Editor: **Daniel B. Neill,** H.J. Heinz III College, Carnegie Mellon University, neill@cs.cmu.edu

Subtyping: What It Is and Its Role in Precision Medicine

Suchi Saria, *Johns Hopkins University*
Anna Goldenberg, *SickKids Research Institute*

Many diseases—for example, neuropsychiatric, cardiovascular, and autoimmune disorders—are difficult to treat because of the remarkable degree of variation among affected individuals. *Precision medicine*, also known as personalized medicine or P4 medicine,[1] is an emerging approach for individualizing the practice of medicine.[2] It takes into account individual variability in genes, environment, and lifestyle with the goals of better defining health or wellness for each person, predicting disease progression and transitions between disease stages, and targeting the most appropriate medical interventions.

In the autoimmune disease scleroderma, for example, as many as six different organ systems may be involved. Organ involvement trajectories can vary greatly across individuals from no involvement to rapid and aggressive decline.[3] This uncertainty associated with an individual's disease progression makes treatment planning challenging. Furthermore, the current evidence base for guiding an individual's treatment is insufficient in several ways. First, clinical practice guidelines overemphasize simplicity so that healthcare providers can easily implement them without computerized decision support. Thus, it is rare to see decision criteria combining many different types of data about the individual (such as molecular, genetic, and clinical) to make a therapeutic recommendation. Second, most of these guidelines are derived from randomized controlled trials for single disease treatments, which can exclude patients with significant complications; the evidence base derived is not tailored to the granular characteristics of each individual, but rather the "average" patient in the recruited cohort. Consequently, the knowledge needed to provide appropriate therapy for patients with complex cases, who also consume the lion's share of healthcare spending, is largely lacking.[4]

These challenges motivated the idea of disease subtyping as a central tenet of precision medicine.[1] Broadly construed, disease subtyping is the task of identifying subpopulations of similar patients that can guide treatment decisions for a given individual. (When the subtypes have been established to be causally associated with the underlying mechanism, these are also called *endotypes*.[5]) Boland and colleagues describe the concept of a "verotype" (the Latin word *vero* means "true") to represent the true population of similar patients for treatment purposes.[6] What constitutes these verotypes and how they should be discovered remains an open question. An active and growing body of work has explored different approaches for identifying homogeneous patient subgroups ranging from qualitative—based on clinical observations alone—to quantitative models that integrate measurements from diverse high-throughput biotechnologies. Cancer, autism, autoimmune diseases, cardiovascular diseases, and Parkinson's are examples of diseases that have been studied through the lens of subtyping.[3,7,8]

The discovery and refinement of disease subtypes can benefit both the practice and science of medicine. Clinically, by refining prognoses based on similar individuals, disease subtypes help reduce uncertainty in an individual's expected outcome. Accurate prognoses can thereby improve treatment decisions. For example, administration of a therapy with strong side effects could be well justified on an individual prognosticated to decline rapidly without this treatment. Beyond prognoses, subtypes can also inform forecasts about the expected costs of care. In complex diseases, where there is

1541-1672/15/$31.00 © 2015 IEEE
Published by the IEEE Computer Society

tremendous heterogeneity in disease presentation, subtyping can help improve the effectiveness of clinical trials by enabling targeted recruitment.

Scientifically, subtypes can drive the design of new genome-wide association studies.[9,10] For example, by finding subgroups whose clinical manifestations differ, researchers can conduct targeted studies to identify the molecular determinants of these differences. Such analyses can allow clinical scientists to understand the causes of related diseases.

In this article, we provide an overview of the diverse approaches to subtyping, from early accounts based on clinical practice to more recent approaches that focus on computationally derived subtypes based on molecular and electronic health record (EHR) data. This field is expansive and growing rapidly—thus, a comprehensive review is not our focus here. Instead, we juxtapose approaches taken by different communities and emphasize the significant open computational problems that remain.

Disease Subtyping: Overview

Traditionally, disease subtyping research has been conducted as a by-product of clinical experience, wherein a clinician noticed the presence of patterns or groups of outlier patients and performed a more thorough (retrospective or prospective) study to confirm their existence. An early case of such an analysis is the work of James Ewing, a pathologist, who published his observation of a clearly distinct subset of osteogenic sarcoma (OS), a type of bone tumor, nearly a century ago. He observed that a substantial number of his patients with OS experienced spontaneous fractures and swellings.[11] All these patients had very characteristic radiographic features that were substantially different from his typical patients with OS. Subsequent microscopic analysis of the tumor tissue revealed that these tumors were indeed of a different (endothelial) origin, were rather common among younger subjects, and were clearly distinct from the mainstream OS. This subtype later came to be known as *Ewing's sarcoma*.

Early examples of subtyping were limited by the power of individual doctors to detect patterns among the patients they had observed. In the last decade, the advent of high-throughput biotechnologies has provided the means for measuring differences between individuals at the cellular and molecular levels. The cost of measuring various "–omic" data (such as genomic, proteomic, and metabolomics data) has dropped significantly, letting scientists collect such data on a large number of patients. The number of measured variables in these data ranges from tens of thousands (for example, expression levels of messenger RNA, or mRNA) to millions (genetic data in the form of single nucleotide variants); thus, research has shifted toward computationally driven approaches to identify subtypes.

Molecular Subtyping

One of the main goals driving the analyses of high-throughput molecular data is the unbiased biomedical discovery of disease subtypes via unsupervised clustering of either individual or multiple sources of molecular data. Using statistical and machine learning approaches such as nonnegative matrix factorization, hierarchical clustering, and probabilistic latent factor analysis,[12,13] researchers have identified subgroups of individuals based on similar gene expression levels. More recent approaches have targeted data integration. Toward this, researchers have tried a broad range of techniques,[14] spanning from ad hoc combinations of individual datastream analyses to latent variable models[15,16] to more recent network-based fusion approaches.[17]

One of the biggest drawbacks of this line of work is that depending on the type of data used, the resulting conclusions about disease subtypes differed. Glioblastoma multiforme (GBM), a very aggressive form of brain cancer, is a good example of different analyses producing a range of conclusions. An earlier analysis of GBM identified two subtypes based on the loss of a chromosome.[18] An integrative analysis of GBM driven primarily by mRNA expression data identified four different subtypes, which were not strict subsets of those previously identified.[7] A recent DNA methylation-based approach[19] identified a subtype, characterized by a mutation in a particular gene (IDH1), with a significantly better survival prognosis. Although methylation data was available in the earlier analysis, their conclusions were different—the IDH1-subtype was not identified because the subtypes were largely based on clustering mRNA expression data.[7] From a technical standpoint, this is not surprising, because the recovered subtypes are a function of the data, the clustering approach, and the associated notion of similarity used. When the integrated data are high dimensional and heterogeneous, defining a coherent metric for clustering becomes increasingly challenging.

To ensure identification of clinically relevant subtypes, others have started to model distinct subgroups based on clinical hypotheses and perform follow-up analysis to identify molecular determinants of differences between these subgroups.[20] Naturally, as the molecular-level characterization of human diversity becomes continually more detailed—not only in terms of genetic information but also molecular measurements over time—the phenotypic manifestations of all these details

are often insufficiently represented by available broad disease categories, which capture only the end points rather than the exhaustive trajectories of disease development.[21] The advent of EHRs significantly advances our ability to describe subtypes that are based on more precise and detailed descriptions of the disease.

Electronic Health Records and Phenotyping

The Health Information Technology for Economic and Clinical Health (HITECH) Act, which was part of the American Recovery and Reinvestment Act of 2009, incentivized the adoption of EHRs. As a result, today much of an individual's health data—such as demographics, personal and family medical history, diagnosis codes, current and past treatments, history of allergic reactions, vaccination records, laboratory test results, and imaging results—is stored in an EHR. Academic medical centers such as Vanderbilt University have succeeded in linking EHR data to biobanked blood samples that were accrued during routine clinical care. Over repeat clinic visits, from a research perspective, such integrated patient data constitute a computable collection of fine-grained longitudinal clinical profiles.

However, data contained within EHRs present numerous computational challenges. For example, even the seemingly simple task of extracting the list of conditions that an individual has been diagnosed with is nontrivial. Although ICD-9 codes—those indicating the presence or absence of a condition—are routinely encoded for billing, they are incomplete and noisy. Thus, much research has focused on developing new algorithms for annotating patient records with this information; in the literature, this task of extracting various clinical attributes for each individual is referred to as *phenotyping*. Phenotyping algorithms implement probabilistic rules[22] that combine information from the structured data (for example, demographics, diagnoses, medications, and laboratory measurements) as well as unstructured clinical text (such as radiology reports, encounter notes, and discharge summaries) to annotate attributes such as the presence or absence of specific diagnoses or medical adverse events (see the bibliography of a recent *JAMIA* paper by Jyotishman Pathak and colleagues[23]). The Electronic Medical Records and Genomics (eMERGE) consortium—a network of nine academic medical centers—has demonstrated the successful use of these EHR-derived phenotypes for cohort identification to conduct genome- and phenome-wide association studies.[9,24] These efforts have replicated genetic risk factors for many diseases, including Alzheimer's, type 2 diabetes, and arrhythmia.[24]

It is important to remember that phenotyping facilitates cleaner, but still broad, categories of disease. The power of EHRs lies in their detailed and longitudinal data, which makes it possible to obtain more refined subtypes. Thus, recent efforts have built on phenotyping work to move away from analysis of broad disease categories and instead provide more descriptive definitions of the disease over time.

Clinically Enriched Subtypes

Existing approaches have used different hypotheses to cluster individuals into subtypes using EHR data. In autism, for example, using hierarchical agglomerative clustering over individuals based on their set of comorbid conditions (secondary complications) as defined by ICD-9 codes, Finale Doshi-Velez and colleagues show different clinical subtypes: patients with some of the subtypes are more likely to experience seizures, whereas others are more likely to experience auditory complications.[25] Others have clustered vectors summarizing the procedure and diagnoses counts for each individual using tensor factorization to identify clusters of diagnosis codes and medications that co-occur.[26] Joyce Ho and colleagues show that this approach can automatically discover mild and severe forms of a disease based on a combination of commonly observed comorbidity and treatment patterns.[26] These approaches leverage treatment and diagnosis codes. The main advantage of using such data is that they are readily available in administrative billing records, which have standardized formats. However, subgroups based on ICD-9 codes and treatment data are highly sensitive to practice patterns.

Others have thus sought to leverage the rich array of clinical and laboratory data (such as glucose levels, blood counts, and functional test results) for measuring disease activity. For example, David Chen and colleagues show that by clustering the *clinarray*—a vector containing summary statistics such as the mean value of a marker over time—in diseases like Crohn's and cystic fibrosis, they can distinguish mild from severe forms of disease.[27] Alternatively, in disease progression modeling,[28] generative models of the disease progression—which characterize disease as a continuum or as a set of discrete stages and observed clinical data as a function of the stage—are learned from data. Past works have developed dynamic models that characterize progression as a function of comorbidities, or individual markers. These are typically developed with the motivation of tracking and predicting the progression (or severity) of an individual's disease over time. However, these models can also help identify patients with similar disease-progression patterns and thus enable

discovery of molecular variations that might be driving individuals to express different forms of the disease.

For example, Peter Schulam and colleagues devised an approach that uses continuous laboratory and clinical markers measured over time to identify subtypes with similar disease trajectories.[3] Figure 1a shows four subtypes based on their lung progression patterns, ranging from healthy and stable lung function (top left) to poor and consistent decline (bottom right). In their work, rather than clustering the raw measurements themselves, Schulam and colleagues incorporate knowledge of the measurement process and known disease process to account for sources of variation that affect the measured markers but are unrelated to the underlying subtype. For example, two individuals might show similar progression, but a chronic smoker will tend to have slightly worse lung function (modeled using random effects by a global shift). Thus, Schulam and colleagues removed these nuisance sources of variability when inferring subtypes.[3]

Figure 1b shows three different subtypes that Schulam and colleagues uncovered using joint analysis of their lung and skin markers (the two key complications of scleroderma). These recovered subtypes are associated with different autoantibody—a type of protein produced by the immune system and known to be a cause of autoimmune diseases—markers (see the top of Figure 1b) and comorbidity patterns (see the bottom of Figure 1b).

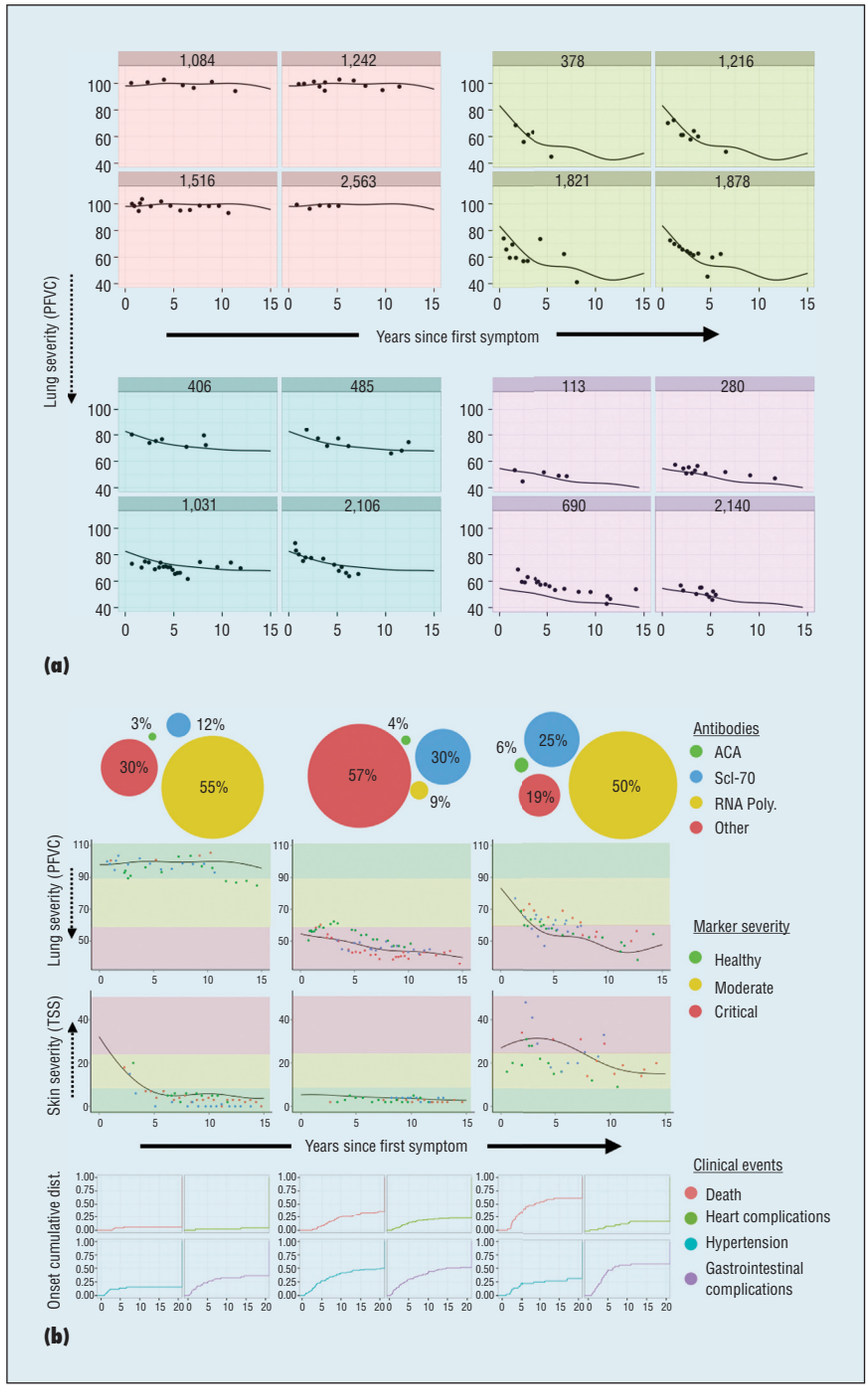

Figure 1. Subtyping driven by disease activity trajectories. (a) Four subtypes of scleroderma based on lung disease activity trajectories tracked over 15 years. The top left subtype shows stable lung function, whereas the bottom right shows active decline. These are inferred using the probabilistic subtyping model by Peter Schulam and colleagues based on continuous, sparse, and irregularly sampled measurements (shown as black dots) of the forced vital capacity, a marker of lung disease.[3] These measurements were taken during clinical visits as part of routine care. (b) Joint analysis of the lung and skin trajectories yields subpopulations that show distinct autoantibody profiles and comorbidity patterns (three distinct subtypes are shown). The top panel shows autoantibody prevalence (using size).

Next Steps: Integrative Analysis

In the next decade, affordable access to molecular data collection linked with rich clinical data has the potential to significantly advance our understanding of how diseases are defined and treated. In the near term, subtype defi-

nitions that enable accurate prognoses of disease trajectories can enable treatment planning and prognosis. In the longer term, as molecular mechanisms governing different disease subtypes are discovered, novel biomarkers that are predictive of the disease course, and novel treatments inspired by the mechanistic pathways, will become possible.

To achieve these goals, we will need careful integration of the diverse data surrounding an individual's health—molecular, clinical, and environmental data. To summarize, the goals of integrative analysis are twofold. The first goal is to identify naturally occurring subpopulations whose presentation in the clinic differ so that one can tailor treatments to these subgroups. For example, by being able to detect individuals at high risk for a given complication early, one can tailor the use of more aggressive therapies, when available. The second goal is to identify causal pathways associated with the phenotypically differentiated subpopulations. Causal pathways facilitate development of treatment programs appropriate to each subtype.

Although we have made tremendous progress toward leveraging the diverse molecular and clinical datasets, much remains to be done. For example, most of the existing methods for joint analysis of more than one data type[16,17] tend to treat the different measurement types as independent sources of information. However, in practice, we know that many of these measurements are interdependent. For example, DNA methylation affects levels of mRNA expression, and miRNA regulates gene expression post transcriptionally. Thus, inferences made using an assumption of independence are likely to be biased. Moreover, exploiting the relationship between these measurements can help reduce the effective dimensionality of the problem. This is especially useful when integration is being done in a supervised setting (for example, predicting markers of drug response) and the number of samples corresponding to the individual subpopulations is small. In works that have tackled joint analysis and model dependencies, typically only a small range of data types is modeled within any given study (see references within a paper by Marylyn Ritchie and colleagues[20]). As consortia-based efforts are gearing up to collect whole-systems level data for many patients, availability of data may no longer be a bottleneck, but rather the availability of analytic approaches that can integrate data at multiple resolutions, from the cell to the organ level.

Another challenge for integrative analysis arises from the heterogeneity of the different data types: some markers are continuous while others are categorical; some are measured only once (for example, gender or DNA sequence) while others are measured repeatedly (such as blood cell counts and functional lung tests). Different sources of measurement noise and bias can also affect the measurements made. For example, functional measurements (such as those shown in Figure 1a) can vary depending on the individual making the measurement, the altitude at which the measurement was made, and whether the patient is experiencing temporary inflammation. Similarly, the mRNA expression levels can vary depending on the measured sample's number and composition of cell types. Finally, the healthcare process itself governs which measurements are taken and recorded, and when.[29] These are nuisance sources of variability that are unrelated to subtype and should be accounted for in inferring meaningful subtypes.[3] Multiresolution models that integrate diverse markers incorporating knowledge of the measurement process and the biology are likely to be most fruitful.

On a pragmatic level, researchers who wish to analyze large amounts of patient data still face the technical challenges of integrating scattered, heterogeneous data, in addition to ethical and legal obstacles that limit access to the data. Infrastructure investments at the national, regional, and institutional levels are needed to make integrated data sources readily available for research.

As molecular data linked with rich clinical data are becoming easily accessible, new integrative methods for subtyping have the potential to significantly advance our understanding of how diseases are defined and treated. We call on the computational community to participate in this exciting computational task that can ultimately improve the quality of healthcare for us all. ⬒

Acknowledgments

Saria thanks Google Research, the Gordon and Betty Moore Foundation, and the National Science Foundation for their support and Peter Schulam for creating Figure 1. Goldenberg thanks the SickKids Foundation for support. Both authors thank Daniel Neill for valuable feedback on the article.

References

1. L. Hood and S.H. Friend, "Predictive, Personalized, Preventive, Participatory (P4) Cancer Medicine," *Nature Rev. Clinical Oncology*, vol. 8, no. 3, 2011, pp. 184–187.
2. R. Mirnezami, J. Nicholson, and A. Darzi, "Preparing for Precision Medicine," *New England J. Medicine*, vol. 366, no. 6, 2012, pp. 489–491.
3. P. Schulam, F. Wigley, and S. Saria, "Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery," Am. Assoc. for Artificial Intelligence, 2015; http://pschulam.com/papers/schulam+wigley+saria_aaai_2015.pdf.

4. S. Saria, "A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery," *IEEE Intelligent Systems*, vol. 29, no. 4, 2014, pp 82–87.
5. J. Lotvall et al., "Asthma Endotypes: A New Approach to Classification of Disease Entities within the Asthma Syndrome," *J. Allergy and Clinical Immunology*, vol. 127, no. 2, 2011, pp. 355–360.
6. M.R. Boland et al., "Defining a Comprehensive Verotype Using Electronic Health Records for Personalized Medicine," *J. Am. Medical Informatics Assoc.*, vol. 20, 2013, pp. e232–e238.
7. R.G. Verhaak et al., "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, 2010, pp. 98–110.
8. S.J. Lewis et al., "Heterogeneity of Parkinson's Disease in the Early Clinical Stages using a Data Driven Approach," *J. Neurology, Neurosurgery, and Psychiatry*, vol. 76, no. 3, 2005, pp. 343–348.
9. A.N. Kho et al., "Electronic Medical Records for Genetic Research: Results of the Emerge Consortium," *Science Translational Medicine*, vol. 3, no. 79, 2011, doi:10.1126/scitranslmed.3001807.
10. I.S. Kohane, "Using Electronic Health Records to Drive Discovery in Disease Genomics," *Nature Rev. Genetics*, vol. 12, no. 6, 2011, pp. 417–428.
11. J. Ewing, "Diffuse Endothelioma of Bone," *Proc. New York Pathological Soc.*, vol. 21, 1921, pp. 17–24.
12. J.P. Brunet et al., "Metagenes and Molecular Pattern Discovery Using Matrix Factorization," *Proc. Nat'l Academy Sciences*, vol. 101, no. 12, 2004, 4164-4169.
13. C.M. Perou et al., "Molecular Portraits of Human Breast Tumours," *Nature*, vol. 406, no. 6797, 2000, pp. 747–752.
14. V.N. Kristensen et al., "Principles and Methods of Integrative Genomic Analyses in Cancer," *Nature Rev. Cancer*, vol. 14, no. 5, 2014, pp. 299–313.
15. R. Shen, A.B. Olshen, and M. Ladanyi, "Integrative Clustering of Multiple Genomic Data Types using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis," *Bioinformatics*, vol. 25, no. 22, 2009, pp. 2906–2912.
16. P. Kirk et al., "Bayesian Correlated Clustering to Integrate Multiple Datasets," *Bioinformatics*, vol. 28, no. 24, 2012, pp. 3290–3297.
17. B. Wang et al., "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale," *Nature Methods*, vol. 11, no. 3, 2014, pp. 333–337.
18. J.M. Nigro et al., "Integrated Array-Comparative Genomic Hybridization and Expression Array Profiles Identify Clinically Relevant Molecular Subtypes of Glioblastoma," *Cancer Research*, vol. 65, no. 5, 2005, pp. 1678–1686.
19. D. Sturm et al., "Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma," *Cancer Cell*, vol. 22, no. 4, 2012, pp. 425–437.
20. M.D. Ritchie et al., "Methods of Integrating Data to Uncover Genotype-Phenotype Interactions," *Nature Rev. Genetics*, vol. 16, no. 2, 2015, pp. 85–97.
21. P.B. Jensen et al., "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care," *Nature Rev. Genetics*, vol. 13, no. 6, 2012, pp. 395–405.
22. S. Saria et al., "Combining Structured and Free-Text Data for Automatic Coding of Patient Outcomes," *Proc. AMIA Ann. Symp.*, 2010, p. 712.
23. J. Pathak, A.N. Kho, and J.C. Denny, "Electronic Health Records-Driven Phenotyping: Challenges, Recent Advances, and Perspectives," *J. Am. Medical Informatics Assoc.*, vol. 20, no. e2, 2013, pp. e206–e211.
24. J.C. Denny et al., "Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data," *Nature Biotechnology*, vol. 31, no. 12, 2013, pp. 1102–1111.
25. F. Doshi-Velez, Y. Ge, and I. Kohane, "Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis," *Pediatrics*, vol. 133, no. 1, 2014, pp. e54–63.
26. J.C. Ho, J. Ghosh, and J. Sun, "Marble: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization," *Proc. 20th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2014, pp. 115–124.
27. D.P. Chen et al., "Clinical Arrays of Laboratory Measures, or 'Clinarrays,' Built from an Electronic Health Record Enable Disease Subtyping by Severity," *Proc. AMIA Ann. Symp.*, 2007, pp. 115–119.
28. D. Mould, "Models for Disease Progression: New Approaches and Uses," *Clinical Pharmacology & Therapeutics*, vol. 92, no. 1, 2012, pp. 125–131.
29. G. Hripcsak and D.J. Albers, "Next-Generation Phenotyping of Electronic Health Records," *J. Am. Medical Informatics Assoc.*, vol. 20, no. 1, 2012, pp. 117–121.

Suchi Saria is an assistant professor in the Departments of Computer Science and Health Policy and Management at Johns Hopkins University. Contact her at ssaria@cs.jhu.edu.

Anna Goldenberg is a scientist in genetics and genome biology at the SickKids Research Institute and an assistant professor in the Department of Computer Science at the University of Toronto. Contact her at anna.goldenberg@utoronto.ca.

cn *Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*