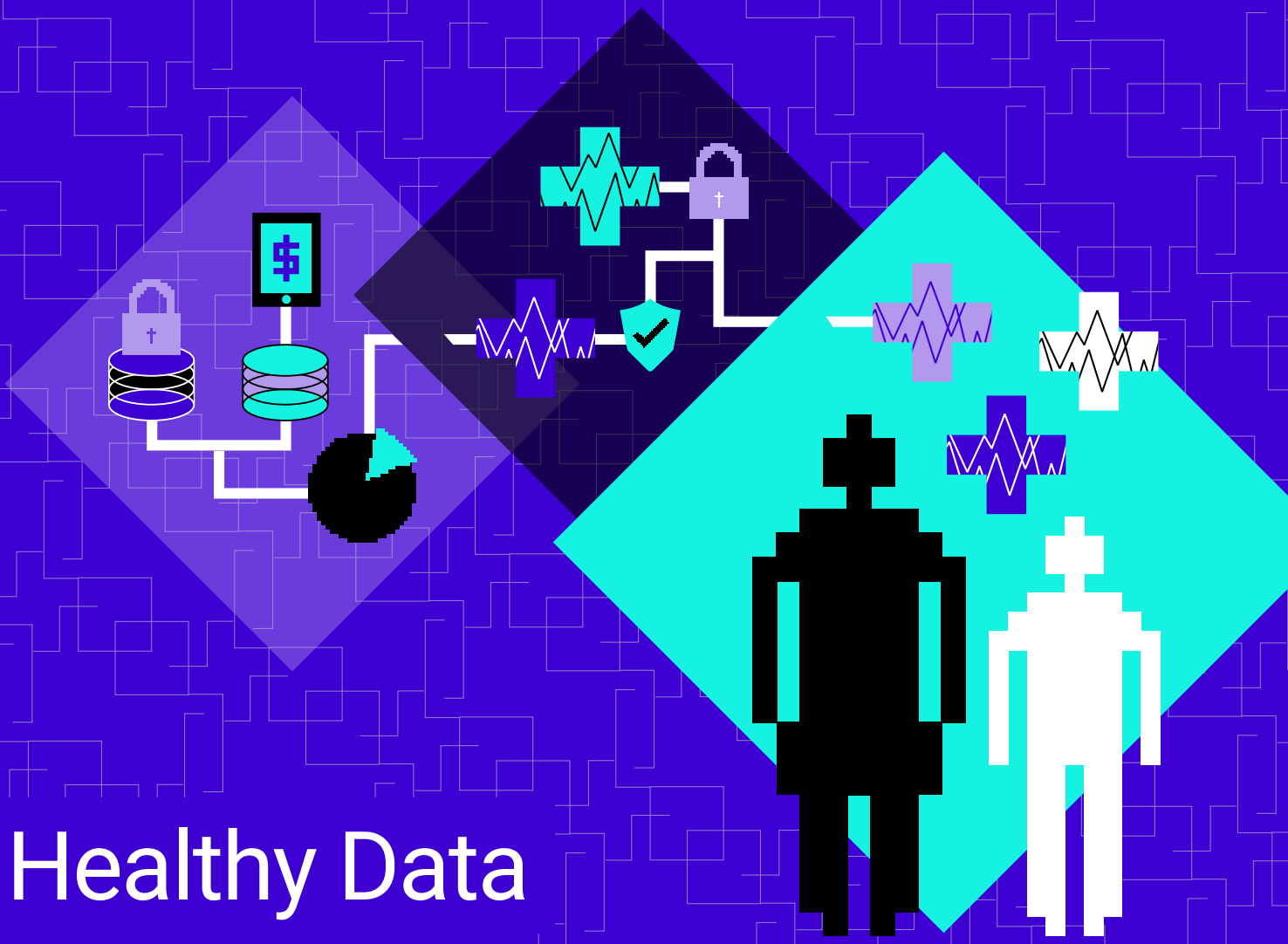
MOWAT RESEARCH #179 | DECEMBER 2018



Healthy Data

Policy solutions for big data and AI innovation in health

BY CATHERINE STINSON

Mowat Centre
ONTARIO'S VOICE ON PUBLIC POLICY

MUNK SCHOOL
OF GLOBAL AFFAIRS
& PUBLIC POLICY

UNIVERSITY OF TORONTO

Acknowledgements

The author would like to thank those who participated in this research project as key informant interviewees and critical reviewers. Thanks also to Andrew Parkin, Michael Crawford Urban, Reuven Shlozberg and Sunil Johal for their review and Elaine Stam for her design work on this report. Special thanks to Rebecca Hallam and Andrew Thies for their help with the interviews and background research.

Author

CATHERINE STINSON

Senior Policy Associate

Catherine joined the Mowat Centre as a Senior Policy Associate in May 2018, bringing experience from the intersection of technology and health. They hold a Masters degree in Computer Science from the University of Toronto and a PhD in History & Philosophy of Science from the University of Pittsburgh. Catherine has worked as a machine learning researcher, web developer, resident philosopher in a virtual reality lab, and lecturer at universities in Canada, the USA, and Germany. Their Postdoc research at the Rotman Institute of Philosophy focused on how to reform psychiatric classification and treatment to reflect the broad scope of influences on mental health, from genetics up to social factors. Catherine has contributed to public debate about ethics in AI and health in *The Globe and Mail* and *CBC Radio*, and has an article about consent forthcoming in *Chatelaine*.

Mowat Centre
ONTARIO'S VOICE ON PUBLIC POLICY

The Mowat Centre is an independent public policy think tank located at the Munk School of Global Affairs and Public Policy at the University of Toronto. The Mowat Centre is Ontario's non-partisan, evidence-based voice on public policy. It undertakes collaborative applied policy research, proposes innovative research-driven recommendations, and engages in public dialogue on Canada's most important national issues.

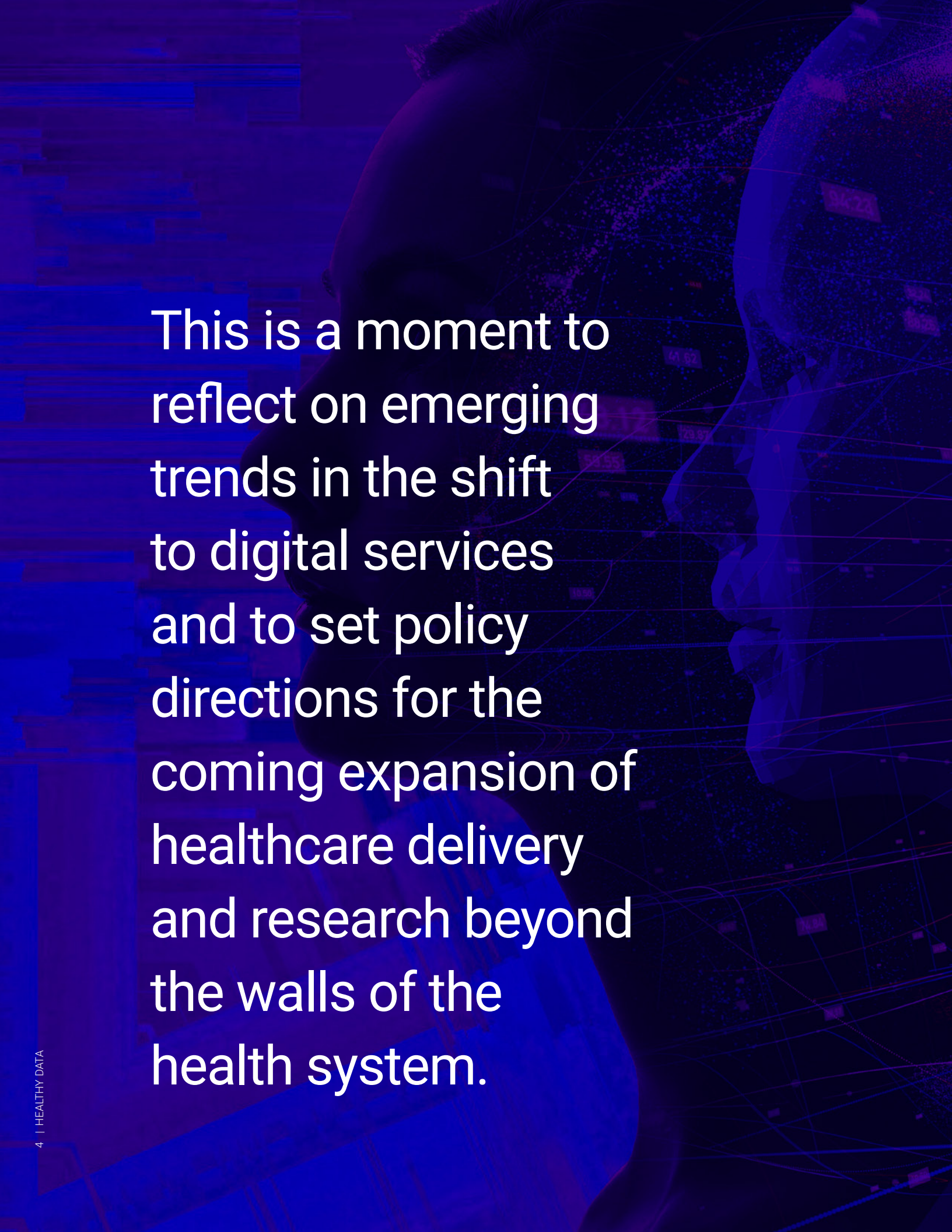
MOWATCENTRE.CA
@MOWATCENTRE
439 UNIVERSITY AVENUE
SUITE 2200, TORONTO, ON
M5G 1Y8 CANADA

MUNK SCHOOL
OF GLOBAL AFFAIRS
& PUBLIC POLICY

UNIVERSITY OF TORONTO

 ISBN 978-1-77259-080-7

Contents



This is a moment to reflect on emerging trends in the shift to digital services and to set policy directions for the coming expansion of healthcare delivery and research beyond the walls of the health system.

EXECUTIVE SUMMARY

The health sector is undergoing a digital revolution. Electronic Health Records (EHRs) are being rolled out across the country. Hospitals and medical researchers are sharing data across silos and borders. The tech industry is producing health apps and wearable wellness gadgets. Artificial intelligence (AI) researchers and doctors are forming partnerships to collect vast banks of health data for analysis. And patients are increasingly seeking help with their health problems online.

These changes promise to transform healthcare by enabling the discovery of new treatments, streamlining diagnoses, making care pathways more efficient, cutting costs, giving patients more control over their care and ultimately improving quality of life. But there are also risks involved in greater data sharing. Without oversight, we could end up with a system where chronic treatments are prioritized over cures, drug prices are wildly inflated, patients lose control over their personal information and health inequities increase.[1][2]

We have reached a point where some digital services like EHRs are available to the vast majority of Canadians and data sharing for research purposes is taking off. Technology companies and AI researchers are actively vying for greater access to health data, while policymakers are looking for data governance models that will strike a balance between encouraging innovation and protecting patients' rights. This is a moment to reflect on emerging trends in the shift to digital services and to set policy directions for the coming expansion of healthcare delivery and research beyond the walls of the health system.

This report surveys the changes afoot, weighs projected benefits against potential harms, integrates stakeholder opinion, maps out the policy challenges and proposes actionable recommendations. Although the focus is on the Province of Ontario, the insights are drawn from a wider set of jurisdictions, and many of the action points will be applicable for any jurisdictions seeking to develop policy for AI in health care.

The current hype about AI combined with social media's commercialization of personal data, and the technology industry's moves into surveillance capitalism, have the public and policymakers concerned.

1 Ross, Ángel (2017). *Powering Health Equity Action with Online Data Tools: 10 Design Principles. Policy Link & Ecotrust.* http://nationalequityatlas.org/sites/default/files/10-Design-Principles-For-Online-Data-Tools.pdf.

2 Zhang, X., Pérez-Stable, E. J., Bourne, P. E., Peprah, E., Duru, O. K., Breen, N., Berrigan, D., Wood, F., Jackson, J. S., Wong, D., Denny, J. (2017). Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century. *Ethnicity & disease, 27*(2), 95-106. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5398183/.

There is an inherent tension between the desire for innovation-friendly data platforms and the need to protect individuals and communities from the harm or unfair treatment that could result.

1 INTRODUCTION

People are scrambling to find out what they need to know about AI in order to protect their interests. This anxiety is heightened in health, given the sensitivity of health data and importance of health to quality of life. But the reality is that AI and big data have the potential to revolutionize health for the public good if the digital transition is governed skilfully.

There is good agreement that to kick-start innovation in health we need more data, of better quality, held in more accessible formats, with integrated tools for performing sophisticated analyses. But there is an inherent tension between the desire for innovation-friendly data platforms and the need to protect individuals and communities from the harm or unfair treatment that could result.

Focus of Report

The focus of this report is on how to strike the right balance between innovation and patients' rights. The main themes covered are privacy, consent, trust, and data governance.

Privacy

In most cases, health data is anonymized before it is used for *secondary purposes*, i.e., purposes that go beyond the delivery and administration of healthcare. Given how much personal information is now shared and stored online, maintaining privacy of health records is becoming more difficult, even after records have been stripped of personally identifying information. Sophisticated methods of de-identifying data have been developed, but they do not eliminate all privacy risks. Successfully anonymized records are also less useful to analysts, because individuals can no longer be tracked across datasets. Anonymization is an important tool for the protection of patients' rights, but it has limits and is not the only tool needed.

Consent

Since the introduction of the Nuremberg Code in 1947,[3] informed consent has been an important tool for protecting patients' rights in medical research. But with more widespread data sharing and larger-scale studies, seeking consent for each use becomes onerous to providers and overwhelming to patients. With advanced analytics, it is impossible to predict the purposes to which data may be put, making it difficult to keep patients informed. Even where uses are known, some are so complicated that the average patient does not have the scientific or technical literacy to understand the benefits and risks. Vulnerable populations require even more careful consent processes. We may have reached a point where the consent model is breaking down.

3 National Institute of Health Office of History and Stretten Museum (no date). *The Nuremberg Code*. Bethesda, MD: National Institute of Health. https://history.nih.gov/research/downloads/nuremberg.pdf.

Trust

Patients willingly participate in research when they trust data stewards to protect their interests. Trust requires transparency and accountability, but existing structures have important gaps in their coverage. Hospitals are generally careful with health data, but health data collected by industry is not subject to the same privacy laws. Research Ethics Board (REB) interpretations of the Tri-Council[4] Policy Statement[5] (which sets out expectations for ethical and methodological standards in research funded by the Tri-Council, including decisions about when research with human participants is permissible) vary widely, education programs for tech workers typically include no training on how to manage sensitive health data and healthcare providers are not always given the resources they need to protect patients' rights. These gaps must be addressed if trust is to be maintained.

Education programs for tech workers typically include no training on how to manage sensitive health data

Data Governance

Building platforms for innovation involves significant work in data collection, storage, sharing, analysis and governance. Many parties are involved in one way or another, ranging from patients, to advocacy groups, healthcare providers, technology companies, academic researchers, governments and independent bodies, each of which have a stake in how the digital transformation plays out. Patients are not in a position to manage the process themselves, while healthcare providers are busy caring for patients. There is space for technology companies to take responsibility for some of this work, if policies are in place to ensure that they act in the public interest. However, some roles might best be played by independent bodies insulated from both private interests and shifting political winds.

These four themes are intertwined with practical concerns like building an interoperable EHR system that meets all needs, compelling providers to use it, implementing access controls and ensuring compliance with existing policy. Throughout this report, practical solutions that work for busy healthcare providers working with limited budgets are favoured over unrealistic "ideal" solutions.

4 The Tri-council is made up of the Social Sciences and Humanities Research Council, the Natural Sciences and Engineering Research Council, and the Canadian Institute of Health Research.

5 Canadian Institute of Health Research; Natural Sciences and Engineering Research Council of Canada; Social Sciences and Humanities Research Council of Canada (2005). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. http://www.pre.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf.

Report Structure and Methods

The rest of this report is organized as follows. Section 2 covers the projected benefits of AI and big data innovation in healthcare, the potential harms and the strategies currently in use. Since these innovations depend critically on health information being digitized and accessible to researchers, Section 2 begins by considering EHRs. The remainder of the section covers the three arenas where innovation is happening: the technology industry, public sector medical research and independent data governance and research centres.

Section 3 draws policy lessons from the challenges and strategies discussed. Based on insights from the literature, the views of key informants, and existing policies that govern this area, this section draws out best practices for managing consent, privacy, trust and data governance. It further notes the practical barriers healthcare providers and researchers face in implementing those best practices, and the policy gaps that allow for potentially harmful uses of data. A series of policy responses are proposed that would support ethical, responsible innovation in health care by removing those barriers and closing the gaps.

The Conclusion highlights the key themes that emerge from the preceding sections' analysis and summarizes the action points developed throughout the paper.

Research for this report is based on two strands of inquiry:

» Semi-structured interviews with 24 stakeholders working across nine jurisdictions, primarily in Ontario.[6] Our sample included primary care physicians, medical researchers, healthcare administrators, public health workers, privacy officers, health informatics researchers, technology developers, Ontario and Federal public servants, ethicists, patient advocates, open data specialists and legal scholars.[7]

» A broad literature review covering AI projects underway in Ontario hospitals, health data privacy, AI ethics and data governance.

6 These jurisdictions are Cambridge (UK), Guelph (ON), Hamilton (ON), London (ON), Montreal (QC), Northern BC, Ottawa (ON), Toronto (ON), and Vancouver (BC).
7 The direct quotations from interviews are referenced in footnotes, but the anonymity of the individuals concerned is preserved. This paper also benefitted from informal conversations with physicians, informatics workers, a legal scholar, and a First Nations health administrator.

2 BIG DATA AND AI'S INFLUENCE ON HEALTHCARE

This section examines three arenas where big data and AI intersect with healthcare, and the EHRs that act as the basic building blocks for innovation. It outlines the improved health outcomes that this transformation promises, the risks that stakeholders are concerned about, and the strategies in use to manage risk.

Electronic Health Records

There is some disagreement about the distinction between EHRs and electronic medical records (EMRs). Here we will use EMR to indicate the records stored in hospital or doctor's office computer systems. The encoding and other particulars of EMRs vary considerably depending on the software used, and may not be readable by the EMR systems used by other providers. We will use EHR to indicate medical records more generally, in contexts where the records from multiple sources and formats might be combined in larger collections in an interoperable format suitable for broader access. EMRs are being rolled out across Canada. Although adoption has been slow compared to other countries,[8] as of 2017, 85 per cent of primary care physicians were using EMRs.[9] Efforts to build a pan-Canadian EHR system that would establish a common standard for EMR systems as well as a central repository of health records are also underway. Some of the benefits already evident are efficiency gains, reduction of duplicate tests, quicker referrals, improved continuity of care and fewer adverse drug events.[10] EMRs allow doctors to quickly and automatically share records with specialists, lab technicians or hospitals working with their patients and, where EMR systems are interoperable, for records to follow patients when they move. Several provinces have EHR systems in place that allow patients to access their health records online, and options like telephone and Internet access to care are expanding.

8 In 2014, 79% of physicians in Canada were using EHRs, compared to 90% internationally, according to this survey: Canadian Medical Association (2014). *How can Canada achieve enhanced use of electronic medical records*. https://www.cma.ca/Assets/assets- library/document/en/advocacy/Enhanced-Use-of-EMRs-Discussion-Paper-Final-May-2014.pdf.

9 Canada Health Infoway (2018). *Connected Health Information in Canada: A Benefits Evaluation Study*. https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/benefits-evaluation/3510-connected-health-information-in-canada-a-benefits-evaluation-study.

10 Canada Health Infoway (2013). *The emerging benefits of electronic medical record use in community-based care*. https://www.infoway-inforoute.ca/en/component/edocman/1224-the-emerging-benefits-of-electronic-medical-record-use-in-community-based-care-full-report/view-document.

Digitizing medical records is a necessary precondition for AI and big data's move into health. The power of these technological tools comes from gathering together a large amount of data in a machine-readable format. Algorithms can only uncover patterns that might help improve health if health information is encoded in digital records, and many AI algorithms require huge amounts of data. Digitizing health records is the first step, however, the records also need to be stored in a format that can be read by other machines, and combined into larger datasets.

There are a number of challenges to address to make EMRs suitable for AI innovation:

» EMRs created using proprietary software often cannot be exported into an accessible format, making them difficult to share.

» Different healthcare providers use different EMR systems that code information in incompatible ways.

» Front-line healthcare workers do not always enter information accurately and completely, so records contain potentially misleading gaps and typos, which may be harmless in clinical settings, but could cause problems for algorithms.

» Unneeded fields in EMRs are sometimes used idiosyncratically to record unrelated administrative data.

» Some kinds of information that is highly relevant to health is not present in EMRs, such as income, working hours, availability of family support and local transportation options.

» All of these challenges make combining EMR records into interoperable EHR datasets a laborious process of translation.

These considerations point to the need for a standardized, centralized EHR system. Already in 1999, the Advisory Council on Health Infostructure commissioned by the Federal Minister of Health envisioned integration and standardization of data in a pan-Canadian EHR system.[11]

Informants from very different fields all agreed on this point. A physician and genomics researcher said that, "Government should be responsible for interoperable EHR systems... Nobody else can do it."[12] A social worker overseeing the rollout of EMRs in a rural health network wished that "government could support people at community levels to figure out what's best for their needs"[13] and worried that without that direction, private companies are going to win contracts to make systems that are incompatible with neighbouring health authorities, simply because doctors are too busy caring for patients to negotiate with software developers.

The director of a health informatics company noted that there's a need for "standardization and a cooperative approach across hospitals"[14] because currently they all have different rules, but was hopeful about the Ontario Ministry of Health's openness to creating a central system and making records available to research. They also noted that some companies who make EMR systems have access to all the data stored in them, because physicians signed away those permissions. Access to patient data by third parties ought to be more carefully controlled, and could be if standards were in place. A lawyer working for the Ontario government also agreed that a centralized EHR system would be helpful to have.

11 Advisory Council on Health Infostructure (1999). *Canada Health Infoway: Paths To Better Health final report*. Ottawa: Health Canada Publications. http://publications.gc.ca/collections/Collection/H21-145-1999E.pdf.
12 Interview, 2018.
13 Interview, 2018.
14 Interview, 2018.

There were concerns raised about how widely and under what conditions health records should be shared. In Ontario, explicit consent is not always required before records are shared with other providers. A 2015 clarification[15] of the Personal Health Information Protection Act (PHIPA) details how records can be shared within a "circle of care" that includes all healthcare professionals involved in a patient's care, on the basis of assumed implied consent. In practice this means that clinics and hospitals have posters and pamphlets on display informing patients that their information will be shared, but in most cases there is no follow-up with patients to make sure they saw the information, understood what it means for them, do agree to the sharing, and are aware that they can withdraw consent.

A worker at a women's mental health clinic, where patients often have histories of trauma and/or drug use, noted that assumed implied consent can cause problems for their patients. If a traumatic experience is shared with a trusted healthcare provider it can be re-traumatizing to have it brought up by another, untrusted provider. Stigmatized information like mental health diagnoses and drug use can negatively affect the quality of care patients receive from other providers, and if disclosed to outside agencies can affect child custody cases in ways that may not be in the best interests of the patient or their children. That clinic went back to their pre-2015 practice of asking for explicit consent to data sharing to mitigate these problems.[16]

In theory, patients can opt out of sharing within the circle of care. But many patients do not know who their information is being shared with, nor what their rights are. Furthermore, a lawyer working in public health reported that the opt-out features are not fully functional in the software healthcare providers use: "There's supposed to be 5 levels of granularity, but it's difficult to implement."[17] So even when patients do ask not to have their data shared with certain people, that is not always technically possible. Similarly, Alberta gives patients the opportunity to request that some information be withheld from the provincial EHR system, yet does not have the administrative tools to support those requests.[18]

Stigmatized information like mental health diagnoses and drug use can negatively affect the quality of care patients receive.

Privacy audits are another area where implementation lags behind regulation. Healthcare providers are legally required to audit access to their patients' health records to ensure that inappropriate access is not made. According to the privacy officer at a community health centre, however, there are few guidelines on what a privacy audit should entail, administrative roadblocks that get in the way of accessing the necessary information, no accountability mechanism in place to ensure they get done, and a lack of funding to actually do the audits.[19]

15 Office of the Information and Privacy Commissioner of Ontario (2015). *Circle of Care: Sharing Personal Health Information for Health-Care Purposes*. https://www.ipc.on.ca/wp-content/uploads/Resources/circle-of-care.pdf.
16 Interview, 2018.
17 Interview, 2018.
18 Armstrong, Wendy (2011). "Getting Lost in Doing Good: A Societal Reality Check" in *Data Data Everywhere: Access and Accountability?* Ed. C.M. Flood. McGill-Queen's University Press.
19 Interview, 2018.

Several informants expressed concern that once health data are collected in an EHR system, patients lose control over how their data will be used. An open data advocate quoted Lucy Bernholtz's mantra, "If you can't protect it, don't collect it." An Ontario public servant warned that "collection of data isn't always used for the initial intention," citing how the Nazis used census data in The Netherlands to identify Jewish people as an extreme illustration of how data can be misused.[20] In a 2007 report on secondary uses of health data, the main distinction between EHRs and EMRs was their completeness; it was assumed that EHRs would be much more selective in which data are included.[21] A number of informants were unaware of this distinction, reflecting the currently popular assumption that EHRs will include as much data as possible, with the only distinction being how locally the data are held.

Once health data are digitized and stored in interoperable EHRs, the expectation is that it will be made available for a variety of secondary uses that go well beyond healthcare delivery. It is in this domain of secondary uses that AI will have its biggest impact.

20 Interview, 2018.
21 Roy, David & François Fournier (2007). *Secondary Use of Personal Information Held on National Electronic Health Record Systems: Key Developments, Issues and Concerns*. Montreal: Centre for Bioethics, Clinical Research Institute of Montreal.

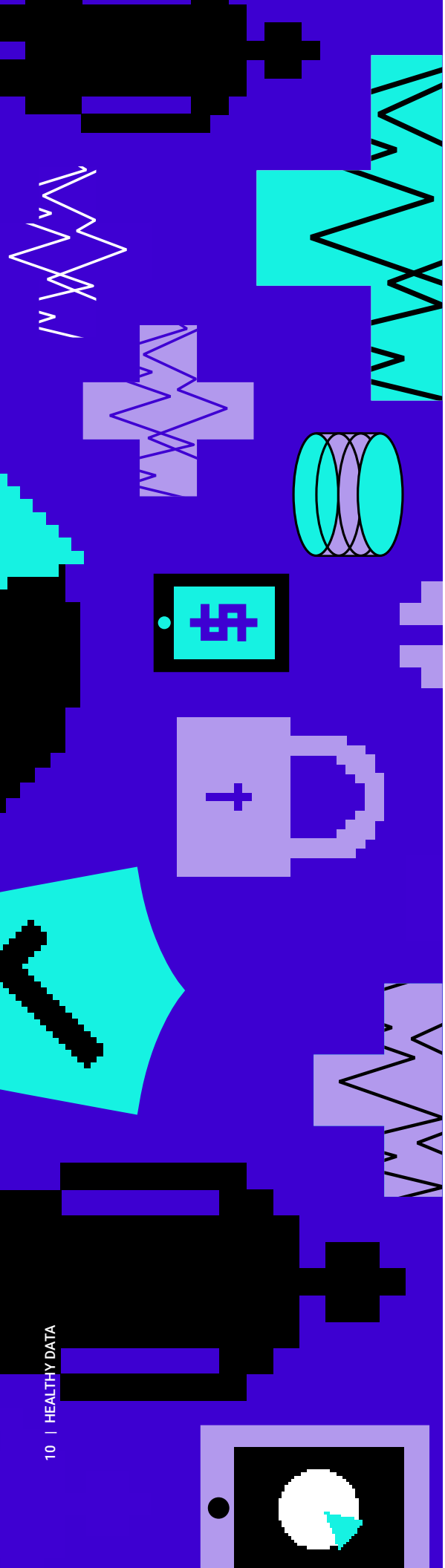
Digitization's Slippery Slope

Jurisdictions where health data infrastructure is more advanced than in Ontario have seen a gradual slide away from protecting patients' rights as digital services take hold.

Alberta's province-wide EHR system, which began development in 1999,[22] exemplifies this trend. According to Armstrong (2011), when the system was rolled out, one-time patient consent was required to upload records into the provincial EHR system, but that requirement was removed in 2003. In 2006, disclosure to police and insurance companies began to be allowed without patient consent or knowledge. By 2007, tens of thousands of healthcare providers, administrators and others could browse the records based on an honour system, with fines as the only privacy protections. In 2008, IT security was reported to be woefully inadequate. In 2009, the Alberta government proposed a bill that would require doctors to upload their records, with refusal punishable with massive fines. This move finally resulted in push-back from medical associations and individual Albertans, prompting amendments to Alberta's Health Information Act.[23]

The reasoning often given for these erosions of rights is that better protections would be an administrative burden. Researchers observed that "The likely effect of deferring questions concerning secondary uses will be an exacerbated policy dilemma that drives solutions further away from the well-established norm of voluntary and informed consent as a core component of privacy protection."[24] This prediction is proving accurate.

22 Alberta Netcare EHR (no date). *The History of the HER*. Government of Alberta. http://www.albertanetcare.ca/History.htm.
23 Armstrong, Wendy (2011). "Getting Lost in Doing Good: A Societal Reality Check" in *Data Data Everywhere: Access and Accountability?* Ed. C.M. Flood. McGill-Queen's University Press.
24 Kosseim, Patricia & Brady, Megan. (2008). *Policy by Procrastination: Secondary Use of Electronic Health Records for Health Research Purposes*. McGill Journal of Law and Health, 2:5—45. https://mjlhmcgill.files.wordpress.com/2017/07/mjlh_vol2_kosseim-brady.pdf



Public support for secondary uses of EHRs is high, but conditional. Surveys consistently show that 80 per cent or more of Canadians are comfortable and even expect that their health data will be shared with hospitals, university researchers, organizations like Statistics Canada and departments of health for secondary purposes. According to a 2012 HarrisDecima survey, however, about 70 per cent of Canadians are not comfortable sharing health records with private businesses, including pharmaceutical and insurance companies. Only half of Canadians are comfortable with their identifying information being retained on their records when the records are shared for secondary uses. 66 per cent insist that their consent should be required if their records are going to be shared with identifying information included, and an additional 26 per cent say they would prefer to be asked for consent. The number of respondents who say they would give that consent is on the rise, with only 9 per cent saying they would refuse. 71 per cent would be more comfortable with a lack of consent if secondary research had to be approved by an independent agency that assessed privacy risks. 86 per cent of Canadians expect to be notified about when and how their information is used.[25] Despite there being a clear expectation that secondary use of health data requires notification and consent, or at least independent oversight, common practice in health research is moving away from the informed consent model.

Canada Health Infoway lists three “emerging areas” where EHRs can have an impact on healthcare: patient-generated data from consumer medical technologies, predictive analytics and data governance partnerships.[26] We will look at each in turn.

25 HarrisDecima (2012). *2012 Public Opinion Research: Canadian Views on Electronic Health Records*. https://www.infoway-inforoute.ca/en/component/edocman/1687-harris-decima-survey-on-electronic-health-records/download?Itemid=101.

26 Canada Health Infoway (2018). *Connected Health Information in Canada: A Benefits Evaluation Study*. https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/benefits-evaluation/3510-connected-health-information-in-canada-a-benefits-evaluation-study.

Privacy Laws

Privacy laws relevant to the Ontario health context are spread among several federal and provincial acts, including:

» The federal *Privacy Act*[27], which covers how the federal government handles personal information.

» The federal *Personal Information Protection and Electronic Documents Act*[28] (PIPEDA), which covers how businesses handle personal information.

» The Ontario *Personal Health Information Protection Act*[29] (PHIPA), which covers the health information custodians (doctors, hospitals, etc.) involved in the delivery of healthcare services.

Under PIPEDA, identifiable personal information has special status, restricting its collection, handling and disclosure. Some of the rules that apply to personal information are:

» It may be collected, used and disclosed only for "purposes that a reasonable person would consider appropriate."

» Collection, use and disclosure without knowledge or consent may only be done under exceptional circumstances, such as in case of a security threat, in compliance with legal proceedings or for research that cannot be done otherwise.

PIPEDA includes a list of ten principles organizations are required to follow: accountability; identifying purposes; consent; limiting collection; limiting use, disclosure and retention; accuracy; safeguards; openness; individual access and challenging compliance.

PHIPA defines the additional responsibilities that health information custodians bear in their handling of health data. These include the requirement to record access to records, report any data breaches and not disclose health information to individuals who are not directly involved in a patient's care, with a few exceptions. PHIPA also sets out guidelines for health research projects, research ethics boards and the prescribed persons and prescribed entities to whom custodians may disclose data without consent, for purposes like health system administration, disease registries, cancer screening and public health.

Health data that is generated outside the healthcare industry, such as by wearable technology, is considered sensitive data under PIPEDA, but is not covered by PHIPA. There are borderline cases where it can be difficult to tell which jurisdiction is responsible, for example when businesses work with hospitals on health technology, or hospitals collaborate across provincial borders. When custodians transfer health data to industry partners, strict data sharing agreements must be put in place.

The federal Office of the Privacy Commissioner (OPC), and the provincial Office of the Information and Privacy Commissioner (IPC) have different powers of enforcement. Individuals may file complaints with the OPC, or the OPC may audit organizations with reasonable grounds. The OPC has the power to investigate, and compel persons to give evidence. Court hearings may be brought either by individuals or the OPC, on the basis of their report, with the possible outcome that the organization is ordered to correct its practices, and/or pay damages to the complainant. The IPC may investigate complaints, gather evidence and make orders requiring compliance. Health information custodians are required to report privacy breaches to the IPC and notify the affected individuals. Only the Attorney General may initiate prosecution under PHIPA, which may result in fines.

27 Privacy Act (1985). R.S.C., c. P-21. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-privacy-act/.
28 *Personal Information Protection and Electronic Documents Act* (2000). SC 2000, c 5. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/.
29 *Personal Health Information Protection Act* (2004). SO 2004, c 3 Schedule A. https://www.ontario.ca/laws/statute/04p03.

The Tech Industry

In Toronto alone there are dozens of technology startups developing new applications that gather data from users to provide health and wellness services. There are apps to help treat depression, renal failure, heart conditions, eating disorders and speech pathologies. In addition to wearable gadgets like the FitBit, there are devices to help children with spinal injuries walk upright and headsets that help you concentrate. There is an appetite among entrepreneurs to innovate in this arena, and an appetite among consumers to try more flexible, responsive tools to manage their health.

While the data these apps and gadgets gather is technically health data, PHIPA only covers health data held by, or originating from, health information custodians (doctors, hospitals, public health authorities, etc.). Although PIPEDA still applies, commercial health data is left "operating in a legal gray area" according to a lawyer working in health research.[30] Likewise for data collected from social media activities, like status updates that discuss medical symptoms or reveal psychological states,[31] and genomic data held by commercial DNA profiling services.[32] Some of the companies that hold health data use it for research purposes[33] or share it with third parties, and their consent procedures often involve little more than clicking past a privacy policy. Another worrying trend is for insurance companies in the US to require customers to share health data in order to qualify for policies.[34]

A legal scholar noted that technology companies have no accountability in how they use these data, because "there's little that can be done to enforce the agreements that are made for how data will be used." Fines for privacy violations are becoming just a cost of doing business. In the public sector this is not so. Researchers working in hospitals and universities need to get approval from an REB before undertaking experiments with human participants, or they risk losing their jobs and funding. An AI researcher working in health said, "AI workers and tech companies need more knowledge of REBs and the importance of consent for human experimentation… They are required to get REB approval, but outside of universities and research institutions there is no oversight," which means that tech projects that do not involve university or hospital partners often proceed without an ethics process. Some larger companies are setting up internal REBs due to increasing consumer pressure, but without independent oversight, these may do little more than rubber-stamp proposals.

An open data expert called for a highly visible, easy-to-use framework to protect consumers from having their data collected and used without their knowledge or consent. They suggested a system analogous to nutritional labels on food: "you can look at the label in 30 seconds and understand what you're getting."[35] The majority opinion was that the regulations that apply to public sector health data should apply to private sector health data too, and that better enforcement mechanisms are needed.

30 Interview, 2018.

31 Kramer, Adam D. I. et al. (2014). *Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks*. Proceedings of the National Academy of Sciences. http://www.pnas.org/content/111/24/8788.full.

32 Seife, Charles (2013). *23andMe Is Terrifying, but Not for the Reasons the FDA Thinks*. Scientific American. https://www.scientificamerican.com/article/23andme-is-terrifying-but-not-for-the-reasons-the-fda-thinks/.

33 Eriksson, Nicholas et al. (2010). *Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits*. PLOS Gewnetics. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000993.

34 Barlyn, Suzanne (2018). *Strap on the Fitbit: John Hancock to sell only interactive life insurance*. Reuters. https://www.reuters.com/article/us-manulife-financi-john-hancock-lifeins/strap-on-the-fitbit-john-hancock-to-sell-only-interactive-life-insurance-idUSKCN1LZ1WL.

35 Interviews, 2018.

Despite these worries about privacy and consent, several informants were enthusiastic about using data collected by consumer technologies for improving health. There is great potential in being able to consistently monitor conditions between doctor's office visits, correlate symptoms with lifestyle factors to learn about the social determinants of health and intervene immediately when signs of distress are detected. A consultant working in health technology suggested that we have a moral duty to make use of all the tools and data available to us if they can help save lives.

AI health businesses are also seeking out partnerships with hospitals, university researchers and publicly funded data custodians to get access to Canadians' health data. As mentioned, some private businesses are getting access to health data through the EMR platforms they sell to doctors. A project manager considering a new EMR system reported that, "nobody brought up the topic of access to info by the tech companies bidding to upgrade our system. Doctors don't have time to think about that." Just as doctors may not know that they're giving away access to their patient data, tech companies "seem very unaware of their obligations" and do not know that "they can't use data for secondary uses" under PHIPA, according to a public servant who works in privacy. A consultant to the tech industry reported that they regularly have to "tell people in startups to learn about health data," because they assume they'll be able to get access to any data they want.[36] Several informants identified this lack of knowledge on the part of tech workers about how they can and cannot use health data as a problem, and called for better education and outreach.

The consensus view is that we need to do better at teaching applied ethics in STEM subjects like computer science. Some suggested that all computer science programs should have a required ethics class. An AI ethics researcher said, "just a single unit of ethics doesn't go far enough… it needs to be across all courses students take" to be effective. A few people stressed the importance of showing how strategies like increasing the diversity of the workforce, collaborating in interdisciplinary groups, and paying attention to fairness, accountability and transparency, could directly benefit tech companies. We heard, "Make clear it's not obstructionist," "frame it as a mutual interest," and "in machine learning the most valuable data points are the outliers." There was agreement that ethics ought to be taught from a "practical angle," "not starting at, like, Plato." Partnerships between AI researchers and clinicians were also recommended: "it's important to have an interdisciplinary conversation;" "collaboration is essential."[37]

AI in Medical Research

There are also a significant number of AI analytics projects happening within the public sector. Every hospital we spoke to had at least one such project underway, with AI researchers and clinicians working in partnership to collect hospital data, link it with other datasets to create large-scale models and use those models to discover patterns that might lead to health improvements. Academic researchers in fields like epidemiology and public health are also working in partnership with health agencies to perform large-scale data analyses on patient data. These researchers are also working to collect data about groups who are not represented in clinical datasets, either because they are reluctant to disclose information directly to clinicians, or have limited access to healthcare.

36 Interviews, 2018.

37 Interviews, 2018.

AI methods in health

The subfield of AI currently making a big impact on healthcare is machine learning. Contemporary machine learning methods (deep learning, regression methods, dimensionality reduction, Bayesian networks, among others) typically require large datasets to work well. In the training stage when the model is being built, these algorithms pore through the data picking out patterns and regularities to build a model. (Some types of algorithms continue to refine the model after it's in use.) The model represents the entirety of the dataset used in training. Once a model is built, it can then be used to better understand new data. For example, when a patient arrives in the ER with a set of symptoms and needs a diagnosis, a model that has learned associations between symptoms and diagnoses for millions of patients might do better at guessing the most likely diagnosis than can a physician who has only seen hundreds of similar patients. AI models can also be used to read diagnostic scans, prioritize patients through triage, find genetic markers associated with disorders and predict the effects of drugs in particular patients, among many other actual and projected uses.

In order for AI models to perform well, they not only need a large quantity of data, but also high-quality data. High-quality health data means medical records that are accurate, relatively complete, recorded in a format that is easy to work with numerically, and covering an unbiased sample of patients. Models built on data from one hospital may not always perform well at different institutions with different patient demographics. A dataset that includes only patients from a private university clinic would cover a different range of ages and socioeconomic statuses than a mobile clinic covering remote northern communities, for example. The dataset used to train the model should include the same diversity as the population the model will be used to serve, otherwise some kinds of patients might not be well represented. When models built for one population are to be used more widely, they may need to be recalibrated, or used only with patients who fit certain parameter ranges.

Researchers and clinicians working in partnership made the strongest case for the benefits of increased data sharing and applications of AI methods in health. An AI researcher working in genomics said, "If you don't have open data, you'll never realize what kinds of rare genetic disorders are out there." Both an AI researcher working in mental health, and the privacy officer in a large urban hospital talked about improving patient outcomes by developing "data-driven care pathways." That means using solid evidence to decide questions like which medication to try first and at what dose, which patients to admit to hospital and for how long. One hospital is working on integrating lifestyle data, like income, mobility and frequency of exercise, with its ER data, to redirect unnecessary visits to preventative community care. A public health researcher talked about the potential to improve people's lives by linking health data to details about "where you grew up, were you prepared for entering kindergarten, did you graduate, did you get a job and where?" to support the development of more effective public health interventions.[38] It is important to emphasize that this sort of large-scale linking of detailed data about multiple aspects of people's lives and use of powerful analytical tools to find patterns in that data is a novel development. Aside from the national census and a handful of longitudinal studies, this scale of project has rarely been attempted before.

38 Interviews, 2018.

Researchers and clinicians felt strongly that patients want to participate in research. One clinician said, "We're collecting data as we're resuscitating patients and saving their lives… We'll ask for permission to follow the patient's recovery over years, and so far no one has turned us down." An AI researcher said, "There are massive patient advocacy groups that give up all kinds of detailed information if they think it will get them help." Another AI researcher noted, "People are looking for ways to provide impact by sharing their data." An epidemiologist working with a vulnerable population noted that even where privacy risks are highest, "people are often willing to give lots of detailed information if they really believe that it's for the public good."[39]

Two of the institutional structures in place to ensure that data are not misused are data sharing agreements and REBs. Data sharing agreements between healthcare providers and researchers seeking to do secondary research include clauses about what security measures are required, how long data can be kept on file, who will have access to the data and sometimes conditions on publication. These agreements are "getting longer and more complicated," according to the privacy officer at a large hospital, who admitted that the job of "tracking the projects over multiple years, with changing agreements" was becoming overwhelming, even with a team of lawyers. An AI researcher complained that getting access to patient data was impossible for small tech firms without specialized health lawyers on staff, because data sharing agreements are prohibitively complex. An academic researcher noted that although universities have "templates for data sharing agreements and informed consent contracts, these differ, even within a university," and some of them do not give adequate protection to participants and researchers.[40] All three called for standardized language and protocols for these agreements. Several people worried that there is no audit mechanism to ensure that data sharing agreements are being honoured, and wondered whether Canada should have regulations similar to the European Union's recently enacted General Data Protection Regulation (GDPR).[41]

Approval from REBs is one layer of protection against inappropriate uses of data in research. Some of the protections an REB might demand include that personal data be stored anonymously and securely, destroyed after the end of the study or not collected at all. Several informants pointed out weaknesses in the REB system. The most common complaint was that REBs at different institutions hold researchers to very different standards. One interviewee reported that a hospital they work with was "using independent research boards for research, because they're less stringent than the hospital's ethics board."[42] While that may sound like an attempt to cut corners, several researchers claimed that REBs in smaller organizations are overly cautious, because they don't have access to lawyers who are experts in privacy legislation. We heard several calls for better training or certification of REB members, and suggestions that "we may need to move away from a voluntary format" or that we might "be better off with centralized or specialized [REBs]."[43]

39 Interviews, 2018.

40 Interviews, 2018.

41 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016). OJ L 119, 4.5.2016, p. 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A32016R0679.

42 Interview, 2018.

43 Interviews, 2018.

Privacy

One potential harm of greater data sharing and AI analysis falls on the individual offering up their data, should that data not be kept private. For most people, a data leak would be a minor annoyance or embarrassment. As one informant put it, "Identifying an individual isn't a concern; people don't care."[44] But in the case of people who have highly stigmatized information in their medical records (like HIV status, drug use, disability, mental illness, overweight, LGBTQI* identity), having that information leaked could have serious consequences for jobs, housing, relationships, insurance, personal safety or immigration status. The risk of a privacy breach is not evenly distributed across the population, so strategies for protecting privacy need to focus on the people actually at risk.

When hospitals enter into data sharing agreements with research partners, typically the first step is to de-identify the data. At a bare minimum, de-identification means stripping off personally identifying information, like name, full birth date, full address and health card number. Thorough de-identification also requires masking any data points that are rare in the dataset. For example, ages over 90 years are rare enough that they could give away a person's identity, so the top range of ages might be replaced with "85+". Details are inevitably lost in the process, so de-identification needs to be sensitive to context. A study focused on the health of seniors should not blur all ages over 85. Rare combinations of characteristics likewise need to be masked, and re-identification risks carefully measured.[45]

Vulnerable individuals tend to be the ones who are most easily identifiable in anonymized datasets. An accessibility advocate pointed out that "people with disabilities are the outliers, the anomalous, so it's easy to find out who the data is about. We're also the people whose data is abused or misused." Another researcher concurred that among the gender diverse population they work with, "Some people have really unique characteristics." Both argued that de-identification does not serve the needs of the populations they work with, because "De-identification strategies that deal with outliers eliminate us from the dataset. Then people are making decisions about you without you there."[46] This adds to the invisibility caused by differential access to healthcare, and surveys not including questions specific to vulnerable groups.

Re-identification risks are never zero, so the benefits of particular research projects need to be carefully weighed against the potential harms. Typically, the benefits will be spread across the population and over a long timeframe, while the risks are more pointed and specific, making a direct comparison difficult. For some types of data, de-identification is not possible; genomic data is one example where even a small amount can be used to re-identify an individual. Brain scans may also fall into that category.

Consent

When health data is not de-identified, researchers require the patient's informed consent to use the data for secondary purposes (with a few exceptions, like health system administration, cancer registries and public health uses). Three main concerns about consent were raised in interviews: barriers to getting consent, erosion of consent standards and consent bias. AI poses additional problems for consent.

44 Interview, 2018.

45 El Emam, Khaled (2013). *Guide to the De-identification of Personal Health Information*. Boca Raton: CRC Press.

46 Interviews, 2018.

The main barrier to getting informed consent is ensuring that patients really understand how their data is going to be used and what the consequences are. The Personal Genome Project,[47] for example, requires participants to go through a comprehensive study guide that explains how genomic studies work and how the data might be used, then take an enrolment exam before consenting. If you want patients to be truly informed, this places an extra burden on participants in terms of time and education. Requiring an entrance exam about genetics ensures that consent is well informed, but effectively bars most of the population from participating, making the dataset biased. More elaborate informed consent procedures are also more time consuming and expensive for researchers.

AI applications in health share with genomic studies a high bar for patients' understanding as well as great uncertainty about the possible results, because the uses of a model cannot be easily predicted before it is built. For large AI projects, patients are asked to consent to unknown future uses of their data. This is in tension with the principle of informed consent, but REBs have been approving this kind of study where future uses of data are unknown.

Once a model is built it is often very difficult to figure out what the basis of its decisions are, or to audit the outcomes. If the training data was biased, the model's decisions will likewise be biased, which, according to an AI researcher, "can lead to harmful medical findings."[48] Cancer and asthma research are just two well-known areas where racial bias in clinical trials has led to the development of treatments that do not work well on the populations where prevalence is highest.[49] When embedded in an AI model, this bias becomes largely invisible, making it harder to question the reliability of a machine's decisions.

Because biased datasets generate biased models, AI projects are particularly concerned with avoiding *consent bias*, where the consent process itself creates bias. There is some evidence to suggest that willingness to consent to research is not consistent across sectors of society. El Emam (2013) reviews 27 studies that report consent biases. Effects of age and gender on consent were mixed. Unmarried people, African Americans and those with lower education levels were found to be less likely to consent across several studies. Patients with the disease under study were more likely than others to consent. While those results suggest that consent bias is a pervasive phenomenon, the details of the studies reveal a more nuanced picture. Opt-in consent had lower participation rates in lower education and socioeconomic groups than opt-out, and likewise written consent resulted in lower participation rates than verbal consent.[50] This suggests that for many, the reason for not consenting is that the paperwork is burdensome, either because of time constraints or literacy.

Consent bias and the burden of consent on patients have been cited as reasons to relax consent standards. As mentioned already, many healthcare providers have moved from an explicit to an implicit consent model within the circle of care. Broad consent, where a patient gives one-time, opt-in consent for a variety of future research purposes, is becoming more prevalent. A public health researcher said hospitals are moving toward "blanket consent, where patients opt-out

47 Personal Genome Project Canada (no date). *About*. Toronto: Personal Genome Project Canada. https://personalgenomes.ca/.
48 Interview, 2018.
49 Konkel, Lindsey (2015). Racial and Ethnic Disparities in Research Studies: The Challenge of Creating More Diverse Cohorts. *Environmental Health Perspectives, 123*(12), A297–A302. http://doi.org/10.1289/ehp.123-A297.
50 El Emam, Khaled (2013). *Guide to the De-identification of Personal Health Information*. Boca Raton: CRC Press, pp 49-50.

rather than opt-in." Several researchers called consent "a hurdle" and one expressed the wish to "not have to keep going back" for every use. An expert on health policy noted that in cases where whole communities or families have a stake, "informed consent for a single individual doesn't safeguard everyone who might be affected." Conversely, others reported that "consent is well-recognized as being necessary" or insisted that "getting no consent is not an option," although "we need better education about consent for certain populations."[51]

Trust

A direct and personal approach can be effective in reducing the burden associated with consent, mitigating consent bias. That more personal approach also helps to instil trust, which is also essential for consent. A human rights advocate stressed the role of communication in building trust, "you have to walk communities through why you're doing what you're doing." A paediatric health researcher shared that "families don't fully understand [the research they're consenting to]. They give consent based on trust." Researchers and caregivers working with vulnerable populations (including First Nations communities, people with disabilities, mental health patients, and gender diverse populations) also cited trust as the essential ingredient in getting consent.[52]

Trust is also important to getting good quality data. A data privacy specialist reported that "certain people will lie about medical information if they feel the data is being misused or mishandled." An epidemiologist working with vulnerable populations concurred, citing trust as essential for the "ability to generate high-quality data… so that we're not trading off validity in research."[53]

Beyond honest and open communication about intentions and risks, trust depends essentially on making good use of the data. This means using the data to generate health benefits, refusing to use the data for purposes that might stigmatize or otherwise harm the population offering the data, and sharing the results openly with the data subjects. That the public has limited understanding of AI methods, and that AI has a reputation for ethical blunders, makes trust especially difficult to establish for AI projects in health.

Trust can be earned through efforts to improve digital literacy and develop ethics oversight in AI. How much knowledge of AI we need at the policy level was a point of disagreement among informants, with one insisting that "technical stuff is driving policy issues, you can't give policy recommendations without it" and another claiming that you "don't need to be tech savvy in order to fix the gaps in healthcare services."[54] There was agreement that countering the sensationalized narratives about AI appearing in the popular media is important. The consensus was that better understanding between policymakers, healthcare providers and AI workers is best achieved through collaboration, where various stakeholders are at the table together, working on a common problem. Industry-led AI ethics initiatives were met with suspicion by most informants. Two informatics workers suggested an accreditation or certification program through which AI researchers could become trusted individuals.

51 Interviews, 2018.
52 Interview, 2018.
53 Interviews, 2018.
54 Interviews, 2018.

Data Governance

Another kind of harm that can come from greater data sharing and analysis operates at the community level. Both over-studied populations like First Nations members and refugees, and under-studied populations like trans people and people with disabilities are concerned with community level harms, and have begun developing responses that protect their communities.

The First Nations Information Governance Centre (FNIGC) has a well-developed set of principles designed to bring ownership, control, access and possession (OCAP) of data about communities back into the hands of those communities. Their OCAP principles were developed in response to decades of injustices brought about by outside researchers and government agencies. On multiple occasions, outsiders have collected data or biological samples from First Nations communities, ostensibly for public health purposes, like diabetes or cancer screening, then without consent used the data for secondary purposes that stigmatized communities, led to loss of services or profited the researchers without giving anything back to the community.[55]

Such uses are allowed, because only a few First Nations communities are designated as "governments" under Canadian law, thus the health data that government agencies collect from most Indigenous communities are not protected by provincial health privacy laws. Instead, those data are subject to public disclosure under the federal *Access to Information Act*. Those disclosures are anonymized, but First Nations, Metis or Inuit identity is not removed from records.

Once an individual has been dead for more than 20 years, even their personal information becomes publicly accessible.[56] This lack of privacy of health information is out of line with the protections other Canadians enjoy. Although the details differ in other countries, similar patterns of data injustice occur in many places, and the response by Canada's First Nations is part of a global Indigenous data sovereignty movement.[57]

Outsiders have collected data or biological samples from First Nations communities, ostensibly for public health purposes, like diabetes or cancer screening, then without consent used the data for secondary purposes that stigmatized communities.

FNIGC now runs regular health surveys of their own, and provides the data in restricted form to Health Canada and other agencies under strict data use agreements. They also charge fees for data access and use, which ensures that communities profit alongside researchers. FNIGC also has a partnership with the Institute for Clinical Evaluative Sciences (IC/ES), wherein IC/ES acts as data custodian, holding FNIGC data and allowing it to be used for research purposes subject to OCAP principles. IC/ES is designated as a *prescribed entity* under PHIPA, which allows it to collect, use and disclose personal health information for some purposes without needing consent and, as it is not a government agency, it is not subject to Access to Information requests.

55 First Nations Information Governance Centre (2014). *Ownership, Control, Access and Possession (OCAP): The Path to First Nations Information Governance*. Ottawa: The First Nations Information Governance Centre. https://fnigc.ca/sites/default/files/docs/ocap_path_to_fn_information_governance_en_final.pdf

56 First Nations Information Governance Centre, 2014.

57 See, for example, Science for Technological Innovation (2018). *Maori Data Futures Hui Report*. Te Herenga Waka Marae Victoria University of Wellington. https://www.sftichallenge.govt.nz/sites/default/files/2018-09/Ma%CC%84ori_Data_Futures_Report.pdf.

A "fortress model" inspired by OCAP is the most viable option for balancing the desire for large, high-quality, linkable datasets that afford innovation, with the privacy protections that more open data governance models cannot guarantee.

One informant described how disability advocates are developing similar systems of data co-ops, where patients (with rare diseases, for example) can pool their data, and co-op members govern who uses the data and how it is shared, "so they can determine how the value from their data is distributed."[58] Ontario's TransPulse[59] project does something similar with survey data from the trans population, which are used and communicated only based on approval from a committee of community members. These communities distrust medical institutions due to "histories of pathologizing research" such that without this measure of control, members of these groups would not be willing to honestly disclose potentially stigmatizing information.[60]

Open data has its uses, but is only appropriate for aggregate health data. For other purposes, like linking across datasets, or being able to follow up with patients after a discovery relevant to their health is made, retaining identifiable information in the dataset is essential, but this requires the security of a closed dataset held by a trusted authority. Several informants, including informatics researchers, healthcare providers, administrators and legal experts all agreed that a "fortress model" inspired by OCAP is the most viable option for balancing the desire for large, high-quality, linkable datasets that afford innovation, with the privacy protections that more open data governance models cannot guarantee. In less militaristic terms, this fortress model is often referred to as a "data trust." Data trusts are being discussed as tools for allowing data access without sacrificing rights or control over data for AI applications,[61] and in regulating smart cities.[62]

One health IT worker called for "secure facilities with better analytics systems… so you can analyze the data on an interactive platform and not be able to download a file of data." Another expected to see "dedicated units within government that do big data analyses… to maximize security of personal information by limiting points of access." A third pointed to the need for "a standardized, single place to hold the data" but noted that instead of government, "maybe it should be held by an independent body that is expert in holding data."[63]

The only hesitation came from a public health researcher who worried that "to look at population-based trends over time, you can't do it with thousands of data repositories across the country, all in the control of different authorities."[64] Given that data trusts like IC/ES and the Ontario Brain Institute have been rapidly expanding their data holdings, the worry that data may become spread around too many repositories may be overblown. Some hurdles to getting automatic access to large datasets are indeed appropriate, since they

58 Interview, 2018.
59 Trans PULSE (2017). *Research and Study Results*. Toronto: Trans PULSE. See http://transpulseproject.ca/research/
60 Interview, 2018.
61 See, for example, Hardinges, Jack (2018). *What is a Data Trust?* London, UK: Open Data Institute. https://theodi.org/article/what-is-a-data-trust.
62 Wylie, Bianca and Sean McDonald (2018). *What is a Data Trust?* Waterloo: Centre for International Governance Innovation. https://www.cigionline.org/articles/what-data-trust.
63 Interviews, 2018.
64 Interview, 2018.

allow for community oversight over unwarranted surveillance, while still enabling population-wide studies where justified, despite the extra effort.

Accrediting individuals to be trusted users of health data is an option several informants raised. A public health researcher suggested using "the StatsCan model where some data is freely downloadable, some is more protected, and to see and use it you have to go through a screening process." An informatics researcher suggested that institutions should have "accredited people, like a data driver's license." However an accessibility advocate expressed frustration at "the notion of a quick fix, like an accreditation system for access to data," arguing that a one-time training course does not ensure that a person is trustworthy. "Privacy isn't stamped once."[65]

65 Interviews, 2018.

A Case Study in Data Governance

A health informatics researcher told us about a case that tested the balance between the desire to innovate and the need to protect patients' rights. A pharmaceutical consortium was looking for a way to do a post-market surveillance survey to gauge the safety of the drugs they produce. This required detailed data about people's health covering a large population. Releasing such a large dataset to a private company who could stand to profit from access would pose privacy risks, and go against the public's disapproval of access to health records by pharmaceutical companies. Aggregate data would not be useful for determining the safety of the drugs. Yet the pharmaceutical company has a duty to perform the study in the interest of public safety.

The solution reached was for the company to hire a professional health data analyst who had trusted status with the data holders. That analyst could access the data and perform the analyses needed to measure drug safety without revealing the raw data to the company. The results of the analysis could then be returned to the pharmaceutical company, and published in a public forum.

The technical challenge is to set up interoperable data collections that are structured so as to encourage innovation, while building in the needed consent and privacy mechanisms. The ethical/ legal challenge is to control access to the data such that only safe, appropriate uses are made of it by trusted entities.

3 POLICY RESPONSES

This section synthesizes conclusions about best practices and policy responses from the strategies and concerns discussed in Section 1.

A recent report from the United Kingdom about data-driven health identifies the need for large quantities of data as the main hurdle AI poses. This can be divided into two "central challenges," one technical, and one ethical/legal. The technical challenge is to set up interoperable data collections that are structured so as to encourage innovation, while building in the needed consent and privacy mechanisms. The ethical/legal challenge is to control access to the data such that only safe, appropriate uses are made of it by trusted entities.[66] We will follow this division.

The Technical Challenge

The first part of the technical challenge is to make available a pan-Canadian EHR/EMR system that providers across the country can adopt.

ACTION POINT 1

Implement a pan-Canadian, interoperable EMR/EHR standard.

The development of such a standard would require funding for both project development and stakeholder consultation. Estonia may provide a useful model to follow.[67] Incentives for providers to implement the common standard will also be necessary, perhaps in the form of grants. Also needed are incentives for third party providers of existing EMRs to make their products compliant with the pan-Canadian standard.

ACTION POINT 2

Provide incentives for providers to adopt the standard EMR, and for 3rd party EMR vendors to make existing systems compliant.

66 Department of Health & Social Care, UK (2018). *Initial code of conduct for data-driven health and care technology*. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology.
67 Heller, Nathan (2017). *Estonia, the Digital Republic*. The New Yorker. https://www.newyorker.com/magazine/2017/12/18/estonia-the-digital-republic.

Built into this standard, should be an implementation of reasonable consent directives. Excuses for letting consent slip away are predicated on it being an administrative burden. That burden can be lightened with a system that builds in consent and privacy protections beginning from the design stages.[68] A secure method for communicating within the circle of care could also be incorporated into the administration of consent directives.

Automatable auditing of access to records should also be built in, so that privacy and security concerns can be easily addressed. Investment up front would be more efficient than needing to continually fund more manual auditing efforts by healthcare providers. Allowing audits to fall by the wayside is not an option. Auditing capabilities could be integrated with a portal that informs patients about the data they have on file, and the uses being made of those data. Several provinces have health portals already, which could be updated with this functionality.

While secure authentication is needed, the process should not be too onerous for busy healthcare workers to use. The introduction of 2-factor authentication, for example, was cited by the privacy officer at a large hospital as leading to worse overall security, because nurses found workarounds to the difficult new system, like sharing passwords.[69] Looser authentication protocols from machines located in trusted environments like hospitals than for machines accessing the portal from elsewhere may provide a good compromise.

ACTION POINT 3

Include integrated consent directives, automated privacy audits, easy-to-use levels of authentication, and built-in de-identification tools in the pan-Canadian EMR/EHR system.

The most challenging question, which is intertwined with the ethical/legal challenge, is how much information from providers' EMR records to include in regional or national EHR systems, and under what conditions. The ability to quickly gather data from large populations is needed in case of public health emergencies like the SARS outbreak, or a measles epidemic. A pan-Canadian interoperable EHR standard would go a long way to making that a reality, even without a national EHR database that includes all the fields from EMRs. In exceptional circumstances, more detailed data could be requested from regional or specialized data centres, then quickly aggregated and analyzed if the systems are interoperable. The infrastructure necessary to make those data transfers in a secure manner should the need arise ought also to be built into the pan-Canadian EHR system.

ACTION POINT 4

Design infrastructure for secure transfer of EHRs between providers and central bodies like prescribed entities.

We already have public health organizations and prescribed entities that maintain databases for the purposes of health system evaluation and public health surveillance. Expanding the holdings of those organizations to cover additional health research needs is the most promising route toward making more health data available for AI-based research without compromising patients' rights. As one public health researcher told us,

68 Cavoukian, Ann (2009). *Privacy by Design: The 7 Foundational Principles*. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.
69 Interview, 2018.

"there has never been a breach in the public health repositories."[70] However, if more data is to be included in those data holdings, concerns about how sensitive data ought to be treated will become more pressing.

State-of-the-art de-identification techniques should be used whenever health data is shared without patient consent, and should also be built into the pan-Canadian system. Greater clarity on what are acceptable methods and standards for calling a dataset de-identified are needed. However, even the best de-identification tools cannot guarantee privacy and safety for the most vulnerable.

ACTION POINT 5

Establish clear standards for what de-identification means.

An additional layer of protection from re-identification could come from strengthening the status of sensitive information. PIPEDA specifies that, "In determining the form of consent to use, organizations shall take into account the sensitivity of the information. Although some information (for example, medical records and income records) is almost always considered to be sensitive, any information can be sensitive, depending on the context." The scope of what counts as sensitive information is left unclear, as are the methods organizations ought to use in taking account of sensitivity.

One might expect sensitive information to include some of the categories protected under the Human Rights Code, such as disability, race and sexual orientation. Additionally, location data from online devices, social service use, mental health diagnosis and drug use might merit extra consideration. A legal scholar noted that, "In other jurisdictions some personal health information is given special status above and beyond other personal health information, such as HIV status, which prevents any secondary use of the data without explicit consent."[71] In cases where individual consent is not feasible, additional protocols, such as approval by a community engagement committee, could ensure an extra layer of protection.

Granting protected status to sensitive information might not require legislative changes. Section 30(2) of PHIPA states that "A health information custodian shall not collect, use or disclose more personal health information than is reasonably necessary to meet the purpose of the collection, use or disclosure, as the case may be."[72] This is informally known as the minimization principle. A clarification document from the IPC could ensure that Section 30 is interpreted as meaning that sensitive information is rarely deemed "reasonably necessary," and could direct REBs to take special care in evaluating research proposals that include collection of sensitive information, so that an extra layer of protection would be required for approval.

ACTION POINT 6

Create protected status for sensitive information, either through legislation or an interpretation document put out by IPC.

70 Interview, 2018.

71 Interview, 2018.

72 *Personal Health Information Protection Act* (2004). SO 2004, c 3 Schedule A. https://www.ontario.ca/laws/statute/04p03.

The new health technologies being developed by tech companies also raise challenges. One of the main problems is that many tech companies are unaware of how privacy legislation applies to the data they handle.

A document by New Zealand's Data Futures Partnership[73] may be helpful in designing procedures for determining what is "sensitive data" and how it ought to be treated. They list features of data use that may require special care in building acceptance, including uses that:

» are novel for the community affected.

» have an impact on Indigenous populations.

» have an impact on vulnerable groups.

» are proposed by an organization with low trust

among others. The document then describes routes for gaining social license, like establishing panels of community representatives, engaging stakeholders through social media, and being willing to change proposed data uses.

The new health technologies being developed by tech companies also raise challenges. One of the main problems is that many tech companies are unaware of how privacy legislation applies to the data they handle. Education initiatives for tech workers, and digital literacy programs for the public, are sorely needed. Many digital literacy programs are run by community groups on a voluntary basis. To expand these offerings, operating grants could be made available so that those programs that already work can reach a wider audience.

ACTION POINT 7

Offer operating grants for community-based digital literacy programs to expand their offerings to reach a wider audience.

For the digital literacy programs that already exist, there is little in the way of evidence-based advice about how to ensure that these programs are effective. There is also insufficient evidence on how to design an effective applied ethics curriculum for university students in STEM subjects. New ethics courses are being rolled out in fall 2018 because of an Ontario Government investment to increase the number of Applied Masters degrees in AI to 1,000 per year, which includes a commitment to education in ethical and social impacts of technology. This program is a significant opportunity to commit to accompanying research on the effectiveness of approaches to ethics education.

ACTION POINT 8

Commission a study to research the effectiveness of ethics education in computer science and engineering programs, and to develop evidence-based curricula.

73 Data Futures Partnership (2017). *A Path to Social Licence: Guidelines for Trusted Data Use*. https://trusteddata.co.nz/wp-content/uploads/2017/08/Summary-Guidelines.pdf.

The Ethical/Legal Challenge

Privacy law also needs to be amended to keep up to date with current technologies. People are feeling pressure to consent to the terms and conditions of digital services without being provided accessible information about what they are consenting to. The Privacy Commissioners' offices lack the enforcement power and budget for outreach that they need to inform tech companies about their responsibilities, and to deal with companies that skirt privacy law.

Some of the needed legislative tools to meet the technical challenge are ready to roll out in Part V.1 of PHIPA. There the duty to maintain EHRs, including privacy and security assurances, auditing responsibilities and reporting of breaches, is defined. This piece of legislation has yet to be proclaimed by the Lieutenant Governor.

ACTION POINT 9

Proclaim Part V.1 of PHIPA.

Informants were split on whether the OPC and IPC's powers to enforce privacy law act as a sufficient deterrent. The OPC recently published a set of recommendations including a renewed call for more oversight powers, the authority to make orders and a budget sufficient to keep up with demand for advisory services to businesses.[74] One area of concern is algorithmic decision-making tools, which are being considered or piloted for a number of uses beyond healthcare delivery, including policing, judicial decision-making, insurance rates, hiring, university admissions and government operations. Public disclosure of these algorithms in any detail would be unfair or unsafe in many instances, but because of the potential for human rights abuses, there needs to be some oversight mechanism. The OPC has been asking for the right to investigate in this area, even where no evidence of wrongdoing exists.

ACTION POINT 10

Grant oversight powers to the OPC, and increase the OPC budget to ensure adequate ability to investigate legislative breaches.

Privacy is not the only ethical concern. Preventing harmful uses of data is of greater concern, regardless of whether the harm involves a privacy breach. The first line of defence against unethical research is REB approval. There is a clear mandate for reforming REBs, but no consensus on what action to take to make sure their decisions are consistent, and that their membership is appropriately chosen and trained. Moving to a system of external REBs staffed by paid members is one option, however, there is concern that that could result in less specific expertise in both the scientific questions under review, and the community concerns at stake.

The Tri-Council is the authority best positioned to mandate changes to REB procedures, and it should take the lead in designing reforms. Closely related is the need for independent REBs to regulate the private sector. There is a developing trend of companies (Facebook, e.g.) setting up their own internal REBs to oversee research plans, but without an independent mandate, these lack credibility. There are no standards for the composition, training or operating procedures required of REBs. Such standards are needed.

ACTION POINT 11

Develop Tri-Council standards for REB composition, training and operating procedures.

74 Office of the Privacy Commissioner of Canada (2018). 2017-18 *Annual Report to Parliament on the Personal Information Protection and Electronic Documents Act and the Privacy Act*. https://www.priv.gc.ca/en/opc-actions-and-decisions/ar_index/201718/ar_201718/.

The OPC's recent guidance documents on obtaining meaningful consent and inappropriate data practices lay out best practices that tech companies ought to follow, but it remains to be seen whether these documents will be treated as binding.

Another line of defence against unethical research is the data sharing agreements that researchers enter into when getting access to data. Researchers affiliated with institutions like universities and hospitals are bound by workplace regulations and funding constraints to follow those agreements. However, one worry is that individual researchers and smaller institutions may not have the legal expertise to compose agreements that adequately protect patients. Guidance from the IPC on how to formulate a comprehensive data sharing agreement would be helpful.

ACTION POINT 12

Provide IPC guidance on composing customizable "off-the-shelf" data sharing agreements that adequately protect patients for smaller institutions and individual researchers.

Another source of worry is agreements that release health data to the private sector, where the main enforcement mechanism available is through the courts. The OPC's recent guidance documents on obtaining meaningful consent[75] and inappropriate data practices[76] lay out best practices that tech companies ought to follow, but it remains to be seen whether these documents will be treated as binding. Granting stronger enforcement and oversight powers to the OPC would help here too. Stronger legislation governing private sector health data may also be called for.

ACTION POINT 13

Enforce new consent and inappropriate data practice guidelines by OPC to stem inappropriate use and collection of health information by private sector firms.

The OPC's 2017-2018 annual report recommends that "consent should continue to play a prominent role" in privacy protection, but note that AI poses a challenge to consent, which may require other forms of privacy protection.[77] The model we advocate moving toward is a more closed data governance system, where there is minimal collection of health data by the private sector, and less *ad hoc* sharing of health data by healthcare providers and researchers. Instead the focus of innovation should be on centrally held datasets that can be better controlled not just for ethical concerns, but also in terms of quality. It is fully in the interests of players who are looking to

75 Office of the Privacy Commissioner of Canada (2018b). *Guidelines for obtaining meaningful consent*. https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gl_omc_201805/.

76 Office of the Privacy Commissioner of Canada (2018c). *Guidance on inappropriate data practices: Interpretation and application of subsection 5*(3). https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gd_53_201805/.

77 Office of the Privacy Commissioner of Canada (2018). 2017-18 *Annual Report to Parliament on the Personal Information Protection and Electronic Documents Act and the Privacy Act*. https://www.priv.gc.ca/en/opc-actions-and-decisions/ar_index/201718/ar_201718/.

innovate to have access to the best quality data. If adequate protections are in place, there is no reason why private interests should not be allowed paid access to detailed analyses of comprehensive, high-quality health datasets for legitimate purposes.

Models for closed data governance systems exist in the OCAP principles, disability data co-ops and PHIPA-defined prescribed persons and prescribed entities. Each of these models demonstrates ways of pooling health data from many individuals in order to support research for the common good, without needing to perform detailed consent negotiations on the individual level. The OCAP principles require community consent, not individual consent, for research use and sharing of data from First Nations communities. Data co-ops likewise are run by community representatives who decide how their pooled data can be used. The reason these systems exist is to encourage research that might help members of the community while exerting enough control such that community interests can be protected. There are many other communities that might likewise benefit from this simultaneous increased availability of data and stronger protection against harmful uses and privacy breaches.

Health information custodians across the province disclose personal health information to prescribed persons and entities like IC/ES and Cancer Care Ontario. Prescribed persons and entities are allowed to hold and use this personal health information, with some limitations, without getting informed consent from the data subjects. However, in order to maintain status as a prescribed person or entity, they are subject to strict security, data management and ethical standards, and compliance is regularly audited.[78] Staff sign confidentiality agreements and get ongoing training. Privacy impact assessments, data linkage protocols, inventories of data holdings, access control procedures, secure data transfer and security and privacy audits are everyday concerns.[79] The privacy impact assessments that IC/ES perform include participation by members of the communities that will be affected by proposed research, similar to community approval under OCAP.

This combination of linkable data (personally identifying information retained) with high security and ethical standards makes prescribed persons and prescribed entities ideal sites for AI research in health. The Vector Institute is already developing a partnership with IC/ES to take advantage of their data management and security infrastructure. More AI researchers should be made aware of this way of accessing health data. The prescribed entity model should be expanded significantly to accommodate additional research purposes, with corresponding budget increases.

ACTION POINT 14

Expand the data holdings of prescribed persons and prescribed entities, and ensure adequate budgets.

ACTION POINT 15

Encourage AI researchers to access health data through prescribed persons or prescribed entities.

78 Office of the Information and Privacy Commissioner of Ontario (2010). Manual for the Review and Approval of Prescribed Persons and Prescribed Entities. https://www.ipc.on.ca/wp-content/uploads/2017/07/process.pdf.

79 Cavoukian, Ann (2005). Review of the Pediatric Oncology Group of Ontario:A Prescribed Entity under the Personal Health Information Protection Act. https://www.ipc.on.ca/wp-content/uploads/2017/07/ent-pogo.pdf.

Some vulnerable groups may also prefer to set up their own independent data trusts, and grants should be made available for that purpose. Since outliers in the population cannot be sufficiently anonymized without prohibitive loses of data, it is reasonable for vulnerable populations to maintain higher levels of control over their data. This need not be done at the individual level, but could be achieved through community consent. The details of what should count as a community, how representatives ought to be chosen, and how to connect data trusts to other health data holdings will require further thought. Research into best practices for data trusts should be commissioned.

ACTION POINT 16

Offer financial support to set up community-governed data trusts for vulnerable populations, and for research into best practices.

Despite Ontario's strong privacy legislation, health information custodians are not always trusted by vulnerable groups. For example, Ontario Public Health raised the ire of HIV activists in 2017 when they began sharing viral load measurements with Public Health Units, contrary to privacy regulations (a practice which they quickly suspended after complaints).[80] Where independent data trusts are not feasible, prescribed entities have demonstrated trustworthy practices for the protection of community interests, as in IC/ES's partnership with FNIGC. Community governance is one way to build the trust that makes research possible, so community involvement in prescribed persons and entities ought to be strengthened. One informant cited the Ontario HIV Treatment Network Cohort Study[81] as an example where a community got on board with a study that links highly stigmatized health data across the province.[82]

ACTION POINT 17

Strengthen community engagement requirements for prescribed persons and prescribed entities.

80 Interview, 2018.
81 Ontario HIV Treatment Network (no date). *OHTN Cohort Study (OCS)*.Toronto: Ontario HIV Treatment Network. http://www.ohtn.on.ca/ohtn-cohort-study/.
82 Interview, 2018.

We are at a moment when informed consent and basic privacy protections are beginning to slip away. There is an opportunity for Ontario to take a leadership position in getting the balance right, and demonstrating the economic benefits of a secure, ethical system of data governance.

4 CONCLUSION

Many of the recommendations we make here echo previous recommendations by the OPC,[83] Canada Health Infoway,[84] the Canadian Institute for Health Information,[85] among others.[86] The new contribution AI makes to the conversation is an urgency to finding the political will to action. AI innovation demands data of a quality that cannot be acquired through anything other than public collection. As one AI researcher said, "the biggest bang for the buck is improving data collection processes and processing of primary datasets… input level intervention is crucial."[87] The possibility of harm is also much greater when large datasets are linked and analyzed using powerful algorithms. The decisions of AI systems often lack transparency, so there is little hope of addressing inequities after the fact.

We are at a moment when informed consent and basic privacy protections are beginning to slip away. Building EHR systems that take privacy and consent seriously must happen now, before unsustainable systems are put in place. Once private health data is leaked, it can never be made private again. Some young people today already believe that there's no point in protecting their privacy, because the information is already out there. It is not yet too late, but that point is rapidly approaching. There is an opportunity for Ontario to take a leadership position in getting the balance right, and demonstrating the economic benefits of a secure, ethical system of data governance.

The digital revolution in healthcare is coming whether we are prepared for it or not. All key stakeholders want a system where data is accessible for legitimate purposes, so that innovation can happen unhindered. But these stakeholders also recognize that open data is too risky; protections need to be put in place to ensure that harmful research does not make it through our data governance structures. Privacy leaks are only one of the concerns, and should not be the main focus of efforts to protect health data. Security is a serious concern when it involves sensitive data, which needs better protection. But making sure that research is performed in the best interests of patients is a broader and more pressing concern. In some situations, informed consent is the most appropriate method, and in these cases it needs to be insisted on starting from the design stages. In other contexts, community oversight can be

83 Office of the Privacy Commissioner, 2018.
84 Canada Health Infoway (2017). *Secondary Use Governance Across Canada: Common Understandings of the Pan-Canadian Health Information Privacy Group*. https://www.infoway-inforoute.ca/en/component/edocman/3356-secondary-use-governance-across-canada-common-understandings/.
85 Canadian Institute for Health Information (2013). *Better Information for Improved Health: A Vision for Health System Use of Data in Canada*. Canada Health Infoway. https://www.cihi.ca/en/hsu_vision_report_en.pdf.
86 Willison, Donald J. (2009). *Use of Data from the Electronic Health Record for Health Research: Current Governance Challenges and Potential Approaches.* https://www.ijpc-se.org/documents/hhs14.pdf.
87 Interview, 2018.

an appropriate way of ensuring that research is not harmful. Independent bodies like data trusts and prescribed entities can do that work.

What innovation hangs on in the end is trust. When they lose trust, patients lie to their doctors or stop accessing healthcare entirely. To ensure that does not happen, we need to build trustworthy processes into the new digital systems we build.

Section 1 outlined the innovations that are on their way in storage of health data, in the technology industry, in AI applications to medical research, and in data governance. Section 2 explored solutions for responsible innovation in two categories: technological challenges, and ethical/legal challenges. The list that follows summarizes the recommendations gathered through interviews with stakeholders and background research, and which were presented in Section 3. It is our hope that these recommendations will spur significant action on these issues in the near term. Our health is simply too important to risk waiting any longer before making these critical policy changes.

ACTION POINT SUMMARY

1. Implement a pan-Canadian, interoperable EMR/EHR standard.
2. Provide incentives for providers to adopt the standard EMR, and for 3rd party EMR vendors to make existing systems compliant.
3. Include integrated consent directives, automated privacy audits, easy-to-use levels of authentication, and built-in de-identification tools in the pan-Canadian EMR/EHR system.
4. Design infrastructure for secure transfer of EHRs between providers and central bodies like prescribed entities.
5. Create protected status for sensitive information, either through legislation or an interpretation document put out by Privacy Commissioners' Offices.
7. Offer operating grants for community-based digital literacy programs to expand their offerings to reach a wider audience.
8. Commission a study to research the effectiveness of ethics education in computer science and engineering programs, and to develop evidence-based curricula.
9. Proclaim Part V.1 of PHIPA.
10. Grant oversight powers to the OPC, and increase the OPC budget to ensure adequate ability to investigate legislative breaches.
11. Develop Tri-Council standards for REB composition, training and operating procedures.
12. Provide IPC guidance on composing customizable "off-the-shelf" data sharing agreements that adequately protect patients for smaller institutions and individual researchers.
13. Enforce new consent and inappropriate data practice guidelines by OPC to stem inappropriate use and collection of health information by private sector firms.
14. Expand the data holdings of prescribed persons and prescribed entities, and ensure adequate budgets.
15. Encourage AI researchers to access health data through prescribed persons or prescribed entities.
16. Offer financial support to set up community-governed data trusts for vulnerable populations, and for research into best practices.
17. Strengthen community engagement requirements for prescribed persons and prescribed entities.