An informatics research agenda to support precision medicine: seven key areas

RECEIVED 31 August 2015
REVISED 25 November 2015
ACCEPTED 24 December 2015
PUBLISHED ONLINE FIRST 23 April 2016

Jessica D Tenenbaum,[1] Paul Avillach,[2] Marge Benham-Hutchins,[3] Matthew K Breitenstein,[4] Erin L Crowgey,[5] Mark A Hoffman,[6] Xia Jiang,[7] Subha Madhavan,[8] John E Mattison,[9] Radhakrishnan Nagarajan,[10] Bisakha Ray,[11] Dmitriy Shin,[12] Shyam Visweswaran,[13] Zhongming Zhao,[14] and Robert R Freimuth[4]

ABSTRACT

The recent announcement of the Precision Medicine Initiative by President Obama has brought precision medicine (PM) to the forefront for healthcare providers, researchers, regulators, innovators, and funders alike. As technologies continue to evolve and datasets grow in magnitude, a strong computational infrastructure will be essential to realize PM's vision of improved healthcare derived from personal data. In addition, informatics research and innovation affords a tremendous opportunity to drive the science underlying PM. The informatics community must lead the development of technologies and methodologies that will increase the discovery and application of biomedical knowledge through close collaboration between researchers, clinicians, and patients. This perspective highlights seven key areas that are in need of further informatics research and innovation to support the realization of PM.

Keywords: precision medicine, informatics, biomarkers, data sharing

The recent announcement of the Precision Medicine (PM) Initiative by President Obama[1] has brought PM to the forefront for healthcare providers, researchers, regulators, and funders alike. In order for PM to be fully realized, we must move toward a Learning Healthcare System model that extends evidence-based practice to practice-based evidence by using data generated through clinical care to inform research (Figure 1).[2] The leadership and members of the American Medical Informatics Association Genomics and Translational Bioinformatics Working Group have identified seven key areas that informatics research should explore to enable PM's vision.

PATIENTS: PAST, PRESENT, AND FUTURE

Stakeholders in the biomedical enterprise include researchers, providers, payers, and patients. But nearly everyone has been or will be a patient at some point. Patients thus are, and must remain, at the heart of the biomedical enterprise.

Key Area One: Facilitate Electronic Consent and Specimen Tracking

In the era of PM, research studies produce more data than they can possibly use and, paradoxically, would benefit from more data than they can possibly generate. As genomic sequencing becomes increasingly available, using de-identified biospecimens for research becomes more nuanced.[3] Research participants may be asked to give broad consent to the future use of their data and biospecimens, and to acknowledge the possible, though unlikely, prospect of sequence-based re-identification.[4,5] To maximize data and biospecimen reuse while protecting study participants' privacy and adhering to their wishes, it is essential to develop machine-readable consent forms that enable electronic queries.[6] As large biorepositories linked to electronic health records (EHRs) become more common, informatics will enable researchers to identify cohorts – both intra- and interinstitutionally – that meet their study criteria and have given the requisite consent. Proper local management of specimens and derived samples enables accurate tracking of chain of custody, sample derivations, processing/handling, and quality control – all of which are key elements of rigorous and reproducible research.[7] Structured and electronically available consent forms can empower study participants by allowing them to access, review, and modify their preferences. A number of large-scale initiatives, including Sage Bionetworks, the Genetic Alliance, and the Global Alliance for Genomic Health, are making progress in this area.

Areas of informatics that can facilitate study participant consent and sample tracking include the development of structured consent forms and the adoption of relevant ontologies,[6,8] user interface design, and infrastructure to enable participant engagement after the point of enrollment. Developing an infrastructure to perform role-based distributed queries over cohorts and sample collections, such as those provided by OpenSpecimen, the Shared Health Research Information Network (SHRINE), and PopMedNet, will also be important.[9–11]

DATA TO KNOWLEDGE

The promise of PM can only be realized by aggregating (virtually or otherwise) and analyzing data from multiple sources. A recent report by the National Academy of Sciences calls for the development of an information commons (IC) that amasses medical, molecular, social, environmental, and health outcomes data for large numbers of individual patients.[12] The IC would be continuously updated, enable data analyses, and serve as the foundation for a knowledge base (KB) (see Key Area Five). Creating an IC would require informatics expertise to develop data standards, ensure data security, standardize processing pipelines, and establish data provenance.

Correspondence to Jessica D Tenenbaum, Box 2721, Durham, NC 27710, USA; jessie.tenenbaum@duke.edu; Tel: +1 (919) 684-7308; For numbered affiliations see end of article.

Figure 1: Informatics methodology enables precision medicine (PM) throughout the Learning Healthcare System cycle. Patients – past, present, and future – are at the beginning and end of the cycle. Both healthcare and research participation result in the generation of data. Informatics methods and tools help turn data into information, and information into knowledge. That knowledge, in turn, influences individuals' behavior and informs patient care. Informatics plays a key role in enabling each stage and transition of this cycle.

Key Area Two: Develop, Deploy, and Adopt Data Standards to Ensure Data Privacy, Security, and Integrity, and to Facilitate Data Integration and Exchange

Transparency, reciprocity, respecting study participant preferences, data quality/integrity, and security are key to obtaining and maintaining the massive data stores needed for the advancement of PM.[13] Data security does not mean data lock-down. Data-sharing can allow a study to proceed despite low numbers of eligible participants at any single institution, and can enable data reuse or meta-analyses. Data and metadata standards are required for data integration and exchange to be successful, but the lack of such standards or inconsistent use of existing standards are frequent barriers to this goal, especially in emergent "omics" disciplines.[14] Data gaps are often discovered when existing standards are adopted for other purposes. Rather than creating yet another standard, those seeking to adopt an existing standard should work with its owners to help extend its scope. Conversely, funders and standards owners should place more emphasis on outreach and education/training for potential adopters of existing data standards. A number of initiatives are working to tackle different aspects of this challenge, including BioSharing, the Center for Expanded Data Annotation and Retrieval (CEDAR), the Biomedical and Healthcare Data Discovery Index Ecosystem (bioCADDIE), and Integrating Data for Analysis, Anonymization, and Sharing (iDASH).[15–18]

Although there have been significant efforts to share molecular datasets publicly, less progress has been made on sharing healthcare data. An emerging strategy is the development of clinical research networks in which EHR-derived data is stored locally, mapped to a common data model, and queried by proxy for members of a consortium or collaboration. Sharing queries rather than data resolves many of the issues that are involved in data standardization and harmonization, data governance, as well as the legal and privacy concerns surrounding other federated or aggregation models. This strategy has been adopted by initiatives such as MiniSentinel, Observational Health Data Sciences and Informatics (OHDSI), and the National Patient-Centered Clinical Research Network (PCORNet).[19–21] Building on these networks to include genomic and other "omics" data, environmental data, and social data is one way forward in the development of ICs for PM.

Work on data and metadata standards should be recognized and incentivized by the organizations that use and benefit from them, including academia, industry, government regulators, and funding agencies. New methods of encrypting and sharing genomic data in a way that enables collaborative research without compromising patient privacy are needed.

Key Area Three: Advance Methods for Biomarker Discovery and Translation

A primary goal of PM is to uncover subphenotypes defined by the distinct molecular mechanisms that underlie variations in disease manifestations and outcomes.[12] One step toward defining subphenotypes is to establish agreed-upon phenotype definitions for existing disease classifications, a surprisingly complex task.[22] A number of different initiatives (eg, the Electronic Medical Records and Genomics [eMERGE] Network and the National Institutes of Health [NIH] Collaboratory) are working to make phenotype definitions computationally tractable and reproducible between sites.[23,24] Although some progress in subphenotyping has been made, new methods, including analyses of high-dimensional data,[25] integration of different types of data (eg, "omics," imaging, clinical, environmental),[26,27] and simulating disease behaviors across multiple biological scales in space and time,[28] are needed to address a number of challenges.

Although molecular biomarkers can help elucidate underlying physiological mechanisms of disease, only a minority of currently known biomarkers are clinically actionable. Moreover, critical disease subtype distinctions may be impacted by nonmolecular factors, such as socioeconomic status.[29] Many questions must be answered before a potentially actionable biomarker can become part of a clinical guideline and translated into practice.[30] Information that is necessary for bridging this gap might include the functional characterization of genes and pathways related to the biomarker, the level of evidence, and data about economic feasibility. Clinical decision making regarding actionable biomarkers would be facilitated by a framework for presenting different levels of evidence regarding whether and how a molecular abnormality, genomic or otherwise, might represent a therapeutically relevant biomarker.[31,32] Variant annotations with actionable clinical information will enable decision support systems to provide interpretable and actionable patient-specific reports.[33–35]

Immediate areas for informatics research to focus on include computational phenotyping, biomarker discovery based on heterogeneous data sources, and frameworks for evaluating clinical actionability and utility.

Key Area Four: Implement and Enforce Protocols and Provenance

Scaling up PM requires complex processing and analytic steps applied to large, heterogeneous datasets. With so many "moving parts," there are many opportunities for errors in the analysis, interpretation, or exchange of information. It is important that both final results and intermediate steps be well documented and fully reproducible. Protocols, and deviations from them, must also be documented. Software versions, analytical parameters, and reference database builds must all be captured as readily available metadata. Although spreadsheets and documents can be useful for informal data exploration, they do not constitute an adequate data management system.

Large projects often share data between groups and may last several years, during which time key personnel may change institutions. All data processing and analysis for final results should be automated and documented so that another researcher can reproduce the work without making assumptions about what was done. There are various tools that enable this approach, including Taverna, preconfigured virtual machines, and Sage Bionetworks's Synapse Platform.[36–38] Though new challenges will always require novel and innovative solutions, the adoption of standard operating procedures when appropriate will facilitate consistency and improve interoperability. In addition, policies must be enacted and enforced to ensure responsible, reproducible, and reusable science.

Processes and protocols for capturing and exchanging metadata and data provenance must be established, standardized, and widely adopted. Furthermore, this information must be considered to be as important as the primary data it describes, and funding agencies and publishers should insist that it be included with any dataset that is produced and released publicly.

KNOWLEDGE TO ACTION

Clinical decision making requires the consolidation of PM knowledge and the development of clinical decision support tools (CDS), which, together with individual patient data, will provide actionable information at the point of care.

Key Area Five: Build a Precision Medicine Knowledge Base

A comprehensive KB will contain information about disease subtypes, disease risk, diagnosis, therapy, and prognosis that emerges from the ongoing analysis of data in an IC. Such a KB must be flexible, scalable, and extensible. Current KBs (eg, on genomic variants) are isolated from one another and do not support federated querying. Informatics solutions are needed for data-sharing and building a consensus on clinical interpretations of disparate, multiscale data. This KB must be machine-readable, as well as human-readable. Knowledge management technologies must enable effective ontological modeling, knowledge provenance, and new methodologies for updating and maintaining the integrated KB. Novel computational reasoning approaches must be utilized to allow efficient federated queries to be run across billions of knowledge units, enabling causal inference and decision support.

New methods and processes must be developed to organize biomedical knowledge into integrated and interconnected KBs that will enable precision diagnostics and therapeutics based on the latest genomic discoveries and clinical evidence. Such KBs must provide federated queries and flexible computational analytics capabilities tailored for use by physicians and researchers.

Key Area Six: Enhance EHRs to Promote Precision Medicine

Commercial EHRs enable CDS for PM that is focused on information about a single gene variant.[39] Informatics challenges for CDS include integrating next generation CDS with PM KBs to provide genome-based risk predictions, prognoses, and drug dosing at the point of care, as well as representing discrete genomic findings and interpretations in a machine-readable format (vs a free-text pathologist or geneticist report). Masys et al.[40] proposed a framework for integrating genome-level data (stored external to the EHR) in which decision support systems are implemented through the EHR. EHRs will need to better aggregate and display patient information in order to allow users to view the heterogeneous data available for each patient, and EHRs will also need to structure and visually display the aggregated knowledge about each patient. Open interfaces that facilitate modular development of genomic CDSs outside of monolithic EHR vendor systems, enabling unencumbered parallel innovation/evolution of each element, should be provided.

EHR systems must provide standards-based programming interfaces that enable the integration of external data and knowledge sources as well as the development of tools that support custom workflows, novel analytics, data visualization, and data aggregation. The informatics community must partner with EHR vendors to author use cases and develop interfaces, such that both parties benefit from the collaboration.

Key Area Seven: Facilitate Consumer Engagement

PM includes more than the medical care administered in a provider's office. Most of the population spends far more time outside of the doctor's office than in it. PM will require explicit acknowledgement of this fact as well as deeper consumer participation, which will involve making consumers aware of their own ongoing health status and engaging them in healthcare decision making. It will also involve collecting more information about a person's environment and lifestyle choices between visits to the doctor – eg, activity level, nutrition information, exposure, and sleep patterns – and incorporating that information into targeted therapeutic and preventive treatments.

Consumer access to genetic testing will increase as provider-ordered and direct-to-consumer genetic tests become more comprehensive and less expensive. Along with the recent announcement from 23andMe that the company will once again offer health-related information and Ancestry's launch of AncestryHealth[41] comes the increased importance of ensuring that consumers understand basic genetic principles and the implications of genetic testing, of trust in the accuracy of genetic tests, and of understanding of how these results, together with family history, will influence treatment decisions.

User-friendly interfaces for the collection, visualization, and integration of consumer data with healthcare information will be key to realizing the potential value of nontraditional data sources. Standards for new consumer data types, as well as patient engagement around ethical, legal, and social issues, will also be important.

CONCLUSIONS

The emergence of PM as a priority in biomedical research and healthcare emphasizes the importance of informatics' contributions to PM. This brief overview highlights essential research directions for both informatics researchers and funding organizations.

ACKNOWLEDGEMENTS

The authors thank our colleagues in the Genomics and Translational Bioinformatics Working Group. Their contributions to discussions online, during formal Working Group meetings, and in casual encounters, both at our home institutions and at annual conferences, have helped shape our thoughts and perspectives as reflected in this manuscript. We also thank Joseph Romano, Peggy Peissig, Carolyn Petersen, Li Lang, and Alexis B. Carter, who participated in early discussions of these ideas. Finally, we thank the reviewers, whose insightful questions and thoughtful suggestions helped to significantly improve the manuscript.

CONTRIBUTORS

All authors contributed to overall intellectual content and specific sections of writing. JDT and RRF edited the manuscript for coherence.

FUNDING

This work was funded in part by NIGMS U19 GM61388 (the Pharmacogenomics Research Network) (RRF), NCATS UL1-TR001117 (JDT), U19-GM61388-13 and R25-CA092049 (MKB), and NLM R01-LM012095 (SV), NLM R01-LM011177 (ZZ), R00-LM010822 and R01-LM011663 (XJ), Delaware INBRE #P20 GM103446 (EC), NCATS UL1-TR000117 (RN), NCI P30-CA51008, NCATS UL1-TR001409 (SM). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

COMPETING INTERESTS

P.A. is a paid consultant for Claritas Genomics.

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. *New Engl J Med.* 2015;372(9):793–795.
2. Califf RM, Platt R. Embedding cardiovascular research into practice. *JAMA.* 2013 Nov 20;310(9):2037–8.
3. Kulynych J, Greely H. Every patient a subject: when personalized medicine, genomic research, and privacy collide. *Slate.com.* 30 December 2014.
4. Hudson KL, Collins FS. Bringing the Common Rule into the 21st Century. *N Engl J Med.* 2015 Dec 10;373(24):2293–2296.
5. Gymrek M, McGuire AL, Golan D, *et al*. Identifying personal genomes by surname inference. *Science.* 2013;339(6117):321–324.
6. Grando Maria Adela, Boxwala Aziz, Schwab Richard, *et al*. Ontological approach for the management of informed consent permissions. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*. IEEE; 2012.
7. Brochhausen M, Fransson MN, Kanaskar NV, *et al*. Developing a semantically rich ontology for the biobank-administration domain. *J Biomed Semantics.* 2013;4(1):23.
8. Zheng J, Stoeckert C, Brochhausen M. Ontology for Biobanking. 2014. https://github.com/biobanking/biobanking. Accessed 25 October 2015.
9. McIntosh LD, Sharma MK, Mulvihill D, *et al*. caTissue suite to OpenSpecimen: developing an extensible, open source, web-based biobanking management system. *J Biomed Inform.* 2015.
10. Weber GM, Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc.* 2009;16(5):624–630.
11. McGlynn EA, Lieu TA, Durham ML, *et al*. Developing a data infrastructure for a learning health system: the PORTAL network. *J Am Med Inform Assoc.* 2014;21(4):596–601.
12. National Research Council Committee on a Framework for Developing a New Taxonomy of Disease. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. US: National Academies Press; 2011.
13. Whitehouse. Precision medicine initiative: proposed privacy and trust principles. 2015.
14. Tenenbaum JD, Sansone SA, Haendel M. A sea of standards for omics data: sink or swim? *J Am Med Inform Assoc.* 2014;21(2):200–203.
15. Field D, Sansone S, Delong EF, *et al*. Meeting Report: BioSharing at ISMB 2010. *Stand Genomic Sci.* 2010;3(3):254–258.
16. Ohno-Machado L, Alter G, Fore I, *et al*. Biomedical and healthCAre Data Discovery Index Ecosystem. https://biocaddie.org. Accessed 25, October 2015.
17. Jiang X, Zhao Y, Wang X, *et al*. A community assessment of privacy preserving techniques for human genomes. *BMC Med Inform Decis Mak.* 2014;14(Suppl 1):S1.
18. Musen MA, Bean CA, Cheung KH, *et al*. The center for expanded data annotation and retrieval. *J Am Med Inform Assoc.* 2015;22(6):1148–1152.
19. Curtis LH, Weiner MG, Boudreau DM, *et al*. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf.* 2012;21(Suppl 1):23–31.
20. Collins FS, Hudson K L, Briggs JP, *et al*. PCORnet: turning a dream into reality. *J Am Med Inform Assoc.* 2014;21(4):576–577.
21. Hripcsak G, Duke JD, Shah NH, *et al*. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574–578.
22. Richesson RL, Rusincovitch SA, Wixted D, *et al*. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc.* 2013;20(e2):e319–e326.
23. Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc.* 2013;20(e1):e147–e154.
24. Richesson RL, Hammond WE, Nahm M, *et al*. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc.* 2013;20(e2):e226–e231.
25. Fernald GH, Capriotti E, Daneshjou R, *et al*. Bioinformatics challenges for personalized medicine. *Bioinformatics.* 2011;27(13):1741–1748.
26. Ritchie MD, Holzinger ER, Li R, *et al*. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet.* 2015;16(2):85–97.
27. Li L, Wei-Yi C, Benjamin SG, *et al*. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med.* 2015;7(311):311ra174–311ra174.
28. Deisboeck TS, Wang Z, Macklin P, *et al*. Multiscale cancer modeling. *Annu Rev Biomed Eng.* 2011;13:127–155.
29. Bradley CJ, Given CW, Roberts C. Race, socioeconomic status, and breast cancer treatment and survival. *J Natl Cancer Inst.* 2002;94(7):490–496.
30. Lander ES. Cutting the Gordian helix – regulating genomic testing in the era of precision medicine. *N Engl J Med.* 2015;372(13):1185–1186.
31. Chen K, Meric-Bernstam F, Zhao H, *et al*. Clinical actionability enhanced through deep targeted sequencing of solid tumors. *Clin Chem.* 2015;61(3):544–553.
32. Vidwans SJ, Turski ML, Janku F, *et al*. A framework for genomic biomarker actionability and its use in clinical decision making. *Oncoscience.* 2014;1(10):614–623.
33. Dorschner MO, Amendola LM, Shirts BH, *et al*. Refining the structure and content of clinical genomic reports. *Am J Med Genet C Semin Med Genet.* 2014;166C(1):85–92.
34. Rehm HL, Berg JS, Brooks LD, *et al*. ClinGen – the Clinical Genome Resource. *N Engl J Med.* 2015;372(23):2235–2242.
35. Richards S, Aziz N, Bale S, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17(5):405–424.
36. Wolstencroft K, Haines R, Fellows D, *et al*. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* 2013;41(Web Server issue):W557–W561.
37. Omberg L, Ellrott K, Yuan Y, *et al*. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet.* 2013;45(10):1121–1126.
38. Dudley JT, Butte AJ. In silico research in the era of cloud computing. *Nat Biotechnol.* 2010;28(11):1181–1185.
39. Crews KR, Hicks JK, Pui CH, *et al*. Pharmacogenomics and individualized medicine: translating science into practice. *Clin Pharmacol Ther.* 2012;92(4):467–475.
40. Masys DR, Jarvik GP, Abernethy NF, *et al*. Technical desiderata for the integration of genomic data into Electronic Health Records. *J Biomed Inform.* 2012;45(3):419–422.
41. Ancestry. *AncestryHealth.* 2015. https://health.ancestry.com. Accessed 24, November 2015.

AUTHOR AFFILIATIONS

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

[2]Department of Biomedical Informatics, Harvard Medical School & Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, USA

[3]School of Nursing, University of Texas, Austin, TX, USA

[4]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

[5]Center for Bioinformatics & Computational Biology, University of Delaware, Newark, DE, USA

[6]Department of Biomedical & Health Informatics, University of Missouri – Kansas City, Children's Mercy Hospital, Kansas City, MO, USA

[7]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[8]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Innovation Center for Biomedical Informatics, Washington, DC, USA

[9]Exponential Medicine, Singularity University; Internal Medicine, System Solutions at Kaiser Permanente, Pasadena, CA, USA

[10]Division of Biomedical Informatics, University of Kentucky, Lexington, KY, USA

[11]Center for Health Informatics and Bioinformatics, New York University School of Medicine, New York, NY, USA

[12]Department of Pathology, MU Informatics Institute, University of Missouri, Columbia, MO, USA

[13]Department of Biomedical Informatics and the Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

[14]Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA