Artificial intelligence or manufactured stupidity? The need for injury informaticians in the big data era

Kirsten Vallmuur

ABSTRACT
The volume, velocity and variety of data collected about individuals have increased exponentially over the last decade, presenting new injury surveillance opportunities to identify risk factors, monitor trends, and evaluate the efficacy of interventions. But does the hype around big data and artificial intelligence (AI) apply to the injury prevention space, and how veracious is surveillance in this era? This commentary discusses the digital transformation of health as applied to injury prevention, but cautions on the challenges of maintaining data quality in integrated systems and discusses the need for an injury informatics strategy moving forward.

INTRODUCTION
Even before the significant digital transformations we are currently witnessing in the health sector, there was no denying that data were a critical foundation of clinical and public health decision-making around injury management and prevention. The basic framework of the public health approach and injury surveillance relies on data for defining the problem and identifying risk/protective factors, drawing on a plethora of data sources including hospital, coronial and vital records, emergency services data, population surveys, specialised injury surveillance systems and so on. Injury prevention programmes and policy development rely on data to identify target areas, design interventions, evaluate programme efficacy and monitor the impact of changes. Trauma clinical improvement programmes use these data to identify at-risk patient cohorts, quantify treatment efficacy, track patient outcomes and enumerate quality of life measures. Traditional approaches to collection of such injury/trauma data are labour-intensive, requiring significant ethics and governance approval processes, lengthy delays to data provision, and sometimes even manual review methods for complete capture of cases. These traditional processes capture data that are often several years out-of-date by the time they are available for analysis due to the data processing, cleaning, verification and authorisation processes, and are often lacking in detail due to masking of key data fields which provide richer contextual information. Ironically, the challenges and complexities of data acquisition in the traditional (or pre-'big data') era exposed the benefit of the knowledge gained through opening data and a rhetoric around the economic and social benefits of using secondary public sector data, which will move the paradigm from one which restricts access to identifiable data to one which authorises release when appropriate data safeguards are in place. The whole of government initiatives, such as the Data Integration Partnership for Australia, is aimed at enabling the establishment of core data infrastructure and integration, developing legislation and governance protocols, preserving data privacy and security, and expanding federal data analytic capabilities and tools. Large-scale data linkage demonstration projects to promulgate this open data value proposition and demonstrate the benefit of the knowledge gained through opening up access to population data resources are being undertaken by the leading statistics agencies (eg, the Australian Bureau of Statistics-led Multi-Agency Data Integration Project and the Australian Institute of Health and Welfare-led National Integrated Health Services Information Analysis Asset). This is further enabled by the increasing sophistication of biostatistical techniques, improvement in privacy-preserving technologies, exponential growth in computational capacity, and availability of cost-effective, cloud-based technological infrastructure to enable storage, manipulation and complex analytics to be performed on these ‘big data’.

However, as the hype around these concepts of big data, data lakes, AI, digital health records, precision medicine/precision public health, Internet of Things (IoT) and real-time surveillance grows, the expectations of stakeholders in regard to the availability, timeliness and accuracy of data are now far

© Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Vallmuur K. Inj Prev 2020; 26:400–401. doi:10.1136/injuryprev-2019-043393

Special feature
exceeding the maturity of the systems. The increasing illusion of there now being vast quantities of high-quality data to rapidly respond to emerging issues is concerning, particularly as the data-as-a-commodity model becomes more mainstream. AI approaches lure many a stakeholder with the promise of new insights which will enable decision support in real time with minimal user input. AI techniques are only valuable insofar as they are based on quality interpretable data drawn from reliable semantically interoperable source systems, and where the AI algorithms can be made transparent and translated to clinicians/practitioners for interpretation and evaluation to verify the validity before they are likely to trusted and changes before implementation. A recent report by Future Advocacy consulted with experts in AI worldwide about the ethical, social and political challenges of AI in the health domain, and a recurring theme of these discussions was that of ‘garbage in, garbage out’, and the need for quality data on which to base any AI techniques.

The defining characteristic of quality data is that it is ‘fit for purpose’, yet as data get further from the raw source through the extraction, transformation and loading phases and without sufficient understanding of the processes for collection and context within which the data are captured and coded, that purpose can be far removed from the initial intention. For example, in the context of data warehouses where the premise is that data are collected once and reused for multiple purposes, how do we determine which purpose we are evaluating the data against? These data warehouses integrate huge amounts of data from multiple systems captured in structured and unstructured formats at various stages of the healthcare pathway, and are aimed at servicing multiple user types including clinical, research, administration and other government department users. Yet so little effort is currently invested in the testing and validation of core data points comprising these systems, or liaison with the clinicians/practitioners to ensure the accuracy of data definitions and precision of selected data points, or in verifying that the interpretation of processes and indicators is appropriate for the context. As stated on the Australian Digital Health Agency’s Digital Health Strategy website, ‘It is imperative that when information is shared between people and systems, its meaning is preserved from one context to another so that information is interpreted in the same way. That is, what was meant is the same as what is understood’.

Understanding the meaning of complex health data and the processes with which they are collected in order to interpret these data accurately and draw meaningful conclusions requires specialist domain and technical knowledge. Health informaticians are at the nexus of this process in these big data lake/warehouse/linkage projects, and are critical roles in increasingly digitised health ecosystem. The health informaticians are the general practitioners of the data world. They need to recognise poor data ‘symptoms’, identify the appropriate specialist required to address these symptoms, liaise between service providers and monitor the ongoing ‘health’ of the data system. These roles are required to have a generalist knowledge of the clinical, technical, research, administration and policy domains. They need to be able to translate the data needs/requirements of the clinicians/practitioners to the technical teams building the data lakes/warehouses and interpret and communicate the technical challenges/queries to the stakeholders, while still understanding the needs of researchers and policymakers in the application of the data to broader practice. There is a deficit of people with sufficient understanding, experience and qualifications in this domain, and there is a call for more investment and capacity building to grow the digital-health-literate data scientist to begin realising the potential benefits of these big data platforms.

The injury prevention community has much to gain from this exponential growth in integrated open data. Injury prevention is an ideal use case to demonstrate the need for integrated data sources across public and private sector agencies. Using Haddon’s Matrix as a framework for understanding the information requirements for injury prevention, we can easily demonstrate the importance of integrating data from the pre-event environment (eg, main roads data regarding road conditions, mental health encounter data, child safety department records) with event data (eg, ambulance records, police crash data, hospital data) and postevent data (eg, vital statistics collections, coronial data, rehabilitation, return to work data, compensation data and so on). Injury prevention practitioners are experts at dissecting injury events into these component parts and understanding the complexities of injury circumstances and comprehensiveness of information needs, as well as appreciating the importance of quality data for evidence-based decision-making. The injury prevention community needs to be more engaged in the digital health big data platform developments and look for avenues to build a workforce of ‘injury informaticians’ to capitalise on the opportunities which are presenting themselves for much richer timely information sources to inform policy and prevention. Data are a powerful commodity, and timely rich quality data are politically persuasive. As we rally together to develop national and international strategies for injury prevention, we must not overlook the need for an agile, future-focused capacity building injury informatics strategy.

Contributors KV wrote the entire submission.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iD Kirsten Vallmuur http://orcid.org/0000-0002-3760-0822

REFERENCES
1 The Australian digital health Agency. Australia’s national digital health strategy. Available: https://conversation.digitalhealth.gov.au/australias-national-digital-health-strategy [Accessed Aug 2019].
2 Dolley S. Big data’s role in precision public health. Front Public Health 2018;6.
3 Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J 2015;14.
4 Bunker D. Going digital. The Health Advocate 2019;55.
5 Department of Prime Minister and Cabinet. Data sharing and release reforms. Available: https://www.pm.gov.au/public-data/data-sharing-and-release-reforms [Accessed Aug 2019].
6 Department of Prime Minister and Cabinet. Data integration partnership Australia. Available: https://www.pm.gov.au/public-data/data-integration-partnership-australia [Accessed Aug 2019].
7 Australian Bureau of statistics. Multi-Agency data integration project (MADIP). Available: https://www.abs.gov.au/ausstats/absweb/dbs/D3310114.nsf/home/Statistical+Data+Integration++MADIP [Accessed Aug 2019].
8 Futures Advocacy. Ethical, social, and political challenges of artificial intelligence in health. 2018. Available: https://wellicome.ac.uk/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf [Accessed Aug 2019].
9 The Australian digital health Agency. Interoperability and data quality. Available: https://conversation.digitalhealth.gov.au/interoperability-and-data-quality [Accessed Aug 2019].
10 Healthcare IT. Building a digital-literate workforce in healthcare. Available: https://www.healthcareit.com.au/article/building-digital-literate-workforce-healthcare [Accessed Aug 2019].
11 Lett R, Kobusingye Q, Sethi D. A unified framework for injury control: the public health approach and Haddon’s matrix combined. Inj Control Saf Promot 2002;9:199–205.