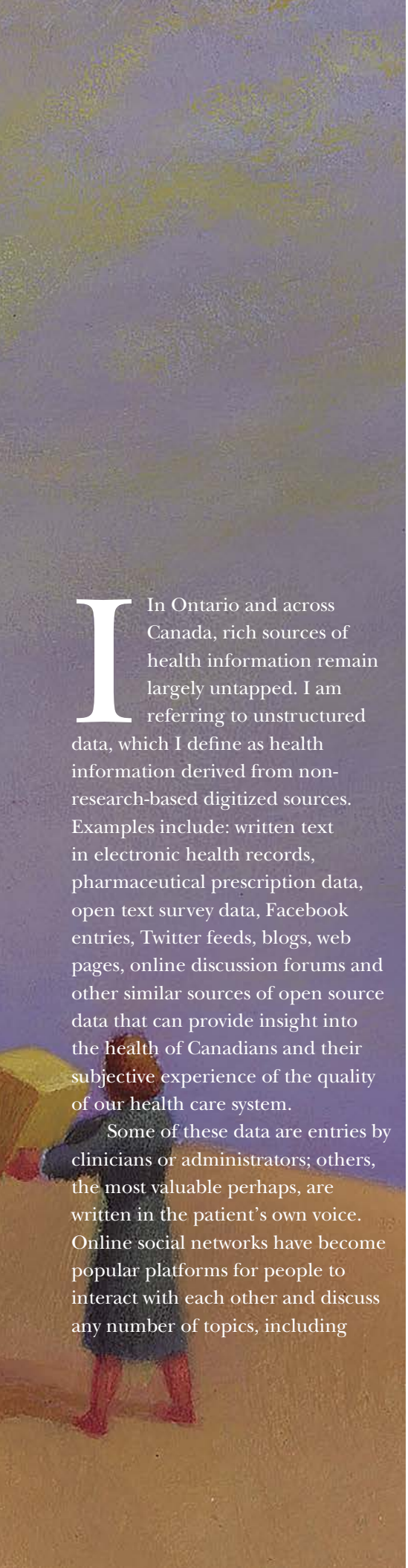
Mining Unstructured Data to Improve Health Care Transitions

By: Neil Seeman



Is there evidence that mining this vast array of unstructured data can positively influence health care? The short answer is "yes".

In Ontario and across Canada, rich sources of health information remain largely untapped. I am referring to unstructured data, which I define as health information derived from non-research-based digitized sources. Examples include: written text in electronic health records, pharmaceutical prescription data, open text survey data, Facebook entries, Twitter feeds, blogs, web pages, online discussion forums and other similar sources of open source data that can provide insight into the health of Canadians and their subjective experience of the quality of our health care system.

Some of these data are entries by clinicians or administrators; others, the most valuable perhaps, are written in the patient's own voice. Online social networks have become popular platforms for people to interact with each other and discuss any number of topics, including health and health system concerns. In this paper, I propose leveraging the use of unstructured data to better understand patients' and caregivers' experiences with transitions in care.

These unstructured data share three characteristics: volume (they are plentiful), velocity (they become available quickly but may also disappear quickly), and variety (they exist in many forms). They often lack a fourth "v," which is veracity – in other words, they are not necessarily accurate or 'evidence-based' (Martin-Sanchez & Verspoor, 2014; Normandeau, 2013). While validity may therefore be in doubt, these data unquestionably complement the more traditional sources of health information and patient experience (e.g., results of clinical trials, case reports, focus group or paper-based patient experience data).

All the traditional methods of gathering information reach their conclusions by studying relatively limited numbers of patients so that generalizability to a larger range of potential patients remains questionable. As the world dramatically shifts from paper to electronic formats, digital data offer a much larger evidence base. Leveraging such data to improve health care quality, however, poses three challenges.

The first challenge is that the data exists in free-form text or, at best, semi-structured clinical documentation, dispersed among different providers, and in different social media. They are, thus, difficult to standardize, analyze, and aggregate. The second challenge is to glean worthwhile insights from this mass of texts even when it is organized, insight that can improve the quality of care, and, specifically, transitions of care.

In this article, I offer a path forward toward solving these two challenges. I suggest that we bring the Ontario health care start-up and institutional communities together to 'hack our way forward', as I explain in the conclusion. This requires overcoming the third challenge, which is to ensure privacy (Kum & Ahalt, 2013). We are talking about data that should enjoy the same legislated privacy protections as do, for example, patient health records. Unstructured data are often tagged with personal identifiers, and therefore, are fraught with privacy challenges. This is a thorny problem and constitutes one of the critical barriers to the full use of these rich data sources.

Return-on-Investment? Show Me the Evidence

Is there evidence that mining this vast array of unstructured data can positively influence health care? The short answer is "yes".

One example is epidemic surveillance. In 2012, Young and colleagues collected more than half a billion tweets and found almost 10,000 that mentioned sexual behaviours and drug use which are risk factors for HIV infection. By placing the geographic origin of these tweets onto a map of the world, potential hot spots for HIV could be discerned, showing how social media can be used to monitor the spread of infectious disease (Young, Rivers & Lewis, 2014). In the same vein, geographic comparisons can be made to assess within which regions patients experience transitions across the continuum of care that are coherent and linked, or experiences of overt or subtle discrimination during such transitions.

As evident in this example, security and privacy risks need to be addressed: even if the data are anonymized and are not traceable to individuals, there should be no way for anyone to be able to know whether particular, identifiable sub-populations in a community are disproportionately at risk for HIV. It is critical to prevent bringing attention, stigma, and discrimination to a population already at risk.

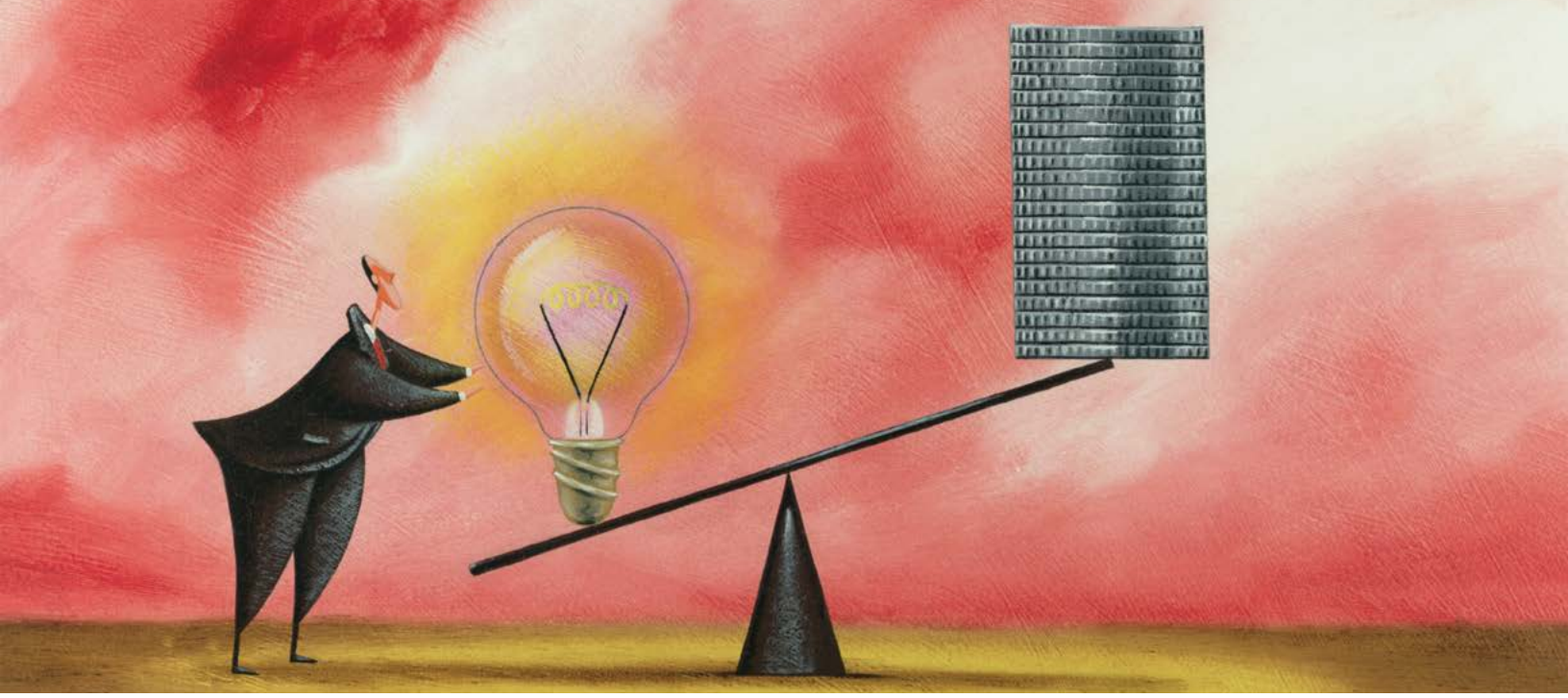
In my view, transitions in care are places where the use of unstructured data can be especially useful, especially since very few tools exist that comprehensively measure patient experience across transitions. Despite this, transitions are crucial points in health care delivery which can impact patient outcomes and the quality of care provided. Transitions are difficult for patients no matter what their illness – first, the transition from well to ill, then the various stages of navigating the health system: accessing help, waiting for referrals, admission to hospital, discharge to home or to a rehabilitation service, home care, or transitioning across the paediatric/adult/geriatric frontiers. These all constitute periods of stress and unmet need often expressively voiced in Facebook entries or in personal blogs that, once organized and analyzed, and safely (i.e., personal identity removed) disseminated, have the potential to provide valuable information that might help to improve the experience of patients as they transition across the continuum of care.

Figuring out how to best leverage unstructured data for any return-on-investment is neither a human resource challenge nor a technology challenge, but a strategic one. Essentially, the issue boils down to improving the quality and subjective experience of transitions in health care within Ontario while preserving the privacy of patient data and guaranteeing their secure storage.

Figuring out how to best leverage unstructured data for any return-on-investment is neither a human resource challenge nor a technology challenge, but a strategic one.

Meeting the First Challenge: Finding the Relevant Data

Since unstructured data exist in so many different forms, the first challenge is to find in a very large haystack the needles of data that are germane to improving transitions in health care. This first step involves drawing on expert stakeholders' input to define more precisely what the problem is we are trying to solve. Are we trying to better understand the experience of transitions of care from the perspective of the patient? Are we trying to fill the gap of unmet needs of patients, families, providers,



or all of these groups? Are we trying to minimize costs, to reduce lawsuits? Is our aim to ensure that patients and families continue to be treated with respect and with skill during care transitions? Is this an attempt at quality improvement?

Depending on the answers to these questions, different forms of unstructured text can be targeted. For example, improving the experience of transitions may involve the use of open text narratives on Twitter to assess the degree to which patients or caregivers express increasing or decreasing frustration online about access points, or parking lot capacity, wheelchair availability, or failures by care providers to transmit important clinical information.

Alternatively, if the goal of the initiative, as defined by expert stakeholders, is to identify the best practices of those most skilled at transitions of care, we may wish to engage in natural language processing of open text comments and online reports that refer to a range of providers and link the most publicly favored among them to clusters of health care facilities in a certain geographic region. We can ensure privacy by excluding identifying remarks and aggregating the data. We can then dig deeper to discover if the preferred provider group shares critical characteristics. Have they adopted a particular model of care that focuses on improving care transitions? Were they trained under the same clinical mentors? What other factors distinguish them?

Relevance to a chosen topic should involve an expert panel consisting not only of patients, caregivers and providers, but also of information technologists who can provide guidance as to methods of extraction by searching for key words, creating a coding scheme, systematically applying the codes to the selected documents, testing for inter-coder reliability, counting the numbers of codes for each document, and analyzing the counts by a variety of statistical methods (Bernard, 2012).

Natural language processing (NLP), which means using heavy computing power to make sense of the semantic or representational meanings hidden in extraordinarily large data sets, is then able to extract relevant transition themes (such as communication across settings and among disciplines and specialties, family input into decisions, cross site co-ordination, the inter-disciplinary nature of care, access to required medications in an uninterrupted way, availability of appropriate transport between settings, anticipation of needs over the evolution of an illness, the efficacy of interoperable electronic health records, maintenance of a safety net throughout life, or education in patient self-management).

This has not yet been done specifically to search for subjective experiences across care transitions, but it has been done to better understand reactions to the general category to which transitions belong, the category of unexpected stressful events (Gaspar, Pedro, Panagistopoulos & Seibt, 2016).

Traditionally, care transitions have been studied through personal interviews or focus groups, mainly of professionals (Davis, Devoe, Kansagara, Nicolaidis & Englander, 2012; Gott, Ingleton, Bennett & Gardines, 2011; Greysen et al., 2014),

although patient voices have also been elicited (Pollack et al., 2016). In order to attempt to identify solutions to care transition problems, Lim, Jarvenpaa and Lanham (2015) reviewed published qualitative studies and came to the conclusion that time pressures were at the heart of many of the reported difficulties. While, as stated above, transitions have not yet been specifically researched on social media, public health questions on a range of topics are beginning to be profitably addressed by mining this rich source of data (Ji, Chum, Wei & Geller, 2015).

Meeting the Second Challenge: Elegant Organization and Operationalization

Organizing the wealth of unstructured data that can be obtained from various sources remains a major challenge. Advances in computer science, such as virtualization and cloud computing have made it possible to accommodate, link, and analyze large, diverse data sets (Raghupathi & Raghupathi, 2014). This requires a staged process by which the information from various sources initially needs to be pooled and distributed on open source platforms available in the cloud. Various statistical methods are used: cluster analysis, decision-tree learning, Bayesian networks, NLP, graph analytics, and other data visualization approaches. The best guidance is to learn from search engine leaders, notably Google, whose leadership in Bayesian algorithms and PageRank analysis has pioneered the modern Internet to guide us in identifying what information is trusted, and what information is not.

Solving this is beyond the scope of any one health care institution, and, as such, I recommend first using free tools made available by Google itself. Public datasets can be analyzed for free in the Google Cloud platform (https://cloud.google.com/public-datasets/). Boolean search strings plugged into Google search, e.g., 'free data sets' AND genes, or 'free data sets' AND 'Ontario' AND 'health care', are a good start. After the data sets are found, a novice unstructured data researcher can insert them into the Google platform. One first searches for correlations and emerging patterns. Predictions can be made from repeating patterns.

The challenge of establishing the veracity of the information requires technology and systematic analysis by experienced researchers. A large variety of methodologies are available for data integration and resolution of inconsistency among data sets (Gray & Thorpe, 2015; Neff, 2013).

Unstructured Data in the Context of Broader Data Needs

In order to improve standards for care transitions and increase the quality of care generally, there is another, higher level of analysis – informed by principles of equity and prioritization of need – that points out questions that require the most urgent attention from a data gaps perspective (Golden, Hager, Gould, Mathioudakis & Pronovost, 2017).

We need to understand the human impact of care transitions on patients, families and social networks; we need to know their economic impact on the health care system in terms of job satisfaction, health service costs, pharmaceutical costs, and costs to the legal system. We need to be able

We need to understand the human impact of care transitions on patients, families and social networks; we need to know their economic impact on the health care system in terms of job satisfaction, health service costs, pharmaceutical costs, and costs to the legal system.

to identify best practices (Naylor et al., 2017), to determine the role of family physicians, the distinctive role of the many health professionals involved in care transitions, the role, if any, of paraprofessionals and alternative care providers. We also need to be aware of obstacles to smooth transitions – currently thought to be time pressures, faulty communication, fragmentation of services and inadequate staff training.

One way to evaluate what we do in Ontario is to analyze data from around the globe as they pertain to transitions – compare successes and failures and costs, and specific regional challenges to determine how we might be able to do things better. Unstructured data lend themselves to such comparisons.

Recent Progress

Seven years ago, on behalf of The Change Foundation, I helped to write a report titled, *Using Social Media to Improve Health Care Quality*, in which we identified the fact that the vast majority of North American health care organizations were not taking advantage of social media data (i.e., the most commonplace form of open source unstructured data), and thus were neglecting a source of patient-driven information (e.g., complaints, expressions of unmet needs) that could, if accessed, potentially lead to insights for quality improvement (The Change Foundation, 2011).

Some progress has been made since then, as described above, but probably not enough. Privacy issues have remained barriers. How much we can rely on the validity of unstructured data is an unanswered question. Numerous bodies, including institution-specific research ethics bodies, are investigating both of these challenges.

Since global public health (notably in the field of pandemic surveillance) has seen some of the earliest usages of unstructured data on a global scale, it is not surprising that organizations such as the Global Public Health Intelligence Network (GPHIN) and the Public Health Agency of Canada (PHAC) are setting forth ethical frameworks and publishing thought leadership to address evolving paradigms in the ethics and privacy needs associated with the use of unstructured data for health care quality generally (Public Health Agency of Canada, 2015).

The broader framework of 'Privacy by Design' (PbD) developed over a decade ago by the Information and Privacy Commissioner of Ontario should, in the author's opinion, guide all application discussions (Cavoukian, 2009). The PbD principle of application "with special vigour to sensitive data such as medical information" is important; however, in the author's view, there is a critical amplification that should be stressed in the context of the unstructured data referred to in this article. That is, in the context of the application of unstructured data to solving transitions-related policy issues, PbD should, in the author's view, be applied with special vigour to sensitive data that is prone to misinformation.

In other words, the challenge of veracity in the use of unstructured data, referred to earlier, is inseparable from the challenge of data privacy. Privacy and security are insufficient; the data have to be accurate.

Hacking a Path Forward

I recommend an Ontario-wide sponsorship of a 'hackathon' – a community of individuals bound by a common and novel mission, in this case, refining the obstacles to using unstructured data to promote aspects of health care quality. Already, the Ontario start-up community has a vested interest and desire to participate. Ontario enjoys a diverse and passionate start-up community developing mobile health apps, and advancing health care data visualization and data mining. Many of these start-ups have faced challenges monetizing their platforms and understanding the needs of potential customers and partners. To address all such problems, it would be strategic to bring together the people most passionate and motivated to solve them; solutions could then be shared. This is a low-cost first start to what may be a very high return on investment for the Ontario health care system.

Many research teams are working on mining unstructured data, with varied success. This is a growing trend in other industries, and health care would stand to benefit greatly from this rich source of data which can present new opportunities for improving care perhaps, never before considered. Doing it well can lead to increased knowledge, decreased obstacles to the receipt of high-quality care, and lowered costs, both human and economic.

Individual practitioners, clinics, and hospitals need to heed what is being said about them and their personnel on online service rating sites. These anonymous criticisms, when taken seriously, can lead to local quality improvement without necessitating large-scale data mining. On a larger scale, only an expert group, sensitive to overarching priorities in the sector of interest and their own institutional priorities, can begin to identify the strategic demands most in need of unstructured data. Along this journey, technologists who are expert in NLP and other machine learning processes can lend guidance to the art of the possible, but should not, of course, be permitted to drive the setting of priorities.

Neil Seeman is Founder, Chairman and CEO of RIWI Corp.; a Senior Fellow of Massey College; Adjunct Lecturer in the Institute of Health Policy, Management and Evaluation, University of Toronto and Senior Advisor to New York-based cyber-law firm Blackstone Law Group LLP.

About RIWI Corp.

RIWI Corp. (www.riwi.com) is a global survey, global messaging and global prediction science firm that has collected opinion data, machine data and unstructured data from more than 1.2 billion web users in every country and territory of the world for organizations such as the United Nations World Food Programme, the World Bank, the Bill and Melinda Gates Foundation, the U.S. State Department, Freedom House, and governments and corporations around the world.

References

Bernard, H.R. (2012). *Social Research Methods: Qualitative and quantitative Approaches.*University of Florida, USA: Sage Publications, Inc.

Cavoukian, A. (2009). Privacy by Design: The 7 Foundational Principles. Information and Privacy Commissioner of Ontario. Retrieved from https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

Davis, M.M., Devoe, M., Kansagara, D., Nicolaidis, C. & Englander, H. (2012). Professional views on hospital to home care transitions. *J. Gen. Intern. Med.* (27), 1649–56.

Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers Hum Behavior* (56), 179e191

Golden, S.H., Hager, D., Gould, L.J., Mathioudakis & N., Pronovost, P.J. (2017). A gap analysis needs assessment tool to drive a care delivery and research agenda for integration of care and sharing of best practices across a health system. *Jt. Comm. Qual. Patient Saf.* (43), 18-28.

Gott, M., Ingleton, C., Bennett, M. & Gardiner, C. (2011). Transitions to palliative care in acute hospitals in England: qualitative study. *Br. Med. J.* (342), d1773.

Gray, E.A. & Thorpe, J.H. (2015). Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. *J. Comp. Eff. Res.* (4), 61–74.

Greysen, S.R., Hoi-Cheung D., Garcia V., Kessell E., Sarkar, U., Goldman, L. & Kushel, M. (2014). "Missing Pieces" – Functional, social, and environmental barriers to recovery for vulnerable older adults transitioning from hospital to home. *J. Am. Geriatr. Soc.* (62), 1556–61.

Ji, X., Chun, S.A., Wei, Z. & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Soc. Netw. Anal. Min.* (5), 13.

Kum, H.C. & Ahalt, S. (2013). Privacy-by-design: Understanding data access models for secondary data. *AMIA Summits Transl. Sci. Proc.* 2013: 126-30.

Lim, S.Y., Jarvenpaa. S.L. & Lanham, H.J. (2015). Barriers to interorganizational knowledge transfer in post-hospital care transitions: review and directions for information systems research. *J. Manag. Inform. Syst.* (32), 48-74.

Martin-Sanchez, F. & Verspoor K. (2014). Big data in medicine is driving big changes. *Yearb. Med. Inform.* 14-20.

Naylor, M.D., Shaid, E.C., Carpenter, D., Gass, B., Levine, C., Li, J., Malley, A & Brock, J.(2017). Components of comprehensive and effective transitional care. *J. Am. Geriatr. Soc.* (65), 1119-25.

Neff, G. (2013). Why big data won't cure us. *Big Data* (1), 117–123.

Normandeau, K. (2013). Beyond volume, variety and velocity is the issue of big data veracity. Inside Big Data. Retrieved from https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

Pollack, A.H., Backonja, U., Miller, A.D., Mishra, S.R., Khelifi, M., Kendall, L. & Pratt, W. (2016). Closing the Gap: Supporting patients' transition to self-management after hospitalization. *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.* 2016, 5324-36.

Public Health Agency of Canada. (2015). Big data and ethics. *Canadian Communicable Disease Report,* 41(9), 219. Retrieved from http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/15vol41/dr-rm41-09/assets/pdf/15vol41_09-eng.pdf

Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Inf. Sci. Syst.* (2), 3.

The Change Foundation (2011). *Using Social Media to Improve Health Care Quality. Change Foundation.* Retrieved from http://www.changefoundation.ca/social-media-healthcare-pt1

Young, S.D., Rivers, C. & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev. Med.* (63),112–5.