Merging heterogeneous clinical data to enable knowledge discovery

Martin G. Seneviratne
Department of Biomedical Data Science, Stanford University
1265 Welch Rd, Stanford
CA 94305, United States
Email: martsen@stanford.edu

Michael G. Kahn
Colorado Clinical and Translational Sciences Institute
Denver, CO 80045, United States
Email: michael.Kahn@ucdenver.edu

Tina Hernandez-Boussard*
Department of Medicine, Biomedical Informatics, Stanford University
1265 Welch Rd, Stanford
CA 94305, United States
Email: boussard@stanford.edu

The vision of precision medicine relies on the integration of large-scale clinical, molecular and environmental datasets. Data integration may be thought of along two axes: data fusion across institutions, and data fusion across modalities. Cross-institutional data sharing that maintains semantic integrity hinges on the adoption of data standards and a push toward ontology-driven integration. The goal should be the creation of query-able data repositories spanning primary and tertiary care providers, disease registries, research organizations etc. to produce rich longitudinal datasets. Cross-modality sharing involves the integration of multiple data streams, from structured EHR data (diagnosis codes, laboratory tests) to genomics, imaging, monitors and patient-generated data including wearable devices. This integration presents unique technical, semantic, and ethical challenges; however recent work suggests that multi-modal clinical data can significantly improve the performance of phenotyping and prediction algorithms, powering knowledge discovery at the patient- and population-level.

Keywords: Data fusion, interoperability, multi-modal data, big data, phenotyping

The quantity of digitized health information has increased exponentially over the past decade, with growing data repositories across all sectors of the health system [1]. The rise of electronic health records has enabled the creation of large datasets containing structured, semi-structured and unstructured data, ranging from diagnostic codes and laboratory results to continuous monitoring signals, clinical notes, medical imaging and pathology. However, there are also rich clinical, molecular and environmental datasets held by government agencies, disease registries, employers, pharmaceutical companies and research organizations. Meanwhile, the proliferation of health tracking apps, wearables and home sensors have created new clinical data streams controlled by the patient, which capture granular information about lifestyle and micro-environmental exposures. Even an individual's social media footprint may be considered as a source of clinical insights. Weber *et al.* have described the spectrum of clinical data available for an individual as a "tapestry of high-value information sources" ranging from the micro (genomic/molecular data) through to the macro (behavioral/lifestyle data) [2].

Many have predicted that the convergence of rich clinical, molecular and environmental data streams will accelerate knowledge discovery in biomedicine and help us to move toward the high-level goal of precision medicine [3,4]. Certainly, larger datasets combining information from numerous sources will improve the performance of diagnostic and prognostic machine learning algorithm, fuelling observational research and improving clinical decisions at the point of care. The critical challenge is how to integrate disparate clinical data streams in a flexible, query-able format while preserving patient privacy and data governance. This integration challenge may be thought of along two axes: data fusion across institutions, and data fusion across modalities.

The first challenge involves cross-institutional data sharing. Federal incentive programs launched through the Health Information Technology for Economic and Clinical Health (HITECH) Act supported the creation of health information exchanges (HIEs) as a platform for clinical data sharing; however based on a 2015 survey, only 23% of HIEs currently supported research, with a further 47% planning to support secondary use in the future [5]. Furthermore, a 2016 review found that the number of HIEs had declined between 2012 and 2014 and only half report being financially sustainable [6]. In 2015, the Office of the National Coordinator of Health IT (ONC) published an Interoperability Roadmap, which outlines a national agenda for improving health information exchange [7]. One key objective is achieving syntactic and semantic interoperability by adoption of common vocabularies, including SNOMED-CT and RxNorm, and common data formats, including consolidated clinical document architecture (C-CDA) and Fast Health Interoperability Resources (FHIR). The roadmap also calls for the adoption of secure transport standards and outlines best practices for matching patient identities between sites. In parallel, there have been a number of academic endeavors to build platforms for observational clinical research, including the Observational Health Data Sciences and Informatics (OHDSI) network [8], SHARPn project [9], and the Informatics for Integrating Biology and the Bedside (i2b2) initiative [10].

An emerging theme throughout these cross-institutional data fusion efforts, from industry to academia, is the power of ontology-driven data integration, inspired by the rise of semantic web technologies [11–13]. This approach has a number of distinct advantages including the ability to synthesize across many disparate data sources via high-level ontologies and the ability to reason over a knowledge base [14]. Ongoing technical challenges include representing data provenance, temporal relationships and data quality [15]; however the prevailing challenge is operational - how to shift organizational culture toward interoperability and data sharing [16]. Beyond this, the infrastructure for interoperability may vary, with successful examples of centralized data warehouses [17], decentralized blockchain-based health records systems [18], and patient-controlled health records [19].

The second major component of data fusion is cross-modality integration. Most EHRs contain a diversity of data types that have traditionally been analyzed independently, ranging from structured diagnosis codes to signal data, clinical notes and imaging. Furthermore, the interoperability advances mentioned above are making it possible to harmonize traditional EHR data with novel clinical data streams including genomic, microbiome, metabolic and patient-generated health data (PGHD). There is an expanding evidence base showing that multi-modal data integration can support precision medicine by stratifying patients based on their ‘deep phenotype’ [20]; improving the performance of clinical decision support algorithms for diagnosis and prediction [21]; and uncovering new phenotypes altogether [22]. For example, Zhao *et al.* developed a risk prediction model for cardiovascular events using EHR data, but found a significant performance boost when those data were fused with

patient-level genomic information [23]. Meanwhile, by using unsupervised learning on a combined dataset of metabolome, microbiome, genetics and imaging data, Shomorony *et al.* were able to identify a signature of biomarkers that identified diabetic patients more accurately than traditional clinical metrics (glucose, insulin resistance, and body-mass-index) - suggesting novel pathways that may be involved in the development of diabetes [24].

The combination of traditional health data with PGHD or social media data has enabled knowledge discovery in the realms of both precision medicine and population health. Santillana *et al.* combined hospital visit data with Twitter, Google searches, and posts on an online health forum to predict influenza incidence [25]. Vilar *et al.* describe efforts to identify drug-drug interactions by combining social media posts with the biomedical literature [26]. On a more granular level, there is a push to integrate patient-reported outcomes (PROs) into EHRs as a way to promote patient-centric care (an example of heterogeneous data fusion potentially driving behavior change) [27] which has fueled interesting insights into the relationship between PROs and clinical outcomes such as mortality [28]. The rise of the 'Internet of Things' in healthcare - the ecosystem of connected monitoring devices that surround a patient - as well as ambient information such as geo-location are creating opportunities for even richer multi-modal datasets [29–31]. These data no longer reside exclusively in hospitals. Private sector initiatives such as Verily's Project Baseline and Apple's HealthKit program are enabling patients to aggregate multiple medical data sources [32,33]. Meanwhile, the *All Of Us* initiative is a National Institutes of Health program to collect molecular, clinical and environmental data on a diverse cohort of volunteers for research purposes [34]. As the pathophysiology behind chronic disease is a complex interplay of clinical, molecular and behavioral factors acting over extended time periods, the datasets required to tackle the global epidemic of chronic disease will need to be similarly layered and sophisticated. There is both a clinical opportunity and an economic one, with increasing evidence to suggest that data integration can reduce overall healthcare costs [35].

Cross-modality data integration is associated with a number of challenges, of which we highlight three below. First, there is the issue of how to harmonize data from distant parts of a knowledge graph reflecting radically different levels of abstraction e.g. diagnosis codes (high-level) with proteomic data (low-level). This creates challenges for data storage and makes it difficult to generate feature vectors to train classifiers. Several recent studies have shown that deep learning can be used to create efficient abstract representations of structured and unstructured EHR data, for example the *DeepPatient* representation using stacked denoising autoencoders [36]. A similar approach might be considered for a broader range of input data. A second caveat is around data stewardship, particularly with respect to privacy and security [37]. Fusion of data streams may accelerate scientific discovery and clinical care, but this comes with an increased risk of patient re-identification. Further work is needed around de-identification, consent processes and access control when data are contributed to shared repositories. The increasing volume of digital health information available to clinicians also raises questions around liability and duty of care i.e. the extent to which clinicians are responsible for the full expanse of information in an aggregated health repository. A third challenge is around equity and inclusion. A 2018 report by Ferryman *et al.* on 'Fairness in precision medicine' highlights the potential for bias in large-scale biomedical training data, stemming from historical discrimination in the health system and recruitment biases at academic medical centers [38]. Data-fusion efforts must be cognizant of the distribution of important demographic variables, such as gender, ethnicity and socioeconomic status in their input data.

The fusion of heterogeneous datasets from different institutions and across different modalities presents a powerful opportunity to drive knowledge discovery in biomedicine. There are technical and operational challenges to enable data sharing across borders of institutional ownership, which we are beginning to overcome with interoperability standards and data sharing platforms. Arguably the more nuanced problem today is how to grapple with extremely diverse data types that encompass the micro and macro scales of a patient's data signature, including how to create flexible data storage and machine learning architectures, and how to design stewardship processes to govern these data appropriately. Holzinger *et al.* claimed in 2014 that "biomedical research is drowning in data, yet starving for knowledge". Today we have more health data than ever before, but the challenge remains how to harmonize, structure and learn from multi-modal datasets [39].

Acknowledgements

This work is partially supported by the National Cancer Institute of the National Institutes of Health (NIH) under Award Number R01CA183962 and by NIH/NCATS Colorado CTSA Grant Number UL1 TR002535. Contents are the authors' sole responsibility and do not necessarily represent official NIH views.

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309: 1351–1352.
2. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA. 2014;311: 2479–2480.
3. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff . 2014;33: 1115–1122.
4. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375: 1216–1219.
5. Parker C, Reeves M, Weiner M, Adler-Milstein J. Health Information Exchange Organizations and Their Support for Research: Current State and Future Outlook. Inquiry. SAGE Publications; 2017;54. doi:10.1177/0046958017713709
6. Adler-Milstein J, Lin SC, Jha AK. The Number Of Health Information Exchange Efforts Is Declining, Leaving The Viability Of Broad Clinical Data Exchange Uncertain. Health Aff . 2016;35: 1278–1285.
7. Connecting Health and Care for the Nation A Shared Nationwide Interoperability Roadmap. Office of the National Coordinator for Health Information Technologies; 2015. Report No.: 1.0.
8. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216: 574–578.
9. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform. 2012;45: 763–771.
10. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17: 124–130.
11. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H-U, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PLoS One. 2015;10: e0116656.
12. Hsu W, Gonzalez NR, Chien A, Pablo Villablanca J, Pajukanta P, Viñuela F, et al. An integrated, ontology-driven approach to constructing observational databases for research. J Biomed Inform. 2015;55: 132–142.
13. Zhang H, Guo Y, Li Q, George TJ, Shenkman E, Modave F, et al. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med Inform Decis Mak. 2018;18: 41.
14. Lezcano L, Sicilia M-A, Rodríguez-Solano C. Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. J Biomed Inform. 2011;44: 343–353.
15. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A Data Quality Ontology for the Secondary Use

of EHR Data. AMIA Annu Symp Proc. 2015;2015: 1937–1946.
16. Ong T, Pradhananga R, Holve E, Kahn MG. A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation. EGEMS (Wash DC). 2017;5: 10.
17. Seneviratne M, Seto T, Blayney DW, Brooks JD, Hernandez-Boussard T. Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2018;6.
18. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: Using Blockchain for Medical Data Access and Permission Management. 2016 2nd International Conference on Open and Big Data (OBD). 2016. pp. 25–30.
19. Chan D, Howard M, Dolovich L, Bartlett G, Price D. Revolutionizing patient control of health information. Can Fam Physician. 2013;59: 823–824.
20. Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. Cold Spring Harb Mol Case Stud. 2015;1: a000372.
21. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1: 18.
22. Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. The Author(s); 2016;64: 168–178.
23. Zhao J, Feng Q, Wu P, Lupu R, Wilke RA, Wells QS, et al. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction [Internet]. bioRxiv. 2018. p. 366682. doi:10.1101/366682
24. Shomorony I, Cirulli ET, Huang L, Napier LA, Heister RR, Hicks M, et al. Unsupervised integration of multimodal dataset identifies novel signatures of health and disease [Internet]. bioRxiv. 2018. p. 432641. doi:10.1101/432641
25. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol. 2015;11: e1004513.
26. Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform. 2018;19: 863–877.
27. Gensheimer SG, Wu AW, Snyder CF, Basch E, Gerson J, Holve E, et al. Oh, the Places We'll Go: Patient-Reported Outcomes and Electronic Health Records. The Patient - Patient-Centered Outcomes Research. 2018; doi:10.1007/s40271-018-0321-9
28. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. JAMA. 2017;318: 197–198.
29. J. Andreu-Perez, D. R. Leff, H. M. D. Ip, G. Yang. From Wearable Sensors to Smart Implants--Toward Pervasive and Personalized Healthcare. IEEE Transactions on Biomedical Engineering. 2015;62.
30. Schinasi LH, Auchincloss AH, Forrest CB, Diez Roux AV. Using electronic health record data for environmental and place based population health research: a systematic review. Ann Epidemiol. 2018;28: 493–502.
31. Saelens BE, Arteaga SS, Berrigan D, Ballard RM, Gorin AA, Powell-Wiley TM, et al. Accumulating Data to Optimally Predict Obesity Treatment (ADOPT) Core Measures: Environmental Domain: ADOPT: Environmental Domain. Obesity . 2018;26: S35–S44.
32. Barr A. Google to Collect Data to Define Healthy Human. Wall Street Journal. wsj.com; 27 Jul 2014. Available: https://www.wsj.com/articles/google-to-collect-data-to-define-healthy-human-1406246214. Accessed 16 Oct 2018.
33. North F, Chaudhry R. Apple HealthKit and Health App: Patient Uptake and Barriers in Primary Care. Telemed J E Health. 2016;22: 608–613.
34. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372: 793–795.
35. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff . 2014;33: 1123–1131.
36. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016;6: 26094.
37. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform. 2014;9: 97–104.
38. Ferryman K, Pitcan M. Fairness in Precision Medicine. Data & Society; 2018 Feb.
39. Holzinger A, Jurisica I. Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. In: Holzinger A, Jurisica I, editors. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. pp. 1–18.