VIEWPOINT

The Evolution of Patient Diagnosis
From Art to Digital Data-Driven Science

Kenneth D. Mandl, MD, MPH
Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts.

Florence T. Bourgeois, MD, MPH
Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts.

Physicians are still taught to diagnose patients according to the 19th-century Oslerian blueprint. A physician takes a history, performs an examination, and matches each patient to the traditional taxonomy of medical conditions. Symptoms, signs, family history, and laboratory reports are interpreted in light of clinical experience and scholarly interpretation of the medical literature. However, diagnosis is evolving from art to data-driven science, whereby large populations contextualize each individual's medical condition. Advances in artificial intelligence now bring insight from population-level data to individual care; a recent study sponsored by and including researchers from Google used data sets with more than 11 000 retinal fundus images to develop a deep learning algorithm that outperformed clinicians for detecting diabetic retinopathy.[1]

However, when clinicians make genetic diagnoses, they are practicing more like Osler than Google. The pathogenicity of a genetic variant is often determined from cohort studies of relatively small numbers of individuals. Even a simple comparison of a patient's variant to a larger population of matched ancestry is generally not possible. Manrai et al[2] illustrated the pitfalls of this approach, showing that monogenic variants considered diagnostic of hypertrophic cardiomyopathy, in fact, have a high frequency in unaffected individuals of African ancestry and, therefore, often apparently represent normal variants among black patients. This problem is more general and many studies implicate pathogenic variants, yet lack sufficient numbers of ancestrally diverse cases and controls. Furthermore, Van Driest et al[3] reviewed electronic health record (EHR) data and electrocardiographs in a cohort of 2022 genotyped patients and found that the majority of participants (41 of 63) with a designated variant in either *SCN5A* or *KCNH2*—putatively associated with cardiac rhythm disturbances—had no identifiable pathological phenotype.

Artificial intelligence will eventually help clinicians extract the maximum knowledge from large genetic reference data sets. But the first step is to simply be able to calculate the statistical genetics that reveal how often a variant is associated with pathology, and how outcomes compare in patients with and without the variant. To make a genetic diagnosis, a physician must evaluate a patient's data against a larger and representative population. A high-functioning health care system not only needs the EHR databases produced as a by-product of care, but also must link EHRs to samples, sequence data, and myriad data sources needed to characterize medical care, lifestyle, and environment.

Initiatives to develop genetic reference data at the population level could be grouped into 3 categories. First are well-known databases of genotype-phenotype relationships as observed and submitted by researchers (eg, Online Mendelian Inheritance in Man, ClinVar, and the National Human Genome Research Institute's Genome-Wide Association Study [GWAS] Catalog). Second are databases, such as the Genome Aggregation Database (gnomAD),[4] the next iteration of the Exome Aggregation Consortium (ExAC) database,[5] and the 1000 Genomes Project,[6] that aggregate sequences collected from other studies for secondary use. Third, patients and other study participants are invited to donate data to registries like GenomeConnect or enroll in cohorts like the National Institutes of Health All of Us initiative, which is recruiting 1 million patients to contribute biological samples and EHR data for research.

All 3 types of databases rely on a research framework rather than a clinical framework for accrual of patients and data. This distinction is important. The populations are selected by researchers or are self-selected, and the data are either deidentified or acquired after a research consent. The bias inherent in these populations may distort the accuracy of data-driven genomic diagnosis. Furthermore, even though deidentification addresses the Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy concerns, it often precludes future linkage of myriad data sets needed to create a robust information commons. The architects of All of Us have instrumented enrollment centers to provide ongoing longitudinal phenotype data from EHRs. Still, relying on consented research participants is not population-based and success is limited by the willingness of individuals to participate and the expense and logistics of consenting.

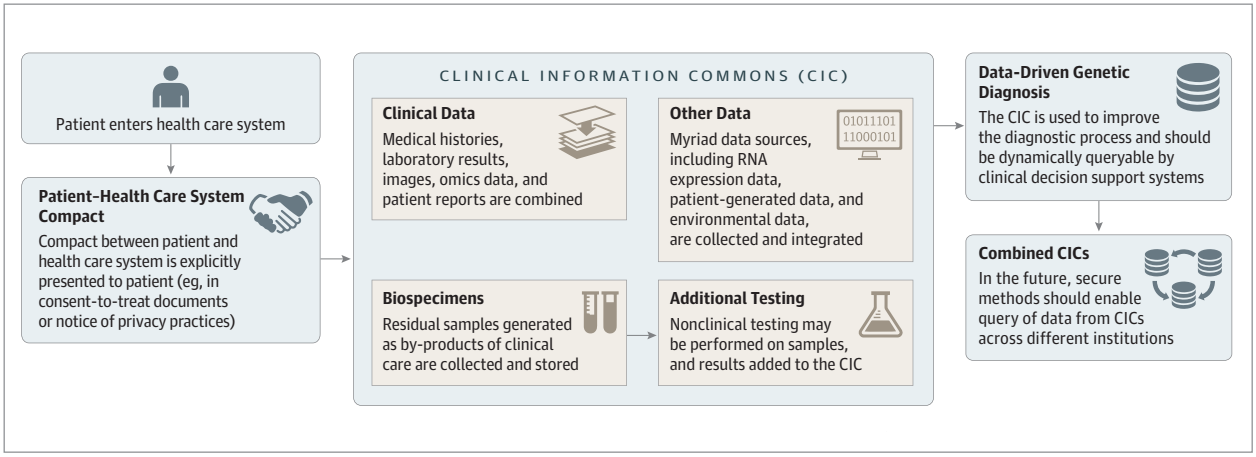
A Patient–Health System Compact to Create a Clinical Information Commons

As the care delivery system generates big data each day, should physicians rely on the exercise of a formal research framework to have data available to perform diagnosis? What dependencies on the research enterprise should patients and clinicians accept for an accurate diagnosis? Even today, a patient's genetic variant should be compared against a population-based reference to assess its pathogenicity and to inform prognosis and treatment based on the care trajectories and outcomes of other patients with the variant or similar variants in a particular gene.

In the National Research Council report, Desmond-Hellmann and coauthors provided the important reminder that it is patients who "uniquely understand the potential value of a social contract in which patients both contribute personal clinical data and benefit from the knowledge gained through the collaboration." Accordingly, there should be a new compact between patients and the health system, such that captured data and biospecimen by-products of the care delivery system should be aggregated and linked to build a clinical information commons (CIC) to aid diagnosis (**Figure**).

Corresponding Author: Kenneth D. Mandl, MD, MPH, Boston Children's Hospital, 300 Longwood Ave, Boston, MA 02115 (kenneth_mandl@harvard.edu).

Figure. The Clinical Information Commons to Support Data-Driven Diagnosis



Patients are engaged in a compact with the health system whereby clinical data and biological samples derived from routine clinical care are aggregated to support innovation and improvement in diagnosis. Data and samples are collected across visits, accumulating over time for individual patients. Institutional data could also be combined across care sites to broaden the patient population and clinical data represented.

For clarity and transparency, the patient should be informed of the compact in writing. This could be accomplished through the consent-to-treat document signed in prelude to care or the notice of privacy practices. The compact requires a patient understanding that residual volumes from blood and tissue samples may be stored, linked to EHR data, and used in a CIC for a patient's own medical care or the care of others. Patients could benefit from the compact by having a care system in which diagnosis may be more accurate.

Hospitals could implement the requisite informatics and laboratory processes to capture residual biospecimens, including blood and pathology specimens (Figure). Additional testing beyond what is used clinically could be performed on these specimens and the results integrated. The CIC should be available to decision support systems in EHRs. In the process of diagnosis, the CIC could be queried in a dynamic fashion, such that an individual can be compared with a population based on her characteristics

The implementation of a CIC is sociologically, financially, and technically complex. Only larger institutions could implement the necessary biobanking. Importantly, the institutional CIC would be reflective of the patient population treated at that institution. Even though an institution-specific CIC is useful, joining deidentified data from CICs across institutions would produce even greater benefit.

Protecting patient privacy and confidentiality is paramount and combined resources that can be queried would need appropriate security, access controls, governance, and privacy protections.

Fortunately, there is a robust informatics and policy basis for privacy-preserving data query and provisioning. As individuals increasingly gain access to their own health system data and direct services based on those data, these linked data may be provisioned to patients for their own use; for example, patients may be able to use apps for managing medications based on pharmacogenomic variants. If additional testing is performed on samples beyond what was clinically indicated, policies and procedures would need to address incidental findings.

A robust CIC is needed for clinical assessments of the utility of burgeoning genetic information that are tied directly to patient outcomes across diverse populations. However, changing the fundamental approach to diagnosis, a core function of medical practice, warrants caution and careful assessment. In medicine, more information is being consistently generated, but little is known about its effect on care.

Conclusions

Incorporating CICs as a routine component of medical treatment facilities could exponentially accelerate the acquisition of samples and omics data linked to medical record data. Rather than making preventable diagnostic errors in the future, clinicians and researchers should engage in a compact with patients to create CICs with maximum representation across the population so the full "normal curve" can underpin digitally driven genetic diagnosis.

ARTICLE INFORMATION

Published Online: October 26, 2017. doi:10.1001/jama.2017.15028

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

Funding/Support: This article was funded by the PrecisionLink initiative at Boston Children's Hospital.

Role of the Funder/Sponsor: Boston Children's Hospital had no role in the preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: We thank Arjun Manrai, PhD, Sek Won Kong, MD, and Patrick Taylor, JD, for insightful input. They did not receive compensation for their contributions.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.

2. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med*. 2016;375(7):655-665.

3. Van Driest SL, Wells QS, Stallings S, et al. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA*. 2016;315(1):47-57.

4. gnomAD browser beta. About gnomAD. http://gnomad.broadinstitute.org/. Accessed October 9, 2017.

5. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60 706 humans. *Nature*. 2016;536(7616):285-291.

6. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.