (2016) Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. GigaScience 5:12. https://doi.org/10.1186/s13742-016-0117-6
12. Bengio Y, Lodi A, Prouvost A (2018) Machine learning for combinatorial optimization: a methodological tour dhorizon. Data science for real-time decision making. https://arxiv.org/pdf/1811.06128.pdf. Accessed 15 Mar 2019
13. Darcy AM, Louie AK, Roberts LW (2016) Machine learning and the profession of medicine. JAMA 315:551–552
14. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ (2015) Big data for health. IEEE J Biomed Health Inform 19:1193–1208
15. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 33:1123–1131
16. Schneeweiss S (2014) Learning from Big Health Care Data. N Engl J Med 370:2161–2163
17. Khoury MJ, Ioannidis JPA (2014) Big data meets public health. Science 346:1054–1055
18. Nguyen QC, Li D, Meng HW, Kath S, Nsoesie E, Li F, Wen M (2016) Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. JMIR Public Health Surveill. https://doi.org/10.2196/publichealth.5869

19. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature. https://doi.org/10.1038/nature07634
20. Ling R, Lee J (2016) Disease monitoring and health campaign evaluation using Google search activities for HIV and AIDS, stroke, colorectal cancer, and marijuana use in Canada: a retrospective observational study. JMIR Public Health Surveill. https://doi.org/10.2196/publichealth.6504
21. Goshen R, Choman E, Ran A, Muller E, Kariv R, Chodick G, Ash N, Narod S, Shalev V (2018) Computer-assisted flagging of individuals at high risk of colorectal cancer in a large Health Maintenance Organization using the ColonFlag test. JCO Clin Cancer Inform 2:1–8
22. Ayers JW, Dredze M, Leas EC, Caputi TL, Allem JP, Cohen JE (2018) Next generation media monitoring: Global coverage of electronic nicotine delivery systems (electronic cigarettes) on Bing, Google and Twitter, 2013-2018. PLoS ONE. https://doi.org/10.1371/journal.pone.0205822
23. Ashrafian H, Darzi A (2018) Transforming health policy through machine learning. PLoS Med. https://doi.org/10.1371/journal.pmed.1002692
24. Cancer Research Alliance (CCRA) (2017) Cancer Research Investment in Canada, 2015. https://www.ccra-acrc.ca/index.php/publications-en/investment-reports-annual. Accessed 18 Aug 2018
25. Invest Canada (2018) http://www.international.gc.ca/investors-investisseurs/assets/pdfs/download/Niche_Sector-AI.pdf. Accessed 18 Aug 2018
26. Canadian Institutes of Health Research. Equitable AI activating opportunities. http://www.cihr-irsc.gc.ca/e/51204.html. Accessed 15 Mar 2019
27. Engchuan W, Dimopoulos AC, Tyrovolas S, Caballero FF, Sanchez-Niubo A, Arndt H, Ayuso-Mateos JL, Haro JM, Chatterji S, Panagiotakos DB (2019) Sociodemographic indicators of health status using a machine learning approach and data from the English Longitudinal Study of Aging (ELSA). Med Sci Monit. https://doi.org/10.12659/MSM.913283
28. Green LW, Ottoson JM, Garcia C, Hiatt RA (2009) Diffusion theory and knowledge dissemination, utilization, and integration in public health. Ann Rev Public Health 30:151–174