Denaxas et al. BioData Mining (2016) 9:29
DOI 10.1186/s13040-016-0109-1

BioData Mining

EDITORIAL Open Access

The tip of the iceberg: challenges of accessing hospital electronic health record data for biological data mining

Spiros C. Denaxas[1,2], Folkert W. Asselbergs[1,2,3] and Jason H. Moore[4*]

* Correspondence: jhmoore@upenn.edu
[4]Institute for Biomedical Informatics, Department of Biostatistics and Epidemiology, Perelman School or Medicine, University of Pennsylvania, Philadelphia, PA 19104-6116, USA
Full list of author information is available at the end of the article

Abstract

Modern cohort studies include self-reported measures on disease, behavior and lifestyle, sensor-based observations from mobile phones and wearables, and rich -omics data. Follow-up is often achieved through electronic health record (EHR) linkages across primary and secondary healthcare providers. Historically however, researchers typically only get to see the tip of the iceberg: coded administrative data relating to healthcare claims which mainly record billable diagnoses and procedures. The rich data generated during the clinical pathway remain submerged and inaccessible. While some institutions and initiatives have made good progress in unlocking such deep phenotypic data within their institutional realms, access at scale still remains challenging. Here we outline and discuss the main technical and social challenges associated with accessing these data for data mining and hauling the entire iceberg.

In January 2015, President Barack Obama launched the Precision Medicine Initiative [1], a $215-million investment aiming to facilitate data-driven precision research by forging a cohort of at least one million participants. Primary data collection includes self-reported measures on disease, behavior and lifestyle, sensor-based observations from mobile phones and wearables, and rich -omics data. Follow-up will be achieved through electronic health record (EHR) linkages across primary and secondary healthcare providers. Historically however, researchers typically only get to see the tip of the iceberg: coded administrative data relating to healthcare claims which mainly record billable diagnoses and procedures. The rich data generated during the clinical pathway [2] (e.g. laboratory measurements, investigations, clinical notes, imaging, medications) remain submerged and inaccessible. While some institutions and initiatives [3–6] have made good progress in unlocking such deep phenotypic data within their institutional realms, access at scale still remains challenging. Here we outline and discuss the main technical and social challenges associated with accessing these data for data mining and hauling the entire iceberg.

It is often said that the field of informatics consists of people and technology intertwined. It comes as no big surprise that the greatest challenges are observed around interacting with clinical informatics staff and information systems. Research is usually not directly within the remit of informatics departments whose primary role is to support patient care through the provision and maintenance of various platforms and

systems. This provision substantially varies between healthcare providers and across clinical specialties: providers might use a single unified EHR platform (e.g. Cerner, Epic) or a set of isolated platforms and systems integrated through bespoke middleware solutions. Often, these systems have been developed by subcontracted external software vendors which leads to substantial interaction costs when attempting to access data outside the standard clinical care use. In both cases however, it is usually the case that access to data for research has not been a key requirement and as a result the deployed platforms critically lack the functionality to facilitate it out of the box.

While the majority of secondary care clinical specialties generate electronic data, the manner in which data get captured and the context under which they are recorded differs. This results in a heterogeneous ecology of healthcare process models that even within a single provider are challenging to identify, integrate and re-use. It is often hard to get the "big picture" and discover the data flows between clinical departments and systems. The irregular utilization of metadata and health data standards makes it challenging to establish data provenance and assess data quality in a meaningful manner. More importantly, given the complexity of healthcare provision, it is difficult to establish the context under which data were generated and which is essentially required to enable the reuse of data for research. For example, the same piece of information, such as a blood pressure measurement or a white blood cell count, can be recorded across multiple systems but at differing temporal and clinical resolutions and in different contexts [7, 8].

Large amounts of information are also often stored in semi-structured or unstructured format. Biochemistry, haematology, microbiology and cellular pathology investigations and results are usually stored as semi-structured reports whose format varies significantly both within and between healthcare providers [9]. In some clinical specialties, such as mental health, the majority of information generated and recorded during interactions with clinical staff is stored as free-text [10]. Unstructured data are increasingly hard to access for research purposes and scalable natural language processing methods [11] and pipelines [12] are required in order to extract, clean and format these data at scale. Developing these tools however is equally difficult as access to large corpora of text which are required for algorithm training is restricted.

Data generated during clinical care are almost exclusively from unconsented patients which leads to ethical and governance challenges [13]. The reuse of such data for research requires a set of complex approvals from multiple governing entities which are challenging to navigate and obtain and operate in an opaque manner. Furthermore, significant concerns are often raised in terms of information security patient confidentiality and minimizing the risk of re-identification [14]. Researchers find themselves between a rock and a hard place. Research-driven environments offer substantially more flexibility in terms of analyzing the data such as for example through the provision of high performance clusters or flexible technology stacks that enable the development and evaluation of novel computational methods and approaches. At the same time, they are considered poorly in terms of information security and governance from healthcare providers who are reluctant to release data for storage there in large numbers or at high fidelity. Researchers often need to choose between working with a limited subset of the data in their own environment or with richer data in restrictive settings that directly hinder their productivity.

The challenges highlighted here underline the urgent need for new clinical informatics tools, theories and approaches in order to bridge the gap between the clinical care and research strata and accelerate the full translational continuum from basic research, to clinical trials and evaluation and integrated provision of healthcare at a population level [15, 16]. The complex and interdependent relationships that are observed between staff, platforms and data pose significant challenges for accessing data for research (e.g. in terms of cost or obtaining contextual knowledge) and performing research within hospitals (e.g. deploying a clinical decision support tool or undertaking integrated pragmatic clinical trials [17, 18]). Meaningful and sustainable relationships with clinical informatics staff need to be developed and nurtured in order to facilitate the bidirectional flow of knowledge. Furthermore, research should inform the requirements of such complex systems early on, enabling the scalable collection and curation of data in a transparent manner early on. Data mining is the key to insights from clinical big data but the data need to accessible and contain the information needed to improve healthcare.

Acknowledgements
None.

Funding
Not applicable.

Availability of data and material
Not applicable.

Authors' contributions
SD, FA, and JH conceived of and wrote the editorial. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Institute of Health Informatics, University College London, London, UK. [2]Farr Institute of Health Informatics Research, University College London, London, UK. [3]Department of Cardiology, Division Heart and Lungs, University Medical Centre Utrecht, Utrecht, Netherlands. [4]Institute for Biomedical Informatics, Department of Biostatistics and Epidemiology, Perelman School or Medicine, University of Pennsylvania, Philadelphia, PA 19104-6116, USA.

Received: 11 September 2016 Accepted: 14 September 2016
Published online: 22 September 2016

References

1. Collins F, Varmus H. A New initiative on precision medicine. N Engl J Med. 2015;372(9):793–95.
2. Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395–405.
3. McCarty C, Chisholm R, Chute C, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genet. 2011;4(1):13.
4. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008;84(3):362–69.
5. Lyons R, Jones K, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K. The SAIL databank: linking multiple health and social care datasets. BMC Medical Informatics and Decision Making. 2009;9(1):3.
6. Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):e008721.
7. Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: opportunities and challenges. European Heart Journal-Quality of Care and Clinical Outcomes. 2015;1:qcv005.

8. Morley K, Wallace J, Denaxas S, Hunter RJ, Patel RS, Perel P, Shah AD, Timmis AD, Schilling RJ, Hemingway H. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One. 2014;9:11.
9. Denny J, Chapter 13. Mining electronic health records in the genomics era. PLoS Comput Biol. 2012;8(12): e1002823.
10. Iqbal E, Mallah R, Jackson RG, Ball M, Ibrahim ZM, Broadent M, Dzahini O, Stewart R, Johnston C, Dobson RJ. Identification of adverse drug events from free text electronic patient records and information in a large mental health case register. PLoS One. 2015;10:8.
11. Wang Z, Shah A, Tate R, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One. 2012;7(1):e30412.
12. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more Out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol. 2013;9(2):e1002854.
13. Boyd D, Crawford K. Critical questions for big data. Information, Communication & Society. 2012;15(5):662–79.
14. Richards N, King J. Big data ethics. Wake Forest Law Review. 2014;49:393–432.
15. Ainsworth J, Buchan I. Combining health data uses to ignite health system learning. Methods Inf Med. 2015;54(6):479–87.
16. Denaxas S, Friedman CP, Geissbuhler A, Hemingway H, Kalra D, Kimura M, Kuhn KA, Payne HA, de Quiros FG, Wyatt JC. Discussion of "combining health data uses to ignite health system learning". Methods Inf Med. 2015;54(6):488–99.
17. Fröbert O, Lagerqvist B, Olivecrona G, Omerovic E, Gudnason T, Maeng M, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med. 2013;369(17):1587–97.
18. van Staa T-P, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. Health Technol Assess. 2014;18(43):1–146.