Envisioning a Data Science Strategy for ICES

September 2017

Envisioning a Data Science Strategy for ICES

Authors

Thérèse A. Stukel
Peter C. Austin
Mahmoud Azimaee
Susan E. Bronskill
Astrid Guttmann
J. Michael Paterson
Michael J. Schull
Rinku Sutradhar
J. Charles Victor

September 2017

Institute for Clinical Evaluative Sciences

Publication Information

INSTITUTE FOR CLINICAL EVALUATIVE SCIENCES
G1 06, 2075 Bayview Avenue
Toronto, ON M4N 3M5
Telephone: 416-480-4055
www.ices.on.ca

How to cite this publication

Stukel TA, Austin PC, Azimaee M, Bronskill SE, Guttmann A, Paterson JM, Schull MJ, Sutradhar R, Victor JC. *Envisioning a Data Science Strategy* for ICES. Toronto, ON: Institute for Clinical Evaluative Sciences; 2017.

ISBN: 978-1-926850-77-1 (Online)

Authors' Affiliations

Thérèse A. Stukel, PhD
Senior Scientist, Institute for Clinical Evaluative Sciences / Professor, Institute of Health Policy, Management and Evaluation, University of Toronto / Professor, Dartmouth Institute for Health Policy and Clinical Practice

Peter C. Austin, PhD
Senior Scientist, Institute for Clinical Evaluative Sciences / Professor, Institute of Health Policy, Management and Evaluation, University of Toronto / Senior Scientist, Sunnybrook Research Institute

Mahmoud Azimaee, BSc, PStat
Director, Data Quality and Information Management, Institute for Clinical Evaluative Sciences

Susan E. Bronskill, PhD
Senior Scientist and Program Lead, Health System Planning and Evaluation, Institute for Clinical Evaluative Sciences / Associate Professor, Institute of Health Policy, Management and Evaluation, University of Toronto / Associate Scientist, Sunnybrook Research Institute / Adjunct Scientist, Women's College Hospital

Astrid Guttmann, MDCM, MSc, FRCP(C)
Chief Science Officer and Senior *Scientist*, Institute for Clinical Evaluative Sciences / Staff Paediatrician, Division of Paediatric Medicine, Hospital for Sick Children / Professor, Institute of Health Policy, Management and Evaluation, University of Toronto

J. Michael Paterson, MSc
Scientist and Program Lead, Chronic Disease and Pharmacotherapy, Institute for Clinical Evaluative Sciences / Assistant Professor, Department of Family Medicine, McMaster University / Assistant Professor, Institute of Health Policy, Management and Evaluation, University of Toronto

Michael J. Schull, MD, MSc, FRCP(C)
President and Chief Executive Officer and Senior Scientist, Institute for Clinical Evaluative Sciences / Professor, Department of Medicine and Institute of Health Policy, Management and Evaluation, University of Toronto / Senior Scientist, Sunnybrook Research Institute

Rinku Sutradhar, PhD
Senior Scientist, Institute for Clinical Evaluative Sciences / Associate Professor, Dalla Lana School of Public Health and Institute of Health Policy, Management and Evaluation, University of Toronto / Affiliate Scientist, Sunnybrook Research Institute

J. Charles Victor, MSc, PStat
Senior Director, Research and Data, Institute for Clinical Evaluative Sciences

Acknowledgements

We are grateful to members of the Scientific Advisory Board at ICES for their support, particularly Professor David Ford for his extensive review of an early draft of this report.

We also thank Ashif Kachra for capturing and organizing our freewheeling discussions into comprehensive meeting notes.

About the Institute for Clinical Evaluative Sciences

Established in 1992, the **Institute for Clinical Evaluative Sciences** (ICES) is an independent not-for-profit corporation with an international reputation as a trusted source of high-quality health and health services research and evidence.

ICES researchers have access to a vast and secure array of Ontario's health-related data, including population-based health surveys, anonymous patient records, and clinical and administrative databases. ICES' unbiased evidence provides measures of health system performance, a clearer understanding of the shifting health care needs of Ontarians, and a stimulus for discussion of practical solutions to optimize scarce resources. ICES research and reports influence the development, implementation and evaluation of health policy and the delivery of health care.

Key to ICES' work is its ability to link population-based health information, at the patient level, in a way that ensures the privacy and confidentiality of personal health information. Linked databases reflecting 13 million of 34 million Canadians allow researchers to follow patient populations through diagnosis and treatment, and to evaluate outcomes. ICES goes to great lengths to protect privacy and is recognized as an international leader in maintaining the security of health information.

ICES receives core funding from the Ontario Ministry of Health and Long-Term Care. In addition, ICES scientists and staff have highly successful track records competing for peer-reviewed grants from federal agencies, such as the Canadian Institutes of Health Research, and from provincial and international funding bodies.

Foreword

In 2015, as part of a strategic planning cycle, ICES reviewed the research priorities for our institute through a consultation process that involved the entire ICES network, including scientists, staff, board members and external stakeholders. The ICES Scientific Advisory Committee also provided key input. As part of our review, we examined recent trends and innovations in health services research, such as evolution in the kinds of data being used, novel methodologies, and new approaches to distributed data analyses, among others.

A number of the innovations we considered fall under the umbrella of data science and apply to many of our existing research priorities. Whereas ICES has traditionally viewed these innovations as enablers of excellent research, our Scientific Advisory Committee asked us to think about how data science could become a priority in its own right at ICES. We accepted the challenge and have begun to develop a data science strategy.

While establishing data science as a priority is new to ICES, integrating novel data and developing new analytic methods are not. In the years following ICES' establishment in 1992, our researchers quickly moved beyond administrative data, augmenting it with population-level data, surveys, registries and trials, and most recently, unstructured data in electronic medical records. Sophisticated statistical methods, such as multistate modelling, and novel applications, such as the identification of virtual networks of physicians and hospitals, are being used in ICES research. ICES is a recognized leader in data linkage and is already involved in the development of international distributed data networks. In the near future, we expect to move forward with a partnership that will provide access to a high-performance computing environment for ICES researchers and create a new model for the secure storage, access, linkage and analysis of research data.

This report, prepared for the ICES Scientific Advisory Committee, marks the first stage in our efforts to develop data science as a priority at ICES. A working group, led by ICES senior scientist Dr. Thérèse Stukel, was struck in 2015 to define what data science is at ICES, identify relevant activities and expertise that exist at ICES, and propose and prioritize recommendations for advancing an ICES data science strategy.

Contents

Introduction

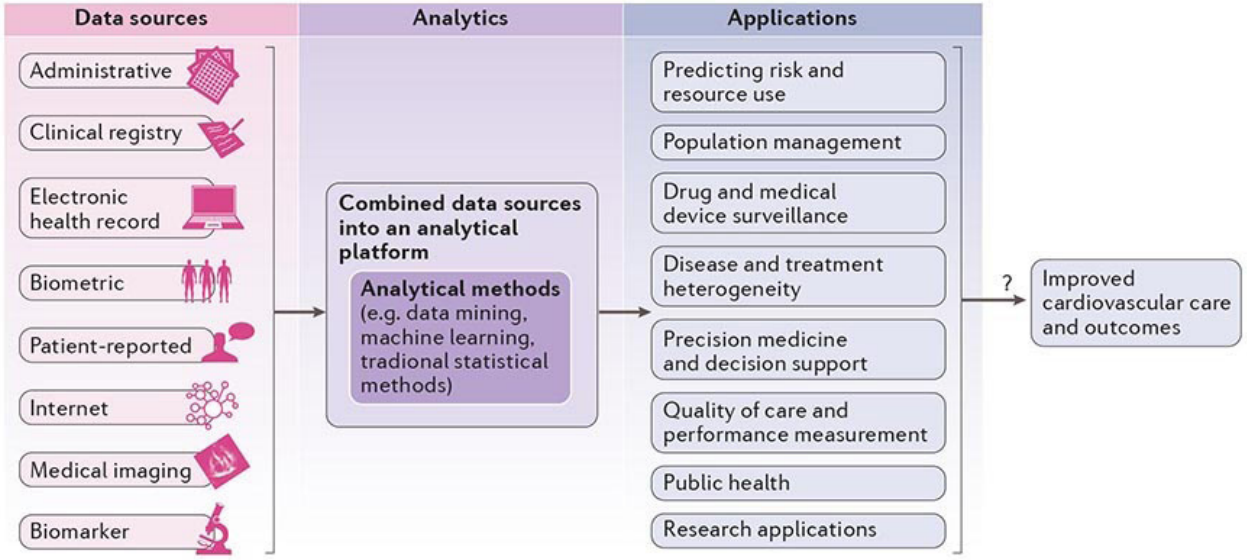
We are facing an explosion of health care data derived from novel sources, including electronic medical records; genomic, biomarker and imaging studies; social media and networked research resources; and patient-reported outcomes. These data, coupled with advances in analytic methodology and computational capability, are creating the promise for big data analytics to identify associations and make predictions that will improve health care quality and patient outcomes.[1,2] Big data pose tremendous opportunities and challenges, requiring us to develop new ways of acquiring, storing and sorting massive amounts of data. Novel statistical methods are needed to analyze very large databases in genetics, imaging and comparative effectiveness research. Researchers in biomedical informatics are developing technologies and software for generating, managing and interpreting biomedical data and knowledge. Biomedical informatics methods are used to merge big data from multiple sources and support the discovery of complex relationships between biomarkers and disease outcomes. There is a need for data scientists who are trained to use these methods, as advanced training in biomedical data science will propel a new era of discovery in this rapidly developing field.

ICES has the ability to link individual-level, health-related information from a variety of independent data sources such as health administrative data, clinical registries, population-based health surveys, and electronic medical records. The ICES data repository is widely seen as a leading model internationally in terms of scope, breadth and access. Administrative data in the repository are broad and contain the billing information for most of Ontario's publicly funded health services since 1991. The repository also includes some non-health administrative data related to citizenship and immigration, social services, Indigenous populations and the census. ICES holds routinely collected biomarker data such as BORN Ontario's prenatal and newborn screening data, genetic and other cancer biomarker data from chart abstraction, and linked electronic medical records profiling primary care received by more than 500,000 Ontarians. An overview of the types of new data, methods and applications is reported in **Exhibit 1.**[1]

Huge developments are occurring in big data analytics across the world. The results will likely

prove transformational. Distributed data research platforms have sprung up in many countries. In the United States, the National Institutes of Health (NIH) have launched the Big Data to Knowledge (BD2K) initiative as "a trans-NIH initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge, and to maximize community engagement."[3] The White House has launched the Precision Medicine Initiative Cohort Program (with funding of $215 million in 2016) to advance research, technology and policies that will enable researchers, providers and patients to work together to develop individualized care.[4] The National Science Foundation has launched a big-data funding opportunity to study foundational issues in data science.[5] The University of Michigan has launched a $100 million data science initiative with the Michigan Institute for Data Science that will serve as the focal point for the new multidisciplinary area of data science at the university.[6] These are among the many new initiatives being launched worldwide.

EXHIBIT I Overview of big data analytics and applications. Examples of the inputs (data sources) and outputs (analytical methods and applications) that can potentially improve cardiovascular quality and outcomes of care. [Adapted by permission from Macmillan Publishers Ltd: *Nature Reviews: Cardiology*. Rumsfeld et al. 13(6):350–9. Copyright 2016.]

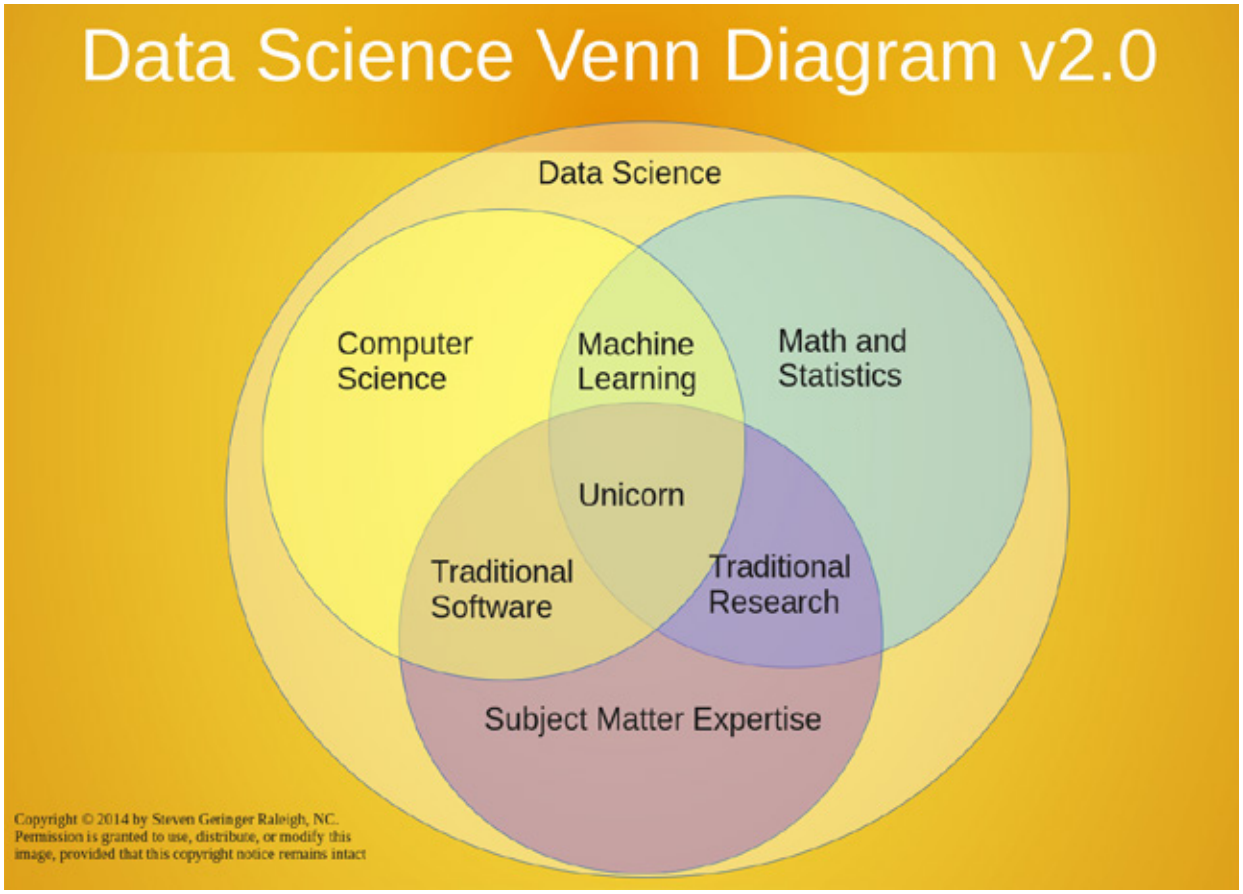


Data science is defined broadly as the science of extracting knowledge and new insights from data. The term was coined in 2001 by Cleveland in *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics* where he introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" and embracing a broader definition of statistics that includes computational and data analytic aspects, in addition to traditional theory and hypothesis testing.[7]

Data science is an interdisciplinary specialty concerned with the processes and systems used to extract knowledge or insights from data in various formats, structured or unstructured, and is an expansion of the traditional data analysis specialties of statistics and predictive analytics. The core elements include statistics, computer science, mathematics and software engineering. Sub-elements include data engineering (database management, data warehousing, data linkage, data transformation, data de-identification for privacy preservation); data visualization; computational methods; data analysis (statistical models, machine learning, predictive analytics, data mining, pattern recognition, signal processing, artificial intelligence); and computational infrastructure (high-performance computing, cloud computing). Data science includes advances in data modelling (inferential and predictive), linkages with "omics" and big data, and related analytic challenges such as methods to handle bias, threats to study reproducibility, unmeasured confounding, missing data and imputation. Data scientists must possess cross-disciplinary skills to extract knowledge and insights from large, complex data. A schematic of the skill sets needed to effectively do data science is shown in **Exhibit 2**.[8] The term "unicorn" in the centre of the diagram is in reference to recent discussions in the blogosphere that data scientists are as hard to find as unicorns.

Donoho wrote an insightful paper[9] on his vision of *greater data* science as a scientific expansion of the current notion of academic data science, the latter being an amalgamation of statistics and machine learning with some technology for scaling up to big data. Donoho defines greater data science as the science of learning from data through the scientific study of data analysis, essentially a new scientific discipline. We will explore the notion of greater data science and borrow heavily from Donoho's ideas as

EXHIBIT 2 Venn diagram depicting the skill sets required to do data science



they are relevant to ICES' goal of broadening its scope to include the science of learning from data.

Donoho traces the history of the scientific canons of data analysis over the last 50 years, from the development of the quantitative programming environment (QPE) to the re-emergence of machine learning and predictive modelling. A QPE allows analysts to *run scripts or workflows* (macros) that codify the steps of an analysis and can be shared and re-executed. The QPE most widely used by statisticians is the open-source software package R. Using a QPE, analysts can understand and critique the logic of another's analysis steps, and tweak the script to improve performance. The QPE has been a game changer in the science of data analysis.

Another re-emerging technology in data science is predictive modelling or machine learning, which, in contrast to classical statistical inference, aims to predict future outcomes and identify patterns in complex data without specifying the underlying mechanism generating the data or placing constraints on the variables used in modelling. The core of the machine-learning culture largely resides within computer science departments. Areas where machine learning has scored successes include natural language processing, text mining, machine translation and optical character recognition, mostly for commercial and marketing applications. The skills involved are primarily informed by information technology.

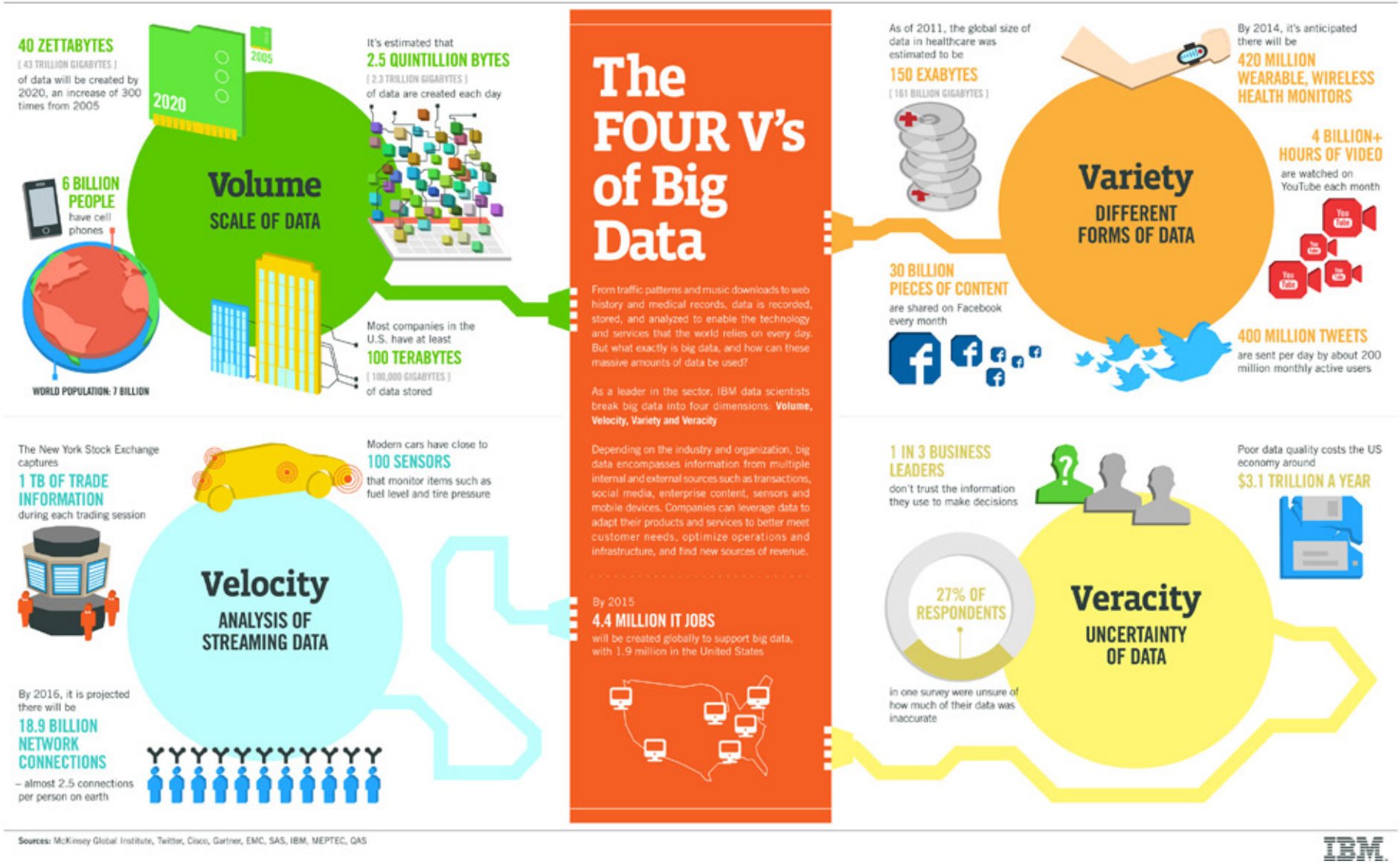
The core elements of greater data science are data exploration and preparation, data representation and transformation, computing with data, data visualization and data modelling. Because they are highly relevant to ICES research, we have generally adopted them as areas of data science expansion at ICES.

A sixth element, the *science of data science*, is a novel component of greater data science. It requires foundational work to create evidence-based data analysis. We mention it here, but it is beyond the scope of adoption by ICES. These new concepts include the science-wide meta-analysis or evaluation of all data analyses published on a specific topic by electronically scraping the scientific literature for clues about meta-problems as to how all science is analyzing data. This is related to Ioannidis' work on estimates of science-wide false discovery rates.[10] Other fertile areas are cross-study analyses, or meta-studies, of collections of data sets and prediction algorithms of the same outcome to understand the shortcomings of various approaches[11] and cross-workflow analyses or studies of the effects of different workflows on study conclusions using the same data sets. For example, Madigan has shown that methodological and analytic variability using the same data set but different workflows can arrive at opposing or contradictory findings about the risk of exposures in observational studies.[12,13]

The emergence of big data has been a catalyst for the new world of data science. Often the two terms are used interchangeably, but they are not synonymous. Big data refers to data sets that are so large or complex that traditional data processing and analysis methods are insufficient. Such data are now routinely gathered from mobile devices, electronic medical records, audio and video data, imaging data, and genetic data. The challenges of big data are often characterized as the four V's: volume, velocity, variety and veracity (see **Exhibit 3**).[14] *Volume* and *velocity* are technical challenges that are being met through sophisticated computer science technology on high-speed parallel processing platforms or high-performance computing networks. *Variety* refers to data that cannot be transformed into the rectangular (structured) format of rows and columns that are commonly used at ICES. *Veracity* refers to the representativeness of the data. Unstructured data arise primarily from text files and include electronic medical records and survey responses, as well as digital, video and audio files and medical, financial, environmental, geographic and social media information. More than 80% of all data is unstructured. Knowing how to retrieve information from unstructured data and combine it with relevant, structured data is critical for obtaining meaningful associations and predictions. Analyses of unstructured data typically require techniques such as data mining, natural language processing and text analytics. Data science methodologies cover extraction of information from both structured and unstructured data.

In the following sections, we review the different aspects of data science in the context of health services and health policy research and provide background and relevance to ICES. Each section concludes with recommendations for ICES that the authors believe are the most innovative and feasible.

EXHIBIT 3 The four V's of big data



Biomedical Big Data

Biomedical big data include the diverse, complex, disorganized, massive and multimodal data being generated by researchers, hospitals, mobile devices, and applications (apps) around the world.[15] These data include imaging, phenotypic, molecular, exposure, health, behavioral and many other types of data, including unstructured data such as free text in medical records, electrocardiogram data or detailed anatomic images, such as magnetic resonance imaging (MRI) scans. Important challenges relate to privacy and security, data federation, electronic data capture and standards, storage and high-performance computing, and high-dimensional data analytics. International examples of large-scale efforts to invest in these data for research include Genomics England (funded to sequence 100,000 whole genomes during routine clinical care), the China Kadoorie Biobank, PCORnet, the National Patient-Centered Clinical Research Network, and the $215 million Precision Medicine Initiative in the US that includes $130 million allocated to the National Institutes of Health to build a national, large-scale research cohort to "extend precision medicine to all diseases by building a national research cohort of one million or more U.S. participants."[3] Big data become transformative when disparate data sets can be linked at the individual person level.[16] Sources of biomedical big data in Ontario are listed in **Appendix A.2**.

There are three related but distinct streams of relevant biomedical big data activity in Ontario. The first is the building of organizations and IT platforms that enable the data management and analysis of this data, largely from research initiatives but with a view to linking to routinely collected health data at ICES. This would include the work of the Ontario Brain Institute in conjunction with Indoc Research. The second is the development of a number of external research groups collecting a range of data on various cohorts and building research infrastructure intended for use by a wide variety of investigators. Many of these groups have been explicit in their intent to link to routinely collected health data. The most prominent of these is the Ontario Health Study, but there are other groups, such as TARGet Kids!, with whom ICES has had discussions. The Ontario

Health Study is a large longitudinal health study investigating causes of chronic diseases, such as cancer, heart disease and diabetes, with over 230,000 Ontario enrollees since 2010. Finally, there is interest in and work to mobilize electronic health records, which contain structured and unstructured clinical data including laboratory results, clinical diagnoses and clinical history. These health records include administrative laboratory data that have greater reach (in some cases, population-based) such as the Ontario Laboratories Information System (OLIS), and prenatal and neonatal screening data such as the Better Outcomes Registry and Network Information System (BORN). In addition, they include primary care records such as the Electronic Medical Record Administrative Data Linked Database (EMRALD), the Canadian Sentinel Practitioner Surveillance Network, and the University of Toronto Practice-Based Research Network (UTOPIAN), as well as hospital data warehouses and electronic health records for discrete patient populations.

Electronic Medical Record Data Expansion at ICES

In recent years, ICES has held an increasing interest in the use of data from primary care electronic medical records (EMRs) to enable the exploration of research questions not previously measurable through linkage with administrative data (including those for primary care). ICES is currently developing a strategy and recommendations for how to increase the number of linked EMRs held at ICES, and expand the use of EMR data in its research. In Ontario, much work has been done to implement pragmatic EMR infrastructure solutions by organizations such as Canada Health Infoway and eHealth Ontario. At ICES, considerable work has been done to create and support the development of EMRALD, as the added capacity of linkage to administrative data has enormous potential. To date, EMRALD has extracted and linked the electronic charts of more than 500,000 patients from 350 family physicians in 42 primary care clinics across various Local Health Integration Networks (LHINs) in Ontario. EMR data could form an enviable platform for observational studies (such as pharmacosurveillance and genetics studies) and clinical trials, and has the potential to support and build capacity for primary care research. A range of trial methodologies could make use of better quality, linked EMR data through practice-based research networks and specialist groups of researchers working at a provincial, national and international level. For example, ICES has recently developed mechanisms to permit, under certain circumstances, the re-identification of individual patients based on linked data so that, for example, patients deemed eligible for a study may be approached by approved researchers for consent to participate, or additional testing of biologic specimens on a defined subset of patients of interest may be undertaken. We have the opportunity for in-house expansion of the EMR data holdings at ICES through key partnerships with the Canadian Primary Care Sentinel Surveillance Network and other collaborations, such as the University of Toronto Practice-Based Research Network. These efforts are focused on primary care, but community-based specialists and hospitals are increasingly using EMRs as well. ICES will continue to prioritize the expansion of primary care EMRs but remains alert to opportunities to expand to EMRs from other providers and institutions.

The development of natural language processing methods will permit the extraction of information on whether a subject has specific diseases or conditions. The use of software and algorithms to extract meaning from freeform text will likely increase as EMRs are imported to ICES. Methods to develop and validate algorithms for identifying subjects with given conditions and diseases will grow in importance as such data proliferate and are used to complement existing electronic administrative data currently held at ICES. Because text mining programs require expertise to adapt to EMRs, EMRALD has partnered with University of Toronto computer scientists to develop text mining algorithms that will validate conditions including hypertension, diabetes, atrial fibrillation, ischemic heart disease, rheumatoid arthritis and brain disorders such as Parkinson's disease, dementia and multiple sclerosis, and is continuing to develop and validate algorithms for use with administrative data. Natural language processing combined with EMRs provides the opportunity for far greater clinical detail and phenotype information than would ever be available in administrative data. However, there are substantial challenges as well. De-identification and

maintaining privacy protection in unstructured and detailed free text, the challenge of storing and accessing massive amounts of data, governance, and the challenges of routinely updating EMRs are all issues that ICES will need to grapple with as we move to enrich our repository with additional records.

Relevance to ICES

Many potential data sources that contain personal health information allow linkage to other data. Much work is already underway at ICES to integrate sources of biomedical big data, such as OLIS and EMRALD. Analytic techniques such as natural language processing for unstructured data found in the free-text portions of EMRALD data and the IT infrastructure to support large, unstructured raw data files will need to be addressed. The privacy implications of uniquely identifiable data, such as detailed neuro-images and whole sequence genomes, will need to be reviewed. Unsupervised analytic methods, which look for clusters and associations without prior hypotheses and often without a specified outcome, may require additional privacy policy review. The current use of ICES data under Ontario's *Personal Health Information Protection Act* requires justification of clear research objectives and documentation of data elements. Finally, linkage with biomarker data can yield interesting results only if we can produce good phenotype and outcomes data.

Recommendations

- ICES should define the scope of use of biomedical big data, for example, adding deeper clinical covariates to traditional epidemiological and health services research vs. unstructured genetic/biomarker association studies.
- ICES should seek the necessary resources to expand the linked primary care EMR records to obtain, at minimum, a sample that is population-based.
- ICES should develop the necessary expertise and tools for exploiting EMRs, namely natural language processing capability, methods for data storage and access, regular updates of EMR records, and appropriate methods for de-identification and privacy protection. ICES needs to also proactively develop approaches to EMR data governance that will be acceptable to stakeholders.
- ICES should focus on existing biomedical big data, such as the Ontario Laboratories Information System, to build the business case for resources needed to incorporate unstructured biomedical big data and make it research ready.
- ICES requires privacy review and opinion around unstructured data analysis and uniquely identifiable data, such as whole genomes.
- ICES should consider participating in pilot projects using biomedical big data and data science methods with scientists from university departments such as computer science.

High-Performance Computing Environment

Essential to the execution of complex interrogation of data, either structured or unstructured, is a robust and efficient computing environment. The current computing, research and analytics environment is a balance between doing high-level research and adhering to the constraints of privacy requirements. Traditional computational environments consisting of a storage layer and a processing layer (with minimal storage) increasingly cannot meet the high throughput and rapid analytics required by complex computational methods, such as natural language processing, machine-learning methods and even some traditional statistical methods such as Bayesian models. One of the key requirements in the development of our computing system and one that may impact our ability to conduct complex data mining is the requirement to audit and log each use of the data. Under our prescribed entity agreement, ICES is obligated to track each time a user accesses the data repository as well as any changes that are made to it. To accomplish this, we use SAS software as the backbone of our system. These privacy requirements do not permit analysts to run common statistical software packages such as R and C directly on the source data.

Beginning in the 1990s and continuing into the early 2000s, the computing speed of single processors represented the rate-limiting step at which greatest efficiency could be achieved. The introduction of multi-threading, parallel processing and massively paralleled processors led to considerable advances in processing speed and throughput. These architectures have become standard for current hardware and software. The current computational infrastructure at ICES employs several of these techniques and has achieved substantial gains in computational efficiency over the past five years. However, both globally and at ICES, the processing speed "weak link" has given way to a second great challenge in computational analytics: data transfer speed (i.e., read/write or input/output speed). While the processors may have little problem executing computations, data flow represents the bottleneck. This issue has been a particular challenge at ICES for

the past decade as our data repository has tripled in size and grown in complexity. This issue will persist as new and complex data sources are acquired. For example, the acquisition of data from the Ontario Laboratory Information System instantaneously doubled the size of the data repository in early 2016. To address this challenge, modern computational infrastructure employs processors with massive memory stores for large-scale in-memory computation. These systems and 'appliances' may currently be cost-prohibitive for small and medium-sized organizations; however, these costs are decreasing and alternative methodologies for data management and query are being developed. For example, the Hadoop software library provides a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Increasingly, novel techniques and approaches are being promoted by ICES researchers and collaborators that involve data mining, social network analysis and natural language processing. These will require ICES to continually invest in computational infrastructure, identify novel approaches to data storage and analytics, and partner with experts, academic and non-academic, to maintain research excellence and efficiency. That said, over 90% of ICES projects can be handled by the existing computing structure. For the remaining 5% to 10%, we will need to be creative about finding environments or partners to assist in expanding our computational capacity.

Developing a Data Safe Haven for Ontario

An important aspect of the data science agenda at ICES involves innovations in data access, linkage and analysis. These elements align well with ICES' institutional goals, which include scientific excellence and improving access to types of data that have not traditionally been housed at ICES.

ICES intends to further improve access to data for researchers and stakeholders through the creation of an Ontario Data Safe Haven. Data safe havens have been variably defined. One useful definition is "a repository in which useful but potentially sensitive data may be kept securely under governance and informatics systems that are fit-for-purpose and appropriately tailored to the nature of the data being maintained, and may be accessed and utilized by legitimate users undertaking work and research contributing to biomedicine, health and/or to ongoing development of healthcare systems."[17]

Data housed in a data safe haven could include routinely gathered, patient-level administrative data from both health and non-health government ministries and agencies, clinical data (e.g., from a practice-based or hospital setting), data captured as part of a specific research study (e.g., randomized controlled trial or cohort), and/or survey, biometric or biologic data which are broadly representative of a population. A data safe haven may contain other types of data such as social services, education, financial and environmental. With appropriate approvals and safeguards, data in a data safe haven can be used to answer research questions and address issues related to policy development and evaluation; potentially, such data could be used in exploratory analyses free of pre-specified hypotheses. In addition, they could be made available for re-analysis and/or linked at the individual level to other data to answer new questions, for example, linking existing data from randomized controlled trials to administrative data to study long-term outcomes. Access to and analysis of data in a data safe haven would be subject to relevant legislation, regulations, data sharing agreements with data custodians, and policies of the data safe haven and ICES. Several key criteria have been proposed for data safe havens: data maintenance and access should be socially acceptable and appropriate, data should be veritable, and data should be safe and secure.

Relevance to ICES

Currently, computational infrastructure meets most of ICES' needs. Recent investments have balanced ICES' obligations as a prescribed entity (using a SAS metadata layer that logs each access to the ICES data repository) with the analytic efficiency required by researchers in a fiscally constrained environment (using a SAS grid computing environment for parallel processing and solid state storage for commonly accessed data). However, the ICES data and analytic infrastructure is currently not sufficient to manage

the workload and computational requirements that may be brought on by the continued growth in the numbers of projects and scientists, the continued expansion of data holdings such as OLIS, and the continued exploration of novel analytic and computationally intensive methods.

Creating the data safe haven in an environment designed for secure management of personal health information will ensure that it can provide services for both health and non-health data holders. **HPC4Health** is a secure, private, cloud-based computational facility that provides on-demand high-performance computational resources to clients for clinical research; it is situated within a secure hospital environment (the Hospital for Sick Children in Toronto) and is authorized to host sensitive medical data sets. ICES is currently piloting the creation of an Ontario Data Safe Haven in partnership with HPC4Health that brings together data assets and provides secure controlled access in an environment of advanced computational capacity. The Ontario Data Safe Haven, hosted at HPC4Health and managed by ICES, would provide several functions:

- Researchers could post, access and analyze individual data sets of which they are the custodian or for which they have research ethics board approval.
- With the use of ICES privacy-preserving protocols, the identifying information could be transformed into coded data and evaluated for the possibility of re-identification, resulting in risk-reduced, coded data sets. Remote access to these coded data sets could be provided once relevant approvals were obtained.
- Through linkage, individual data sets could be enhanced with information from the ICES data repository. These enhanced data sets could be transformed into risk-reduced, coded data and made accessible to the research team once approvals had been obtained. For example, ICES data holdings could enable long-term follow-up for clinical trials.
- At the request of the data custodian, individual data sets could become part of the ICES data repository. Linkage would allow researchers, health system planners, and decision makers to realize the true potential of diverse data assets. Data safe haven services would be established to allow "one-stop shopping" for all functions, resulting in an efficient process for a user to transition from accessing their own data set to linking to other data.

Recommendations

- ICES should establish partnerships with data and computational infrastructure exemplars such as HPC4Health, public sector funders and industrial partners to develop and implement novel data storage and access structures, such as the Ontario Data Safe Haven. Advance the planning by implementing the data safe haven pilot proposal.
- ICES should ensure that the Data Quality and Information Management and Information Technology departments remain informed about current data storage and computational trends, and identify opportunities and technologies appropriate for the ICES environment.
- ICES should ensure that the ICES Research and Data platform develops mechanisms and policies to enable advances in software environment (e.g., a workaround to use STATA, R and C directly on the databases).
- ICES should encourage its Corporate Services team to seek out opportunities for funding (e.g., through the Canadian Foundation for Innovation) to ensure regular renewal of and upgrades to our data infrastructure. In addition, modify the current ICES Finance practice of refreshing budget infrastructure costs each fiscal year rather than earmarking large periodical initiative costs.

Data Integration

Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.[18] It includes record linkage, data anonymization and standardization, data security, development of metadata, and data quality assessments. A comprehensive data integration framework is the combination of all these processes which are streamlined and automated when possible. Such a framework delivers trusted, linkable and reliable data from a variety of sources. The data quality framework distinguishes between database-specific and research-specific quality, and focuses on the former. Research-specific quality implies data quality assessments of project-specific cohorts that combine multiple administrative data and apply specific exclusion criteria.

In 2012, ICES adopted a data quality framework that had been developed at the Manitoba Centre for Health Policy (MCHP).[19,20] Over the last few years, this framework has been adapted to suit ICES' needs. Before adoption, various data quality frameworks were reviewed, including those from the Canadian Institute for Health Information (CIHI), the Public Health Agency of Canada (PHAC), Statistics Canada and the Australian Bureau of Statistics.[21-23] The data quality reports generated by the UK's National Health Service were considered as a model for our reports.[24] The SAIL research group at the University of Wales Swansea has begun to adopt data quality framework concepts and tools into its environment.

The current ICES data quality framework is a streamlined, systematic process to assess the six data quality dimensions of *accuracy* (completeness and correctness), *internal validity* (internal consistency, stability over time, linkability), *external validity* (comparisons with other data or published reports), *timeliness, interpretability* (quality and usability of data documentation) and *relevance* (usability of data). This framework is equipped with a suite of SAS macros to automate and standardize the processes. These concepts are described in more detail in **Appendix A.3**.

There is a need for a repository of validated definitions and algorithms to use with health administrative data. A comprehensive concept dictionary would help investigators carry out methodologically sound work using consistent and validated algorithms and prevent re-inventing the wheel. ICES has created an ICES data wiki and ICES

drug definitions and adapted the MCHP concept dictionary to Ontario data. However, most of this information is outdated and requires major review.

Recently, ICES Data Quality and Information Management formed a working group to create an ICES concept dictionary with the collaboration of MCHP. The group is evaluating the idea of a national, collaborative, web-based application including partners such as MCHP, Health Quality Ontario, Cancer Care Ontario, CIHI, PHAC, Statistics Canada and Population Data BC. Participant organizations would have the privilege of adding validated definitions from their province's data, and approving and publishing added concepts. The application would have advanced search capability to find definitions and filter by province. We are in the process of reaching out to grant funders and potential partner organizations. The group has met to identify potential funding sources.

Current ICES record linkage strategies use probabilistic record linkage methods that include considerable manual effort. Manual efforts are time-consuming and costly, and introduce human error as a source of linkage bias. Part of ICES' strategic plan is to increase access to social data such as housing, education and justice. Cleaning and linking these data to administrative health data will require advanced record linkage methods such as fuzzy matching and machine learning to be accurate and efficient.[25,26] The goal is to minimize manual review while maintaining linkage accuracy.

Finally, from a privacy perspective, privacy-preserving record linkage methods are required to link certain interministerial data to ICES health administrative data. Examples of data that would require such methods are those from the Ministry of Education and the Ontario Brain Institute

Training and dissemination of data quality tools

Scientists from the Manitoba Centre for Health Policy (MCHP), the Public Health Agency of Canada, the B.C. Ministry of Health and ICES jointly received a grant from the Canadian Institutes of Health Research to disseminate the suite of data quality tools developed by MCHP and ICES to other organizations with similar activities across the country. These presentations have created interest in training and consulting support to set up a systematic data quality assessment for other provinces.
ICES and MCHP held a data quality workshop at the International Population Data Linkage Conference in Swansea, Wales, in 2016. Workshop participants were given the opportunity to run ICES data quality tools with simulated data and interpret the results.
MCHP licensed its own version of the tools under a GNU General Public License and made them available on the University of Manitoba's website. The ICES Data Quality and Information Management team is planning to go through the same process in 2017/18.

Recommendations

- ICES should fully implement the data quality framework and develop the necessary tools to automate the outstanding dimensions and components.
- ICES should implement an advanced concept dictionary for knowledge translation.
- ICES should move to modern methods for data anonymization and record linkage, including privacy-preserving record linkage software.

Data Science Methods

Data science methods generally comprise inferential and predictive modelling. Traditional statistical methods focus on inferential modelling, which involves causal inference and hypothesis testing—the primary methods used in most ICES health services research. We have excellent capability in linear, logistic, Poisson and survival regression models, and we employ a diverse range of methods for the analysis of observational data, such as longitudinal data analysis with GEE estimation, multilevel data analysis, case-control matching, propensity score–based approaches and instrumental variable analysis. However, there are areas where our capability in inferential modelling is limited or ad hoc. Examples with particularly high relevance to ICES are *competing risk methods* (since we often analyze the incidence of an adverse outcome in the presence of a competing risk whose occurrence precludes the occurrence of the primary event of interest) and *multistate models* (since we often analyze the occurrence of more than one type of event). While these are not the primary methods we use, they could be used in specific studies if the analytic staff were trained to understand and apply them.

Predictive Modelling

Predicting individuals at high risk for occurrence of an adverse event, incidence of disease, or resource use is an important issue in clinical medicine, health services research, and population and public health. Prediction permits the effective risk stratification of patients and the targeting of interventions to appropriate strata of patients to improve outcomes and prognosis. In epidemiology, the identification of risk factors is important for developing initiatives to reduce the risk of the occurrence of adverse events; an example is the Framingham risk score. Breiman

suggested that there are two approaches or paradigms to prediction: the *model-based approaches* that predominate in the field of statistics and are commonly used by ICES researchers, and *model-free prediction algorithms* that are popular in the field of computer science.[27]

Model-based approaches to prediction assume the presence of an parametric or semi-parametric model underlying the data. When outcomes are binary or counted in nature, the common approaches are logistic and Poisson regression. When outcomes are time-to-event in nature, such as survival, the common approach is to use the Cox proportional hazards regression model. Methods for fitting statistical regression models for predicting outcomes are described by Harrell, Royston and Steyerberg.[28-30]

The algorithmic-based or machine-learning approach to prediction makes no assumptions about an underlying statistical model through which the observed outcomes were generated. Instead, the focus is on developing an algorithm that predicts or classifies outcomes with minimal error. These methods may be suitable even when the number of potential covariates (p) greatly exceeds the number of unique observations (N). There are a large number of such algorithms. One of the oldest and simplest is that of regression trees, alternatively called classification and regression trees (CART) or recursive partitioning.[31] Regression trees are obtained by recursively partitioning the data and estimating the mean or proportion of successes in each node in each partition so that the partitioning can be represented graphically as a decision tree.[32] Extensions and refinements include bagged regression trees (bootstrap aggregation of regression trees), random forests, boosted regression trees and generalized boosting methods.[33-36] These modifications involve aggregating predictions across a large set of regression trees. Aggregating predictions across an ensemble of models has been shown to improve the accuracy of predictions and reduce their variability.

Alternative non-tree-based prediction methods include neural networks and support vector machines.[35] Neural networks involve an algorithm in which there are one or more hidden layers between the predictor variables and the outcome variable. The inner layers are considered latent variables and are simply linear or nonlinear combinations of the input or predictor variables. Hastie writes, "There has been a great deal of *hype* surrounding neural networks, making them seem magical and mysterious. As we make clear, … they are just nonlinear statistical models."[35] *Deep learning* has been used to describe the family of prediction algorithms that include one or more hidden layers between the predictor variables and the outcome variable. Neural networks are one type of deep learning method.

Epidemiologic research has at least two objectives: (1) predicting an individual's risk for experiencing an adverse outcome or developing a given disease, and (2) identifying characteristics or exposures (risk factors) that predispose an individual to an increased risk of an outcome. An algorithmic approach can identify individuals at increased risk of an adverse event but cannot easily quantify the magnitude of the influence of a particular exposure or why these individuals are at increased risk.

In contrast, machine-learning methods can be useful for data mining and hypothesis generation, as well as for dimension reduction. While there appear to be potential benefits to using machine-learning methods for prediction in health research, to date the benefits have been mixed. A systematic review of traditional and machine-learning methods to predict a variety of outcomes (mortality, readmission, return to theatre, outpatient nonattendance) for various conditions (heart failure, acute myocardial infarction (AMI), colorectal and orthopaedic surgery) using UK National Health Service administrative data showed that machine-learning techniques offered little improvement over standard models.[37] ICES studies have shown a similar lack of improvement in predicting AMI mortality and heart failure subtypes.[38-41] A few studies using data from electronic medical records have shown improved prediction of 30-day readmissions and emergency department revisits for congestive heart failure (CHF) using machine-learning methods compared to standard techniques.[42,43] Others have shown the utility of using data mining compared to conventional echographic parameters for the prediction of heart rate variability[44] and the utility of using natural language processing of EMR data for the prediction of 30-day CHF readmissions and post-operative complications compared to traditional discharge records.[45,46] These studies demonstrate the potential and feasibility of the new methods, but to date, there is little evidence that they can be translated into tools that can improve care.[1]

These newer models are generally less understood by stakeholders and users of research evidence, as it is not possible to report the relative rates associated with various risk factors. The models may be likened to a black box and are possibly unreliable. Certainly the well-publicized problems with the Google flu algorithm do not reassure users.[47] As Krumholz notes, "False positive findings from investigations into genomic associations that started with the data are indeed an example of the hazard of pursuing knowledge about causation without theory."[2,48] On the other hand, even within health care services or systems research, common knowledge about what affects outcomes may be wrong, so that some of these new data mining approaches may shed new light, provided the multisource data are available.

Rumsfeld argues that the new machine-learning tools will need to demonstrate clinical utility before they can be considered useful and reliable. In addition, methodological issues such as data quality, consistency, validity and privacy need to be addressed before such methods can be exploited to their full potential. Big data analytics is poised to advance the concept of precision medicine but currently these developments are nascent and the big data era in health care is just beginning.

Data Analysis

Data science requires an open and disciplined approach to ensure validity and reproducibility of research findings. In terms of data exploration and preparation, 65% to 90% of data analysis efforts are expended in cleaning and preparing messy data, or "data wrangling." Common, easily accessible tools are needed to translate messy data into clean, useable data. This involves data representation and transformation to restructure original data into a more revealing and useable form. Traditional methods of running jobs interactively by hand, and copying and pasting into documents are understood to be irresponsible. Optimal practices include reproducible computation, including sharing of code and data, and creation of automated macros to generate all the computations and analyses in a project, including the final publication results. Data scientists should be conversant with several data processing and analysis languages. They should develop macros that can be used for future projects and document the individual steps of an analysis.

Relevance to ICES

Predictive models play an essential role in ICES research in the development of risk prediction models for the occurrence of outcomes in disease-specific cohorts. Examples include the EFFECT-HF mortality model for predicting mortality in patients hospitalized with heart failure, the EHMRG model for predicting short-term mortality for patients presenting to the emergency department with heart failure, the AFTER model for predicting short-term mortality in patients presenting to the emergency department with atrial fibrillation, and the DPORT model for predicting populations at high risk of developing diabetes within 10 years.[49-53] These prediction models were developed using conventional regression -based approaches. Currently, machine-learning techniques are a boutique application for ICES, as our prediction models have tended to use a moderate number of potential predictors that were based on content knowledge and were selected from a pre-specified list of candidate variables.

That said, ICES scientists should be aware of modern prediction methods and the advantages they may offer in specific settings. The relevance of these new prediction methods for ICES researchers may increase when proteomics or similar data sets with a huge number of potential predictor variables (p) are housed at ICES and linked to other health administrative data for outcomes ascertainment. These methods, in particular random forests, are likely to have advantages when the number of potential covariates (p) is very

large relative to the number of subjects (*N*). Linkage with biomarker data may yield interesting results, but only if we can produce good phenotype and outcomes data. The Ontario Health Study has submitted a request to link genomic data with ICES data for predicting, for example, colorectal cancer diagnosis based on targeted genetic markers. This would involve about one million markers on 1,000 people and be an excellent collaboration in which to implement machine-learning methods.

Recommendations

- The core business of ICES is the production of evidence. Since it is not yet clear if these new methods are an improvement on existing risk prediction methods, it is premature to recommend the use of machine-learning methods in standard analyses. However, these methods do have great utility in data mining and hypothesis generation activities and may lead to new insights that can be further tested using conventional models. Therefore, we recommend educating scientists and analysts in such methods as a first step, and partnering with external scientists who wish to lead such studies. Modern prediction methods could be used in specific studies to increase the accuracy and reliability of our findings. ICES could also benefit from increasing its efforts to implement these sophisticated, cutting-edge statistical techniques. The new role of staff methodologist is perfect for the implementation and dissemination of such new methods.

- Methodologist scientists at ICES should have a role in promoting more sophisticated traditional statistical methods by identifying gaps, recommending and implementing new methods and training analysts in their use.

- The education and training of ICES analytic staff in the use of more complex, traditional inferential methods, guided by the methodologist scientists, is recommended.

- Analysts and methodologists at ICES should be exposed to the R and STATA statistical programming languages because new methods often appear in these languages before they appear in SAS. R also has excellent graphical capabilities.

- ICES should increase its capacity in data science methods by collaborating with external scientists who have expertise in machine learning. For example, we could provide access to data in return for expertise. We need to reflect upon which research we wish to lead and which research we wish to enable.

- ICES should explore collaboration with a few research groups interested in linking their genetic or biomarker data with ICES data, since for gene association studies, machine-learning techniques might be useful.

Data Visualization

Data visualization facilitates how we make sense of data. It addresses how we encode and share information using visual objects. It is considered to be both a science and an art. Among other disciplines, it has origins in descriptive statistics. Tukey encouraged statisticians to use visualization as a mechanism for analyzing data in his textbook *Exploratory Data Analysis*.[54] Advancements in computing power for assembling, storing and displaying data have led to an explosion of data visualization in academic, media, business and policy circles.

Data visualization takes advantage of how the human visual system perceives information. In *Data Points*, Yau provides an overview of how elements work together to build a graph.[55] Visual cues such as position, length, angle, shape, area, orientation and colour are used to encode information. By combining these visual cues with coordinate systems (the rules and structures governing where cues are positioned) and measures of scale, data visualizations begin to impart information. While different coordinate systems exist, the most common are Cartesian, polar and geographic. Context also plays a role and can be thought of as the information supplied to characterize "who, what, where, when and why" for data through titles, labels and legends.

Data analyses benefit from data visualization methods in both analysis and presentation. With respect to analysis, data visualization techniques could be used to explore the source data to examine basic relationships using common tools such as line and bar graphs. These techniques can be used to analyze more complex relationships through network analysis, heat maps and dynamic or interactive displays. The principle of matching the graphical method to the analytical purpose is at the core of data visualizations for analysis. Several primers exist to guide these decisions for different audiences.[56] With respect to presentation, the concepts of clarity, precision and efficiency are key.[57] Data visualization techniques are employed to *reveal* and communicate. While these concepts are commonly taught as part of undergraduate and graduate training programs, they are worth revisiting. They include interactive and dynamic data visualization products, such as tree

maps, heat maps and dynamic visuals, while leveraging existing or free software. Notable tools include R, ggplot2 and Shiny for interactive web applications, as well as D3.js, Excel (for sparklines), Google Charts, Gapminder, Tableau and InstantAtlas.

Relevance to ICES

Data visualization techniques can be used to help us process and prepare health administrative data to make them research ready. By visualizing the data and the metrics derived from the data, we can see patterns in data quality, techniques for formatting and parsing data, record linkage rates, the stability of variables over time, and missing information.

Data visualization techniques can support ICES analyses in two main areas. First, they support *research quality* when we are designing studies and building cohorts. Data visualization can be used to examine the distribution of variables, outlying data points and associations.

Data visualization is also useful as an *analytical* tool to explore ways of looking at correlation and causation[58] and to convey multivariable relationships, including comparisons across multiple variables (heat maps, network analysis, radar/star charts, parallel coordinates), dimension reduction (techniques for multidimensional scaling and clustering) and searching for outliers (histograms, box plots). These techniques can also be used to explore variation in exposures and outcomes over time and across regions. Particularly compelling are dynamic, interactive visuals and maps, such as those prepared for the **Global Burden of Disease study**, the **Dartmouth Atlas of Health Care**, the **Institute for Health Metrics and Evaluation** at the University of Washington, Hans Rosling's TED talks (including "**The Best Stats You've Ever Seen**") and The Commonwealth Fund's **Quality-Spending Interactive tool**.

Data visualization is useful for knowledge translation and sharing findings from ICES research with stakeholders. Opportunities exist to build on our use of information graphics in our Research Practice and Communications departments. **Project Big Life** uses web calculators and data visualization to display ICES data in its reports; these include personalized risk calculators for life expectancy, future health care costs and daily sodium consumption. ICES is starting to incorporate enhanced visuals in communications vehicles such as reports, briefing notes, news and social media outreach. For example, ICES is using **infographics** to provide visual representations of our research highlights.

Recommendation

- ICES should build internal capacity in data visualization techniques.

Distributed Data Analysis Networks

Many research questions involving rare exposures or outcomes can be answered only by combining data across multiple jurisdictions or health systems, that is, through a research network. Once established, such networks can also exploit natural experiments that exist across jurisdictions due to differences in health systems, policy or practice. The network infrastructure may also be exploited to study interventions prospectively.

Research networks typically adopt one of two approaches to data combination:

- *Centralized analysis*, which involves the standardization and physical transfer of site-level common data sets, or
- *Distributed site-level analysis*, which involves either an agreed-upon common data set and SAS programs (the Common Data Model) or a common analytical protocol with local translation (the Common Protocol Model).

A key advantage of distributed (rather than centralized) analysis is that existing data sharing agreements often prohibit the physical transfer of record-level data. Other advantages include site-level control and understanding of local data sets and coding practices, and avoidance of the financial and opportunity costs associated with creating and managing a central repository. Both the Common Data Model and the Common Protocol Model of distributed analysis require methods for the eventual pooling of site-level results. There are other approaches to distributed data analysis (**Exhibit A.3** in the Appendix), but we will not mention them further as they are either impractical, not timely or incur privacy issues. **Exhibit 4** reports established distributed data analysis networks in Canada and abroad.

A key distinguishing feature among distributed data analysis networks is whether they employ a Common Data Model or a Common Protocol Model for analysis. **Exhibit 5** compares the main advantages of the two approaches. Key differences are the size and timing of the financial investment, benefits to the data partners, output quality in terms of risk of error and confounding bias, and output timeliness.

EXHIBIT 4 Characteristics of established distributed data networks in Canada, the United States and Europe

Network	Field	Model	No. of Sites; No. of Lives Covered
Canada			
CNODES (Canadian Network for Observational Drug Effect Studies)[65]	Drug safety	Common Protocol	8; 45 million+
CCDSS (Canadian Chronic Disease Surveillance System)	Chronic disease surveillance	Common Protocol/ SAS code	13; 35 million+
United States			
FDA Sentinel Initiative[60]	Drug safety	Common Data	17; 125 million
OMOP (Observational Medical Outcomes Partnership)[66,69]	Methods for drug safety, effectiveness research	Common Data	10; 130 million
PCORnet (Patient-Centered Clinical Research Network)	Comparative clinical effectiveness	Common Data	FDA Sentinel sites and others as required
HCSRN (Heath Care Systems Research Network) [formerly HMO Research Network; extends to Israel]	Various	Common Data	19; 29 million
Europe			
PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium)[70]	Drug safety	Common Protocol	6; 45 million
Farr Institute of Health Informatics Research	Various	Common Protocol	4

EXHIBIT 5 Advantages of Common Data and Common Protocol models for distributed data analysis

Domain	Common Data Model (and SAS Programs)	Common Protocol Model
Financial costs • Set-up • Operation	• Substantial upfront investment in the development of and translation to common data sets, and in the development and testing of common SAS programs. • Annual operating and maintenance costs are comparatively less than for the Common Protocol Model, although staff time is needed to update common data sets, run study code and vet outputs.	• While set-up costs are less than for the Common Data Model, annual operating costs can be substantial, depending on the nature of the study questions/designs and the number of projects.
Benefits to engagement of individual data partners	• Limited	• Substantial opportunity for local input and capacity-building.
Output quality • Risks for error • Risks for confounding bias	• Well-tested, modular programs minimize risks for error. • Most analyses incorporate limited adjustment; however, more sophisticated programs that incorporate matching and stratification on a propensity score are now in use. A limitation is that these programs are typically constrained by the common data set.	• Despite the use of standardized and phased analytical protocols and some shared common code, local translation increases the risk for error relative to a Common Data Model. • Flexibility allows for maximum control for confounding bias by permitting sites to incorporate all available data (e.g., in construction of propensity scores).
Output timeliness	• Hours to days	• Weeks to months

While the Common Data Model represents a substantial upfront investment in the development and maintenance of an agreed-upon minimum data set and code, once established, networks using such a model can carry out extremely rapid, error-free analyses.[59-63] However, this model can suffer from data loss when important data are not collected in all networks. In contrast, the Common Protocol Model requires a relatively small investment, common data elements and code, and shared analytic plans, but analyses can take weeks to months. Curtis describes the challenges of developing four US data networks of combined EMR data and administrative data for research purposes; these involved the Food and Drug Administration, the National Institutes of Health, the Patient-Centered Outcomes Research Institute and a state public health network.[64]

Several ICES research teams, including those involved in the Canadian Network for Observational Drug Effect Studies (CNODES)[65] and the Canadian Chronic Disease Surveillance System, have experience with multicentre distributed analyses using common analytical protocols but with shared SAS programs. An important advantage of this approach is flexibility in both study design and data sources. However, the local translation and implementation of written protocols can be time consuming and error prone. For straightforward research questions and designs that involve a limited number of data sources and data elements, a common data model with shared SAS code can be more efficient.

Relevance to ICES

ICES is playing a leading role in creating the Pan-Canadian Real-world Health Data Network (PRHDN), a distributed data network that will permit researchers and policy and decision makers across Canada to make effective use of linked and linkable administrative data holdings and expertise in multiprovince studies and initiatives without requiring that data leave provincial boundaries. Taking lessons from the common data models of the state-of-the-art FDA Sentinel Project[61,62] and the Observational Medical Outcomes Partnership,[63,66] pilot work involving CNODES and PRHDN is underway to develop a Canadian common data set that comprises core data elements from provincial health insurance registries, hospital discharge abstracts, physician service claims, and prescription drug claims. To capitalize on the strengths of Canada's population-based health data, PRHDN will develop and validate algorithms that generate new harmonized common data and make these available, and develop common analytic protocols that can be used when harmonization of data between provinces is not practical or possible. Pilot studies are expected in 2017. Currently, PRHDN is seeking funding from the Canadian Institutes of Health Research, Health Canada and other sources.

Parallel research is needed on methodological work for sharing and pooling study summary measures and analytic intermediates that satisfy provincial privacy requirements. Such work could include exploring the comparative strengths and limitations of the privacy-preserving methods currently in use, including analysis of propensity score-defined strata, case-centred analyses of risk data, meta-analysis of site-level effect estimates, and methods for pooling individual data without sharing the data.[59,67,68]

Recommendations

- ICES should continue to support networks such as CNODES and PRHDN in their efforts to develop a standardized national common data model for cross-provincial research. Implementation of PRHDN, assuming funding success, should be a key ICES priority.
- ICES should support the advancement of the technical and methodological science of distributed analysis.
- ICES should support the study and dissemination of methods for pooling study summary measures and analytic intermediates that satisfy privacy requirements.

Educational Initiatives

As a new interdisciplinary specialty, data science will involve new interdisciplinary training. Educational initiatives and opportunities for developing expertise in data science have exploded in the United States over the past decade. In most instances, these initiatives have been led by computer science departments, and occasionally engineering and business schools, but typically not statistics departments. In recent years, Canada has seen a small increase in the number of universities offering undergraduate and graduate programs in data science and big data. Canadian universities with existing graduate programs include Simon Fraser University, the University of Waterloo, Carleton University, Dalhousie University and the University of Alberta. In 2016, the University of British Columbia launched a comprehensive master of data science program with 24 one-credit modules; 80% of the course offerings are repurposed from existing statistics and computer science courses and 20% is new material. The University of Toronto will launch an undergraduate program in data science in 2017 and a master's program in 2018. At most Canadian universities, the graduate program in data science is a joint effort between departments, typically computer science and statistics, with a single department taking the lead in formally offering the degree.

Among the few universities that have their own data science institute or department, the faculty consists of professors from a mix of specialties, including computer science, computer engineering, mathematics and statistics. Due to the interdisciplinary nature of data science, requirements include the completion of courses in computer science, machine learning, text mining, data visualization and mathematical statistics. Harvard University's Department of Biostatistics is enhancing computer science training by including big data computing (cloud-based computing, scaling up) in its curriculum; the department is revising its PhD qualifying exam to add computing skills to the mix.

There are many online resources that can be utilized for data science education. For example, MOOCs (massive open online courses) offer open access and unlimited participation and also allow for

user forums to support interactions between students and professors. Data science bootcamps, such as Metis, are accredited, multiweek programs that aim to accelerate the careers of data scientists by focusing on applications and topics in data structures, algorithms and languages (e.g., Python) that are pertinent to data science work. The annual American Statistical Association Conference on Statistical Practice provides courses and tutorials on data science and other novel statistical methods. The Eastern North American Region (ENAR) of the International Biometric Society and the annual Joint Statistical Meetings (JSM) also run excellent workshops on machine-learning methods.

Relevance to ICES

Institutes such as ICES that provide research-based evidence need to make the most of their data and be capable of implementing state-of-the-art analytic methodologies that utilize the benefits of both structured and unstructured data. It is important for ICES scientists to have an understanding of the various types of research questions that have become addressable with new data science methodologies, as well as understand how these new research undertakings could benefit their field. It is likely that such education and ongoing work will require broadening the ICES community such that collaborations with computer scientists and data science experts are feasible. To support this new and additional research direction, it will also be important for ICES staff to learn about the theory and implementation of data science methodologies, as well as receive education on the computing platforms and programming languages required to hold and analyze these data.

Although researchers who use unstructured data, such as electronic medical records, may benefit from collaboration with an ICES-appointed scientist with a PhD in data science or computer science, it is premature to recommend this until educational training programs in data science are more developed, and research in electronic medical records has further matured.

Recommendations

- ICES should develop a staff education strategy with respect to data science. This could include, for example, participating in data science workshops for career development and learning modern data science tools, or hosting in-house workshops.
- ICES should increase capacity in data science methods by collaborating with external scientists having such expertise. Content areas include machine learning, neural networks, text mining and large-scale graphical visualization.
- ICES should consider engaging co-op students in computer science for internships, and potentially hiring computer scientists to carry out the routine elements of data wrangling, data linkage and data de-identification prior to analysis.

Summary Recommendations

The recommendations contained in this report can be classified into five broad themes: forging new partnerships, modernizing data integration, pursuing a data safe haven, exploiting in-house holdings of biomedical big data, and expanding expertise in data science methods.

New Partnerships

There are several areas where ICES does not have the expertise to pursue novel types of analyses. Investing resources in acquiring such expertise, whether in new scientists, software or equipment, is deemed premature as it is not yet clear where the most important data science opportunities will lie for health services research. In addition, this approach would be very costly. We believe that to expand our capabilities in data science, ICES should pursue partnerships with other scientists and institutions that have such expertise in order to learn from them. The partnerships would focus on topics of mutual interest. Through collaborative projects, ICES would offer access to its rich collection of linked health administrative data and its expertise in working with these data in return for scientific partnership. In general, we would need to reflect upon which research we wish to lead and which we wish to enable.

The universities of Toronto and Waterloo have computer science experts in machine learning and text mining. ICES scientists working with EMRALD data have partnered with computer scientists at the University of Toronto to create text-mining algorithms for specific conditions. The algorithms are purpose-built and require manual checking for validity. Computational expertise is required to adapt off-the-shelf text-mining software for medical records. The promise of data-rich electronic medical records is enormous, but before they can be used for research, bioinformatics expertise in natural language processing and text mining is needed to exploit their unstructured data. Historically, we have pursued these avenues on an ad hoc basis.

Similar collaborations could occur with research groups interested in linking genetic or biomarker data to ICES data since machine-learning methods might be useful for gene-association studies. Such studies require high-performance computing, so these collaborations could be fostered through the Data Safe Haven.

Data Integration

The ICES data quality framework has been articulated, but much work and considerable resources are needed for its implementation. Because data linkage is associated with the ICES brand, we aspire to be a leader in the application of modern methods for data linkage, including privacy-preserving record linkage and data anonymization, but at the moment, our reality falls short. ICES linkage is based on outdated probabilistic methods that rely heavily on manual intervention. Moving to modern methods will be resource and labor intensive; however, the overarching goal is to be an international leader in data quality methods.

While the ICES data quality framework formally encompasses database quality, research-specific quality is also important to our mission: research excellence resulting in trusted evidence that makes policy better, health care stronger and people healthier.[69] This requires assessing the validity of project-specific data elements, which can involve combining multiple sources of administrative data and applying numerous criteria to extract cohorts and quality indicators. The new data quality tools and the audit and oversight processes we have developed must be fully integrated into the project life cycle across the organization.

Data Safe Haven

A key organizational priority at ICES is to increase access to data for both ICES and non-ICES scientists, and we have made significant strides in achieving this in recent years with the creation of our Data and Analytic Services (DAS) unit. Existing approaches to data integration and access at ICES are not suitable for all researchers and data custodians. We see the need and the opportunity to create an infrastructure that will allow researchers to securely store and link research data, conduct advanced analytics, and provide for efficient, privacy-preserving data access. ICES is launching an initiative to pilot an Ontario Data Safe Haven that will act as a secure repository for existing research, administrative and, eventually, "omic" data sets, and a platform for data linkage, analysis and access. The data safe haven will be built in partnership with leaders in computer science at the University of Toronto and will leverage their existing high-performance computing environment, allowing it to advance the ICES data science agenda. The recommendation is that ICES pursue the plan to fund and implement the pilot Ontario Data Safe Haven proposal and, based on the results and on obtaining the necessary additional funding, consider fully developing it.

Biomedical Big Data

Rather than acquiring new biomedical big data, we recommend exploiting ICES' existing biomedical big data to build the case for resources needed to extract meaning from large, messy, structured data with deep clinical information (OLIS and BORN) and semi-structured data (electronic medical records) to make them research ready. While we have the necessary in-house expertise to tackle the clinical information, it will require substantial time and effort. For the semistructured data, we will need to partner with external experts to exploit the richness of the text data in electronic medical records. As part of a comprehensive EMR strategy, approaches to governance and privacy protection that are acceptable to stakeholders will need to be developed. We also recommend seeking additional funding to expand the primary care Electronic Medical Record Administrative data Linked Database (EMRALD) to ensure that it is representative of the general population and population-based.

Data Science Methods

ICES should begin by developing an educational strategy to identify gaps and opportunities to train ICES staff and scientists in modern data science methods, and acquiring appropriate statistical software to implement modern data science methods and data visualization techniques. Data science encompasses sophisticated, traditional statistical methods, novel machine-learning applications and data visualization techniques, among others. The combination of ICES scientists with expertise in biostatistical methods and ICES staff methodologists is ideal for identifying and disseminating traditional statistical methods where gaps exist. ICES currently does not have the required expertise to pursue novel machine-learning methods or data visualization techniques, and it would be premature to hire a scientist with this expertise until we have a clearer picture of how they would be useful.

Support for Ongoing Initiatives

There are existing initiatives at ICES that align well with our proposed data science agenda. We recommend that ICES continue to support the efforts of cross-provincial distributed data networks, such as the Canadian Network for Observational Drug Effect Studies and the Pan-Canadian Real-world Health Data Network, because such activities build research capacity and partnerships nationally.

References

1. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016; 13(6):350–9.

2. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)*. 2014; 33(7):1163–70.

3. National Institutes of Health. About BD2K. Accessed December 6, 2016 at **https://datascience.nih.gov/bd2k/about**.

4. National Institutes of Health. About the Precision Medicine Initiative. Accessed December 6, 2016 at **https://www.nih.gov/research-training/allofus-research-program**.

5. Vogelius M, Kannan N, Huo X. NSF Big Data Funding Opportunity for the Statistics Community. Accessed December 6, 2016 at **http://magazine.amstat.org/blog/2015/04/01/nsf-big-data/**.

6. Michigan Institute for Data Science, University of Michigan. Data Science Initiative. Accessed December 6, 2016 at **http://midas.umich.edu/dsi/**.

7. Cleveland WS. Data science: an action plan for expanding the technical areas in the field of statistics. *Int Stat Rev*. 2001; 69(1):21–6.

8. Steve's Machine Learning Blog. Data Science Venn Diagram v2.0. Accessed June 19, 2017 at **http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html**.

9. Donoho D. 50 Years of Data Science. Based on a presentation at the Tukey Centennial Workshop, Princeton, NJ, Sept. 18, 2015. Accessed June 1, 2016 at **http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf**.

10. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005; 2(8):e124.

11. Bernau C, Riester M, Boulesteix AL, et al. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*. 2014; 30(12):i105–i112.

12. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol*. 2013; 178(4):645–51.

13. Madigan D, Stang PE, Berlin JA, et al. A systematic statistical approach to evaluating evidence from observational studies. *Annu Rev Stat Appl*. 2014; 1:11–39.

14. IBM Big Data & Analytics Hub. The Four V's of Big Data. Accessed June 22, 2017 at **http://www.ibmbigdatahub.com/infographic/four-vs-big-data**.

15. National Institutes of Health. What is Big Data? Accessed December 7, 2016 at **https://datascience.nih.gov/bd2k/about/what**.

16. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA*. 2014; 311(24):2479–80.

17. Burton PR, Murtagh MJ, Boyd A, et al. Data Safe Havens in health research and healthcare. *Bioinformatics*. 2015; 31(20):3241–8.

18. Quintero D, Genovese WM, Kim K, et al. *IBM Software Defined Environment*. [n.p.]: IBM Redbooks; 2015. Accessed May 30, 2017 at **http://www.redbooks.ibm.com/redbooks/pdfs/sg248238.pdf**.

19. Azimaee M, Smith M, Lix L, Ostapyk T, Burchill C, Orr J. *MCHP Data Quality Framework*. Winnipeg, MB: Manitoba Centre for Health Policy; 2015. Accessed August 24, 2016 at **http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/protocol/media/Data_Quality_Framework.pdf**.

20. Lix LM, Smith M, Azimaee M, et al. *A Systematic Investigation of Manitoba's Provincial Laboratory Data*. Winnipeg, MB: Manitoba Centre for Health Policy; 2012. Accessed August 24, 2016 at **http://mchp-appserv.cpe.umanitoba.ca/reference/cadham_report_WEB.pdf**.

21. Canadian Institute for Health Information. *The CIHI Data Quality Framework*. Ottawa, ON: CIHI; 2009. Accessed August 24, 2016 at **https://www.cihi.ca/en/data_quality_framework_2009_en.pdf.**

22. Public Health Agency of Canada. *PHAC Data Quality Framework*. Ottawa, ON: PHAC; 2009.

23. Australian Bureau of Statistics. ABS Data Quality Framework, May 2009. Accessed August 24, 2016 at **http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0**.

24. NHS Information Centre. *Data Quality Report for Independent Sector NHS Funded Treatment, Q1-Q2, 2007/08*. Leeds, England: NHS Information Centre; 2008. Accessed August 24, 2016 at **http://collections.europarchive.org/tna/20081211191026/http://www.ic.nhs.uk/webfiles/Key%208/IS%20information/IS%20DQ%20report%20Q2_07_08v1.3.pdf**.

25. Herzog TN, Scheuren FJ, Winkler WE. *Data Quality and Record Linkage Techniques*. New York, NY: Springer-Verlag; 2007.

26. Wilson DR. Beyond probabilistic record linkage: using neural networks and complex features to improve genealogical record linkage. Proceedings of International Joint Conference on Neural Networks, San Jose, CA, July 31-Aug 5, 2011. Accessed August 24, 2016 at **http://axon.cs.byu.edu/~randy/pubs/wilson.ijcnn2011.beyondprl.pdf**.

27. Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001; 16(3):199–231.

28. Harrell FE, Jr. *Regression Modeling Strategies*. New York, NY: Springer-Verlag; 2001.

29. Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. West Sussex, UK: Wiley; 2008.

30. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.

31. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. New York, NY: Taylor & Francis;1984.

32. Loh W. Classification and regression trees. *WIREs Data Mining Knowl Discov*. 2011; 1(1):14–23.

33. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Saitta L, editor. *Machine Learning: Proceedings of the Thirteenth International Conference, Bari, Italy, July 3–6, 1996*. San Francisco, CA: Morgan Kauffman; 1996.

34. Breiman L. Random forests. *Mach Learn*. 2001; 45(1):5–32.

35. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer-Verlag; 2001.

36. Buhlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci*. 2007; 22(4):477–505.

37. Bottle A, Gaudoin R, Goudie R, Jones S, Aylin P. Can valid and practical risk-prediction or casemix adjustment models, including adjustment for comorbidity, be generated from English hospital administrative data (Hospital Episode Statistics)? A national observational study. *Health Serv Deliv Res*. 2014; 2(40).

38. Tu JV, Weinstein MC, McNeil BJ, Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? The Steering Committee of the Cardiac Care Network of Ontario. *Med Decis Making*. 1998; 18(2):229–35.

39. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Stat Med*. 2007; 26(15):2937–57.

40. Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? *Biom J*. 2012; 54(5):657–73.

41. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol*. 2013; 66(4):398–407.

42. Bayati M, Braverman M, Gillam M, et al. Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One*. 2014; 9(10):e109264.

43. Hao S, Jin B, Shin AY, et al. Risk prediction of emergency department revisit 30 days post discharge: a prospective study. *PLoS One*. 2014; 9(11):e112944.

44. Melillo P, Izzo R, Orrico A, et al. Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PLoS One*. 2015; 10(3):e0118504.

45. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care*. 2010; 48(11):981-8.

46. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011; 306(8):848–55.

47. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One*. 2011; 6(8):e23610.

48. Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA*. 2007; 297(14):1551–61.

49. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA*. 2003; 290(19):2581–7.

50. Rosella LC, Manuel DG, Burchill C, Stukel TA. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). *J Epidemiol Community Health*. 2011; 65(7):613–20.

51. Lee DS, Stitt A, Austin PC, et al. Prediction of heart failure mortality in emergent care: a cohort study. *Ann Intern Med*. 2012; 156(11):767–75.

52. Lee DS, Hardy J, Yee R, et al. Clinical risk stratification for primary prevention implantable cardioverter defibrillators. *Circ Heart Fail*. 2015; 8(5):927–37.

53. Atzema CL, Dorian P, Fang J, et al. A clinical decision instrument for 30-day death after an emergency department visit for atrial fibrillation: the Atrial Fibrillation in the Emergency Room (AFTER) Study. *Ann Emerg Med*. 2015; 66(6):658-68.

54. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.

55. Yau N. *Data Points: Visualization That Means Something*. Indianapolis, IN: Wiley; 2013.

56. Few S. Graph selection matrix. Accessed August 24, 2016 at **https://www.perceptualedge.com/articles/misc/Graph_Selection_Matrix.pdf**.

57. Tufte ER. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press; 1983.

58. Yau N. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. New York, NY: Wiley; 2011.

59. Wolfson M, Wallace SE, Masca N, et al. DataSHIELD: resolving a conflict in contemporary bioscience – performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*. 2010; 39(5):1372–82.

60. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative – a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016; 99(3):265–8.

61. Mini-Sentinel. Sentinel Common Data Model v6.01. Accessed July 31, 2017 at **https://www.sentinelinitiative.org/rss/sentinel-common-data-model-v601**.

62. Mini-Sentinel. Routine Querying System. Accessed July 31, 2017 at **https://www.sentinelinitiative.org/sentinel/surveillance-tools/routine-querying-tools/routine-querying-system**.

63. Xu Y, Zhou X, Suehs BT, et al. A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: implications for active drug safety surveillance. *Drug Saf*. 2015; 38(8):749–65.

64. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff (Millwood)*. 2014; 33(7):1178–86.

65. Suissa S, Henry D, Caetano P, et al. CNODES: the Canadian Network for Observational Drug Effect Studies. *Open Med*. 2012; 6(4):e134-e140.

66. Stang P, Ryan P, Hartzema AG, et al. Development and evaluation of infrastructure and analytic methods for systematic drug safety surveillance: lessons and resources from the Observational Medical Outcomes Partnership. In: Andrews EB, Moore N, editors. *Mann's Pharmacovigilance*. Oxford, UK: Wiley; 2014.

67. Stukel TA, Demidenko E, Dykes J, Karagas MR. Two-stage methods for the analysis of pooled data. *Stat Med*. 2001; 20(14):2115–30.

68. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. *Med Care*. 2013; 51(8 Suppl 3):S4–10.

69. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010; 153(9):600–6.

70. Abbing-Karahagopian V, Kurz X, de Vries F, et al. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. *Curr Clin Pharmacol*. 2014; 9(2):130–8.

71. Institute for Clinical Evaluative Sciences. *Data, Discovery, Better Health: ICES Strategic Road Map, 2014/15–2016/17*. Toronto, ON: ICES; 2014. Accessed December 8, 2016 at **http://www.ices.on.ca/~/media/Files/Corporate-Reports/ICES%20StrategicPlan_web_jp01.ashx**.

72. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc*. 1969; 64(328):1183–210.

73. Artificial Intelligence Blog. Data mining. Accessed July 10, 2017 at **https://www.artificial-intelligence.blog/terminology/data-mining**.

74. Wikipedia contributors. Data visualization. Wikipedia, The Free Encyclopedia; July 1, 2017. Accessed July 10, 2017 at **https://en.wikipedia.org/w/index.php?title=Data_visualization&oldid=788368102**.

75. Jurafsky D, Martin JH. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2008.

76. Williams JK, Ahijevych D, Blackburn G, Craig J, Meymaris G. Statistical Learning Methods for Big Data Analysis and Predictive Algorithm Development. Presented at SEA Software Engineering Conference, Boulder, CO, April 2013. Accessed July 10, 2017 at **https://sea.ucar.edu/sites/default/files/StatLearnBigData20130401.pdf**.

77. Wikipedia contributors. Text mining. Wikipedia, The Free Encyclopedia; July 7, 2017. Accessed July 10, 2017 at **https://en.wikipedia.org/w/index.php?title=Text_mining&oldid=789423970**.

78. Wikipedia contributors. Unstructured data. Wikipedia, The Free Encyclopedia; July 4, 2017. Accessed July 10, 2017 at **https://en.wikipedia.org/w/index.php?title=Unstructured_data&oldid=788870782**.

Appendices

A.1 Glossary

Data mining	An interdisciplinary subfield of computer science. The overall goal of data mining is to extract information from a data set and transform it into an understandable structure for further use.[73]
Data visualization	Aims to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Effective visualization helps users to analyze data and evidence, and makes complex data more accessible, understandable and usable.[74]
Natural language processing	The technology for creating usable data from human language as it appears in emails, web pages, product descriptions, newspaper stories, social media and scientific articles, in thousands of languages and varieties.[75]
Statistical learning	A collection of automated or semi-automated techniques for discovering previously unknown patterns in data, including relationships that can be used for prediction.[76]
Structured data	Data that are rectangular regardless of the number of observations. They can be easily organized and are usually stored in databases.
Text analytics	The process of deriving high-quality information from text, which is a form of unstructured data. High-quality information is typically obtained by deriving patterns and trends through statistical pattern learning. Text analytics usually involves the process of structuring the input text and deriving patterns within the structured data.[77]
Unstructured data	Information that either does not have a pre-defined data model or is not organized in a pre-defined manner.[78] It includes data arising from emails, word processing files, blogs, online forums, survey responses, digital images, video and audio files, and social network feeds.

A.2 Biomedical Big Data Sources in Ontario

Biomedical Big Data Platforms

Ontario Brain Institute

- Brain-CODE: Neuro-informatics platform for data management, sharing and analysis
- Areas of focus: epilepsy, cerebral palsy, neurodegenerative disease, depression and neurodevelopmental diseases, with a goal of "deep" data on small cohorts of 5,000+ patients.
- Registries with encrypted OHIP number, clinical assessments, neuroimaging and omics (genomics and proteomics) data.
- Working with ICES on a privacy-preserving protocol (homomorphic encryption) that would allow integration of ICES data with one clinical registry (pediatric epilepsy patients on ketogenic diets).

Indoc Research

- Evolved from the Ontario Cancer Biomarker Network (OCBN).
- OCBN was established in 2005 with support from the Ontario Institute for Cancer Research and the Ministry of Research and Innovation with a mandate to coordinate and amplify the proteomic and genomic biomarker research efforts of its academic and industry partners.
- OCBN expanded its operations to support a broad range of diseases, molecular research technologies, and big data informatics needs, and helped develop Brain-CODE.

Ontario Health Study

- An integrated platform to investigate the complex interplay of environmental, lifestyle and genetic factors that increase individual and community risk of developing cancer, heart disease, diabetes, asthma, depression and other common adult diseases.
- Data include questionnaires, physical measurements and biological samples on a subset of the sample.
- One of the largest health studies in Canada's history and in the top 8% by size of all such studies worldwide. The study has recruited and is engaged with approximately 225,000 participants in Ontario.
- The intent is to follow study participants for their entire adult lifespan through the use of annual online questionnaires and linkage to administrative health data at ICES;
- Study participants are also included in a pan-Canadian initiative (the Canadian Partnership for Tomorrow Project).
- Funded by four organizations: the Ontario Institute for Cancer Research, Cancer Care Ontario, Public Health Ontario and the Canadian Partnership Against Cancer.

Electronic Medical/Health Records Data

PRIMARY CARE

EMRALD (Electronic Medical Record Administrative Data Linked Database)

- Full EMR data from one specific vendor is in-house at ICES from participating practices interested in audit and feedback, with secondary use of the data for research.
- Includes more than 42 clinics, 375 physicians and over 500,000 patients.
- Includes structured data (blood pressure, body mass index) and unstructured data (notes, letters, laboratory reports).

University of Toronto Practice-Based Research Network (UTOPIAN)

- Proposal to develop an Ontario primary care EMR database that will provide a data extraction and analysis service.
- Based initially on the family physician practices participating in UTOPIAN but will be expanded to include other parts of Ontario and other health structures such as Community Health Centres.
- Funding unclear.

HOSPITAL DATA WAREHOUSES

The Ottawa Hospital

- Supports a well-developed hospital electronic health record that has been used for research and linked with ICES data.
- Includes structured and unstructured digital data.

St. Michael's Hospital, Toronto

- Proposal to create an enterprise data warehouse that will integrate St. Michael's clinical and administrative databases.
  - The goal is to "transform data repositories into a comprehensive data warehouse that can be exploited using standardized data extraction and analysis algorithms based on sound scientific principles."
  - Information stored in the enterprise data warehouse will be leveraged for a wide variety of uses, including informing decision-making by senior management, informing clinical quality improvement programs, identifying opportunities to improve the cost-effectiveness of health care services, and facilitating high-quality, clinically relevant research.

OTHER SELECTED BIG DATA COHORTS

The following are investigator-led efforts to develop cohorts capable of answering multiple questions by a larger research community; they collect big data (questionnaires, biomarkers, physical measurements) and have explored linkages to administrative health data at ICES.

TARGet Kids!

- University of Toronto investigator-led research platform studying a cohort of more than 8,000 school-aged children recruited from primary care practices.
- Focus is on healthy weight, nutrition and child development.
- Questionnaire, physical measurements and biomarkers; no "omics" yet but these are under discussion.
- Received consent to link to ICES administrative data; one pilot project is underway and there are ongoing discussions regarding an umbrella data sharing agreement.

Ontario Birth Study

- University of Toronto investigator-led research platform studying a cohort of pregnant mothers (Mount Sinai Hospital with expansion to other hospitals).
- Collection of lifestyle and diet questionnaires plus additional biologic samples at the time of routine care.

- The study plans to follow infants and children through the TARGet Kids! Collaboration.
- Funding is unclear.

PROVINCIAL CLINICAL DATA REPOSITORIES

Ontario Laboratory Information System (OLIS)

- A system that connects hospitals, community laboratories, public health laboratories and practitioners to facilitate the secure electronic exchange of laboratory test orders and results.
- Covers nearly 80% of the annual provincial laboratory test volumes.
- Executed a data sharing agreement with ICES; data arrived at ICES in January 2016.
- Will build on local work in southwestern Ontario using Gamma Dynacare laboratory data.

Better Outcomes Registry and Network (BORN Ontario)

- Established in 2009 as a Prescribed Registry under Ontario's Personal Health Information Protection Act, 2004. Some data go back to 2006.
- BORN is a population-based registry of all pregnancies ending in birth, as well as data on all women who undergo prenatal screening irrespective of whether pregnancy continues to birth
- Prenatal screening data include results of ultrasounds and a variety of serum markers used to screen for genetic and developmental anomalies.
- Newborn Screening Ontario data include the laboratory values (a variety of metabolites, enzymes, hormone levels, etc.) of all live births screened for 29 [current] rare but severe metabolic, endocrine and genetic diseases; these include hypothyroidism, cystic fibrosis, sickle cell disease, severe combined immune deficiency and others.
- Other clinical and demographic data are available for the mothers and infants.
- ICES' current umbrella data sharing agreement with BORN includes most of the BORN data. The new BORN Information System is a relational data set.
- Fertility clinic data housed at BORN are not part of the current data sharing agreement with ICES.

A.3 Data Integration

Data quality

Population-based health administrative data are collected for billing purposes but can be used for health services and health policy research and for monitoring the health care system. It is essential to assess their quality by evaluating completeness, consistency and accuracy. For many administrative data repositories, updates to existing data and new data are continually being received. An automated quality assessment process enables efficient and timely comparisons of the quality of data housed in a single repository over time and between populations, and also facilitates comparisons among repositories, which is necessary to ensure comparability of research across multiple jurisdictions.

Data strategy

Requests for any major data acquisition are discussed and evaluated by ICES' Data Integration and Strategy Committee. The committee, which includes representatives from different departments, considers the value of the data request for ICES research as well as the technical, financial and privacy-related risks and concerns involved in adding the data to the ICES repository. With the committee's approval, an official request is made and negotiations commence with the data custodian, culminating in a data sharing agreement.

Data governance

Understanding the data structure, designing the required infrastructure for holding the data, data conversion, deterministic and probabilistic record linkage, data anonymization, data standardization, building rich metadata, assessing data quality, and applying necessary security and access controls are high-level steps for ICES data governance. Some of these steps, such as building rich metadata and assessing data quality, have their own detailed processes.

Data delivery

When the data are research-ready and all required documentation has been completed, the data are made available. At this stage, critical information is provided to data users. Depending on its complexity, this information can be delivered in person through presentations, rounds and staff meetings or via electronic channels such as blogs and email.

Data anonymization

In the ICES data repository, the primary unique identifier for individuals is the Ontario Health Insurance Plan (OHIP) number, which after being uniquely coded (scrambled or encrypted) is called the ICES key number (IKN). Currently, OHIP numbers are coded using a program written in Fortran. The application does not have an interface and must be called in batch mode, and requires some data preparation. The program can encode fields with only a certain number of digits (other than the OHIP number, there are sometimes other types of sensitive numbers that must be coded). This application does not permit the coding process to be easily reversed, which is required when releasing data back to the data custodians. Because Fortran is an older programming language, it is difficult to find skilled programmers to maintain it.

Record linkage

Record linkage is the process of connecting the same entity (individual, patient, physician, institution) across multiple sources. This could occur by finding an exact match for all identifiers (first name, last name, sex, postal code, date of birth, etc.), which is defined as deterministic record linkage. The idea of considering probable matches by calculating weights for each pair is the basis for traditional probabilistic record linkage, as modeled by Fellegi and Sunter.[70] ICES uses AutoMatch software to implement probabilistic record linkage. Statistics Canada's G-Link software uses Fellegi-Sunter theory with additional matching techniques and is currently being tested for implementation at ICES. Recently, a variety of machine-learning and fuzzy matching techniques for record linkage have been developed that are likely superior to traditional methods.

Database-specific quality

Based on the data quality framework, six quality dimensions are applied to each ICES data holding.

- **Accuracy**. This dimension evaluates completeness and correctness of the data. These components are measured by percentage of valid, invalid, missing and outlier values in each data element. The pattern of missing values over time for each data element reveals coding changes or discontinuation of some fields.

- **Internal validity**. This dimension consists of internal consistency, stability over time and linkability. Internal consistency is measured by logical relationships between fields; for example, a 70-year-old woman would not have a baby, a man would not be scheduled for a Caesarean section, and a four-year-old child would not have an occupation. Internal consistency could also be measured by numeric agreement between the fields using correlations or Kappa statistics. To assess stability over time, a line or curve is fitted to the number of the records over time and unusual observations are identified. Repeated observations with the same value are flagged as a potential data quality issue. For linkability, the linkage rate by type of linkage (deterministic or probabilistic) over a date variable (fiscal year) is presented.

- **External validity**. For this dimension, a value from the database is compared with a value from another source of information; this could be another database available at ICES or a published report or online source.

- **Timeliness**. Timeliness refers to currency of the data. Indicators of timeliness include time to acquisition, time to release and recency of the data. Each of these corresponds, respectively, to the number of days between:

  - the date the data sharing agreement was executed and the date the database was acquired by ICES

  - the date the database was acquired by ICES and the date the database was made available in the ICES data repository for research

  - the last reference date in the database and the date the file was made available in the repository.

- **Interpretability**. This dimension focuses on the availability, quality and usability of data documentation including metadata, the data dictionary, data quality reports, data collection methods, data diagrams and survey forms.

- **Relevance**. This dimension describes the usability of data at ICES by reporting the number of ICES' research projects using the data in a calendar/fiscal year, the number of times the data was accessed through ICES's Research Analytical Environment, and the number of publications that included the data.

Automation

Considering the growing number of data holdings in the ICES data repository and the frequent updates, this process should require a minimum amount of manual interference. For this reason, most data quality reports are created using automated, SAS-based tools. Currently, we have a suite of 26 SAS macros with which to apply the data quality framework to ICES data holdings.

To ensure the automated tools work efficiently, a significant amount of work was allocated to apply a set of standardization rules to all ICES data. One example is the adoption of standard naming conventions and the same length and type for major common variables across different data holdings.

Research-specific quality

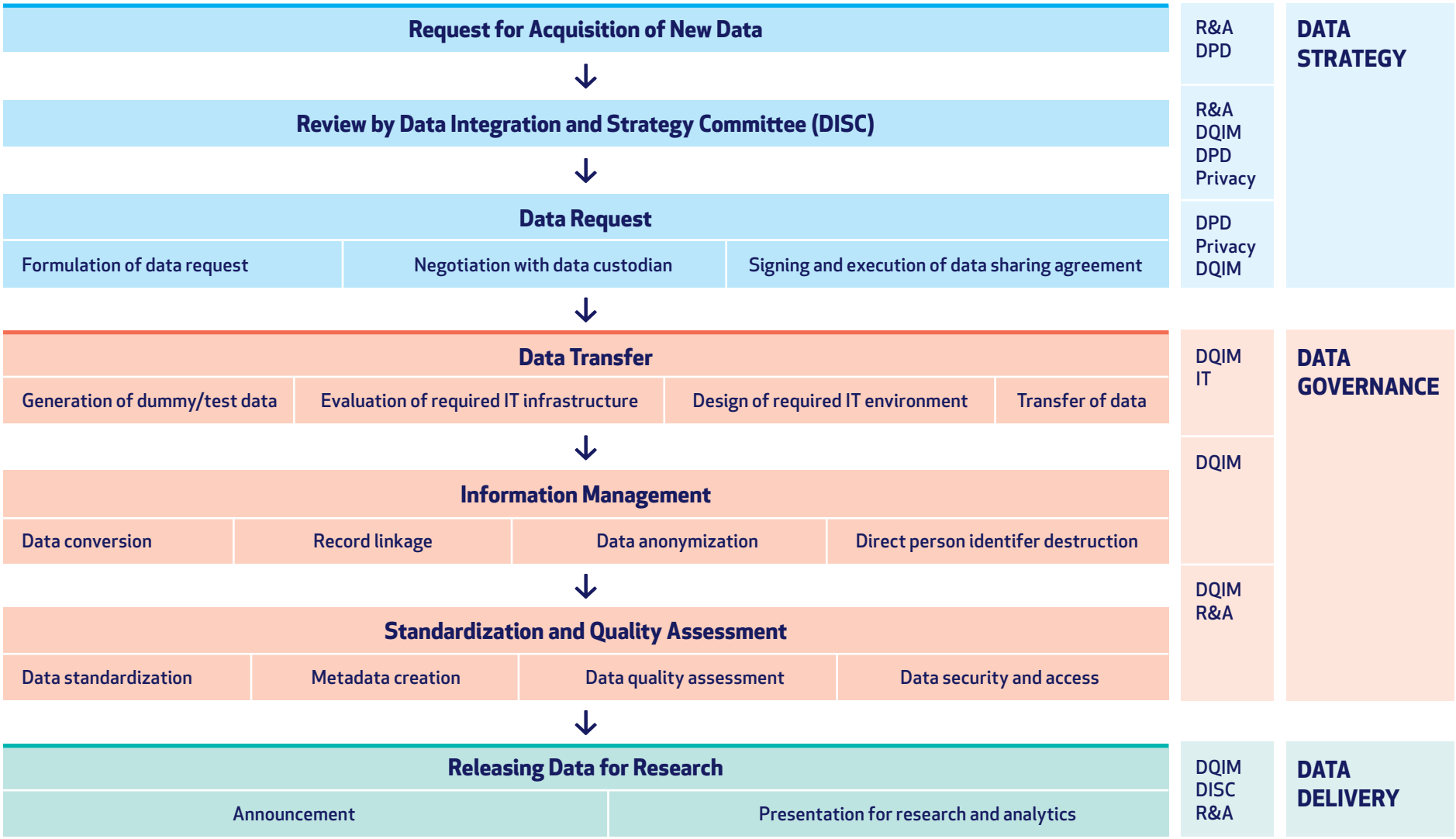
Although the dimensions of research-specific quality may change among projects, this framework recommends two dimensions of accuracy and validity. Accuracy involves completeness, correctness and measurement error, whereas validity includes agreement with other data, internal consistency, stability over time and systematic linkage biases. These assessments are expected to be performed by research project teams on their cohorts; however, the ICES Data Quality and Information Management team has created a user-friendly SAS macro that enables researchers to generate a "lite" version of the standard data quality reports that is generated for data holdings.

Metadata

To evaluate correctness and ensure interpretability, creating and maintaining a rich metadata repository is required. At ICES, we expend a fair amount of effort on enriching our data with as much metadata as possible. This includes adding labels to data sets and data elements and creating value labels that are stored in a SAS Format Catalog; the catalog currently includes more than four million value labels. A metadata data set includes detailed information about the entire ICES data repository. These two sources of information are used to assess the quality of data holdings and to create the ICES data dictionary.

Over time, we have identified secondary uses for this rich metadata. These include providing new tools for generating metrics on ICES data holdings (including file size in megabytes or gigabytes or numbers of observations) and type of access, and for searching for specific keywords in the ICES data repository. Using the latter tool, one could run a search to find which data holdings include information on "homelessness," for example.

EXHIBIT A.1 ICES data integration framework



Abbreviations: DPD: Data Partnerships and Development; DQIM: Data Quality and Information Management; IT: Information Technology; R&A: Research and Analysis

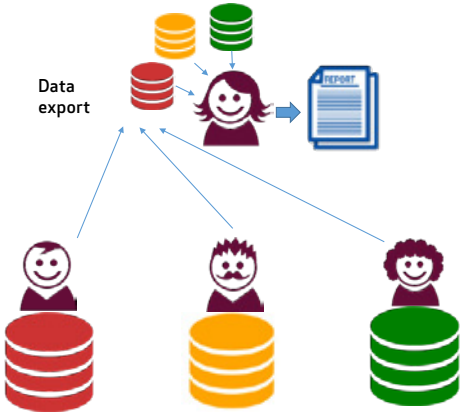
EXHIBIT A.2 ICES data quality framework

Administrative Data Quality							
Database-Specific Quality						Research-Specific Quality	
Accuracy	Internal Validity	External Validity	Timeliness	Interpretability	Relevance	Accuracy	Validity
• Completeness • Correctness	• Internal consistency • Stability over time • Linkability	• Agreement with other data • Agreement with external reports (In development)	• Time to acquisition • Time to release • Recency of the data	• Data documentation - Availability - Quality - Usability	• Usability of data - By projects - By access - By published papers (In development)	• Completeness • Correctness • Measurement error	• Agreement with other data • Internal consistency • Stability over time • Systematic linkage bias
• VIMO macro • TIM macro	• TREND macro • Linkability macro	• Agreement macro (In development)	• Timeliness macro (In development)	• Meta macro • Dictionary macro		• DQ Lite macro	

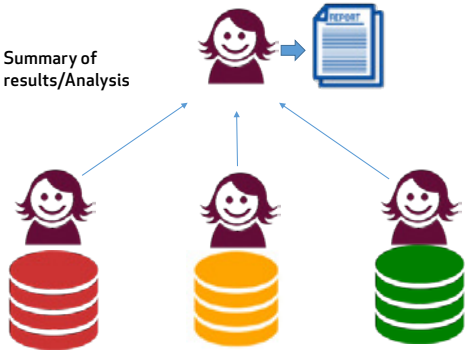
EXHIBIT A.3 Approaches to multi-centre data analysis

[Reproduced with permission from Simon Thompson, SAIL DataBank, Swansea University.]

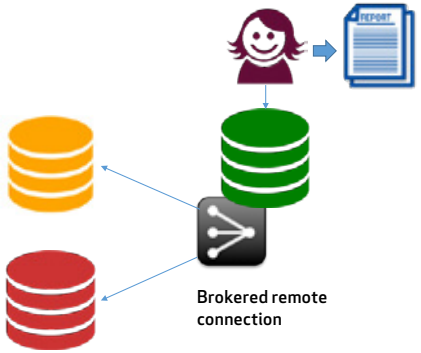
1. **Data moved from 3 centres** – 1 analyst (centralized data model)



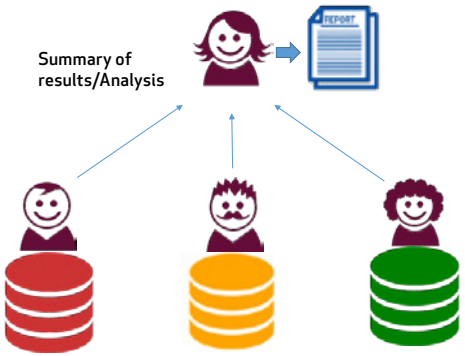
2. **Data at 3 centres** – 1 analyst using each platform, then combining results



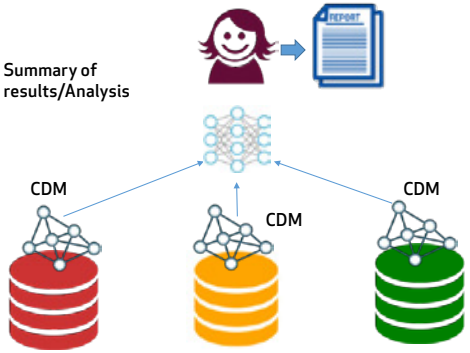
3. **Data at 3 centres** – 1 analyst (remote real-time access model)



4. **Data at 3 centres** – 3 separate analyses (standard replication)



5. **Data at 3 centres** – 1 analyst directing federated queries (DataSHIELD model)



Data
Discovery
Better Health

Institute for Clinical Evaluative Sciences
G1 06, 2075 Bayview Avenue
Toronto, Ontario M4N 3M5
www.ices.on.ca

ICES