Vol. 3, No. 2 | February 2015
RENCI WHITE PAPER SERIES

Clinical Genomics: How Much Data is Enough?

The Team

JAMES P. EVANS, MD, PhD
Professor, Departments of Genetics and Medicine, University of North Carolina at Chapel Hill

KIRK C. WILHELMSEN, MD, PhD
Professor, Departments of Genetics and Neurology, UNC–Chapel Hill

JONATHAN BERG, MD, PhD
Assistant Professor, Department of Genetics, UNC–Chapel Hill;

CHARLES P. SCHMITT, PhD
Director of Informatics and Chief Technical Officer, Renaissance Computing Institute (RENCI)

ASHOK KRISHNAMURTHY, PhD
Deputy Director, RENCI

KARAMARIE FECHO, PhD
Medical and Scientific Writer for RENCI

STANLEY C. AHALT, PhD
Director, RENCI and Professor, Department of Computer Science, UNC–Chapel Hill

Contact Information

STANLEY C. AHALT
Email: ahalt@renci.org
Telephone: 919-445-9641.

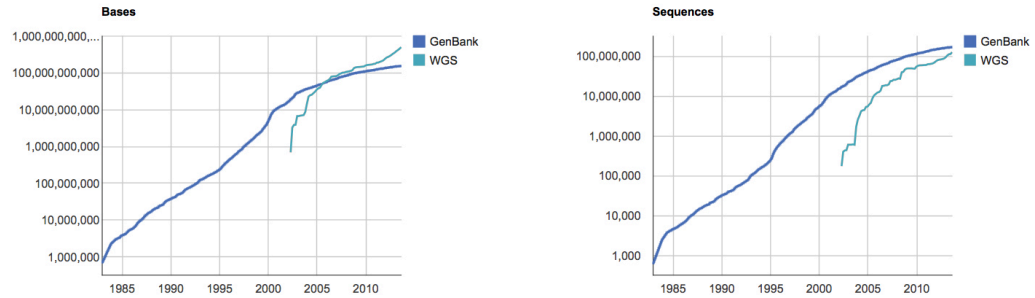
Summary

Genome-scale sequencing data have yet to be widely used in clinical medicine. Indeed, it is not clear whether the routine accumulation of massive amounts of largely uninterpretable genomic data will yield a net benefit in terms of improving health. Nevertheless, the use of genomic data is certain to grow, and the medical community will need to consider how much genomic data to store and how to interpret and communicate those data to both patients and providers. To meet these emerging challenges and ultimately facilitate the optimal use of genomic data, we advocate for the development of a two-pronged Genomic Clinical Decision Support System that encompasses the concept of the *Clinical Mendeliome* and introduces the concept of the *Archival Value Criterion*. The model we propose is designed to stimulate effective clinical use of genomic data, drive genomic research, and meet both current and future needs in medicine and research.

The Challenges

It is now possible to sequence and acquire large amounts of genomic data within a reasonable time frame and at a manageable cost; yet our ability to analyze, interpret, and use genomic findings to guide clinical care lags far behind. The GenBank repository currently contains data on an astonishing ~150 billion bases and more than 150 million sequences, and the number of bases has doubled roughly every 18 months from 1982 to the present (See Figure 1). While the GenBank repository is not restricted to human data, the growth in the data is remarkable and serves as a testament to the pressing need to develop tools to understand and apply those data, both human and non-human, to improve the human condition. Indeed, efforts are underway to annotate, archive, and curate genomic data, and publicly available databases such as ClinVar, dbGap, DGV, and the Genetic Association Database now contain data on thousands of genes and gene variants, many with extensive annotation on functional impacts and validated phenotypes. Moreover, the National Institute of Health's Clinically Relevant Genetic Variants Resource initiative and the Health Level Seven Clinical Genomics working group are pioneering efforts to harmonize individual efforts and databases.

FIGURE 1.
Growth in genomic data stored in GenBank, expressed in terms of bases (left) and sequences (right). WGS = Whole Genome Sequencing (Whole Genome Shotgun). (Figure was freely available from GenBank at http://www.ncbi.nlm.nih.gov/genbank/statistics.)



AT A GLANCE

- Genome-scale sequencing is now economically feasible, yet routine clinical use of genomic data is hindered by challenges related to how much data should be stored in electronic medical records and best practices in the ethical use of those data in order to improve patient health.
- We envision a two-pronged Genomic Clinical Decision Support System that encompasses the concept of the *Clinical Mendeliome* as a patient-centric list of variants that are clinically actionable and introduces the concept of the *Archival Value Criterion* as a decision-making formalism that approximates the cost-effectiveness of capturing, storing, and curating whole genome or whole exome sequencing data.
- Our model addresses the perspectives of all stakeholders, from the patient to the clinical care provider to the genomics researcher, and is designed to meet both current and future needs.
- As a first step toward implementation of our model, we recommend the creation of a *Task Force on the Clinical Mendeliome*, charged with defining context-specific Clinical Mendeliomes and creating an initial set of Clinical Practice Guidelines for their use.

While these data sources are widely used for research, they are not yet applicable for clinical use. Thus, clinical genetic testing typically remains highly targeted, with the majority of genetic tests focused on discrete clinical situations and the relatively small number of genes that exhibit high penetrance and result in recognizable disease when mutated. Those genes for which sequence information clearly provides clinically actionable information (e.g., *BRCA1/2*, *MLH1*, *MSH2*) represent an even more constrained subset.

It remains to be seen whether genome-scale sequencing will be widely embraced in clinical medicine. It may transpire that because of falling sequencing costs and the provision of additional information, Whole Genome Sequencing (WGS) and/or Whole Exome Sequencing (WES) will become the most efficient strategy when any genetic query is clinically indicated and will eventually be applicable to the general population. On the other hand, it is not yet clear whether the routine accumulation of massive amounts of (mostly uninterpretable) data from genome-scale sequencing will yield a net benefit in terms of improving health. Indeed, the pursuits of both clinical medicine and public health have traditionally been best served by narrow testing based on substantiated evidence. The application of overly-broad testing brings with it many potential problems, including inevitable false positive results, ambiguous results that beg for misinterpretation, and considerable downstream costs no matter how inexpensive the upfront cost of testing. Nevertheless, genome-scale analysis is likely to become more common, and the medical enterprise needs to grapple with the central question of how much genomic data should be stored and in what format.

A Framework for Genomic Clinical Decision Support Systems

To meet these emerging challenges and facilitate the optimal use of genomic data, we advocate the development of a two-pronged Clinical Decision Support System (CDSS) for genomics that will: (1) provide the clinician with a dynamic visual snapshot of only those genomic data that are relevant to an individual patient; and (2) capture, store, and curate more comprehensive genomic data for ready access to address future clinical needs and enable genomic research.

The concept of a CDSS dates to a seminal manuscript by Ledley and Lusted (1959) that provided a mathematical foundation for the application of computational techniques to improve medical diagnosis. Subsequent work built on that foundation, including an oft-cited study by de Dombal et al. (1972), who conducted a controlled clinical trial demonstrating that a CDSS produced significantly greater accuracy and reliability in the diagnosis of acute abdominal pain than did an attending physician. Numerous CDSS capabilities are now routinely incorporated into a diverse array of medical devices (as safety alerts and reminders) but have yet to be fully integrated into the healthcare environment (Mitchell, et al. 2011). The emergence of genomic medicine, coupled with increasing utilization of electronic medical records (EMRs) (Charles, et al. 2013), now provides a compelling stimulus for more widespread incorporation of Genomic CDSSs into the healthcare workflow.

Arguments in favor of the development of Genomic CDSSs are not new (Hoffman 2007), and theoretical prototypes have been developed (e.g., Douali & Jaulent 2012; Masys, et al. 2012), as well as a few clinical prototypes (e.g., Neri, et al. 2012; Foreman, et al. 2013). The functional and technical challenges also have been thoroughly laid out and include a lack of data standards, difficulties integrating multiple sources of data, absence of consensus in the adjudication of variants, dearth of standardized phenotype data, incomplete or missing annotation on methodology, and issues related to privacy and quality control (Kawamoto, et al. 2009; Ahalt, et al. 2014; Jacob, et al. 2013; Marsolo & Spooner 2013). However, a critical issue that has not received adequate attention is the amount and type of genomic data that should be incorporated into a Genomic CDSS to satisfy the needs of all stakeholders, including clinicians, patients, researchers, healthcare insurance providers, and hospital or medical center administrators and legal representatives. This issue is central to the implementation of genomic medicine. As with all medical endeavors, the penalty from too little information is readily apparent—but grave penalties also loom when excessive, redundant, and uninterpretable information populates the EMR, risking pervasive harm through distraction, misinterpretation, and increased cost.

Historically, CDSSs have been viewed as tools for the clinician to use to improve patient care. This perspective likely arose and persisted as a result of the initial use of CDSSs as computer-assisted diagnostic tools. We argue that this clinician-centric focus, while valuable, should be expanded from providing mere passive alerts for the clinician to a proactive model that encompasses the perspectives of the diverse stakeholders who participate in both patient care and research. Such a change in framework can in turn guide decisions about which genomic

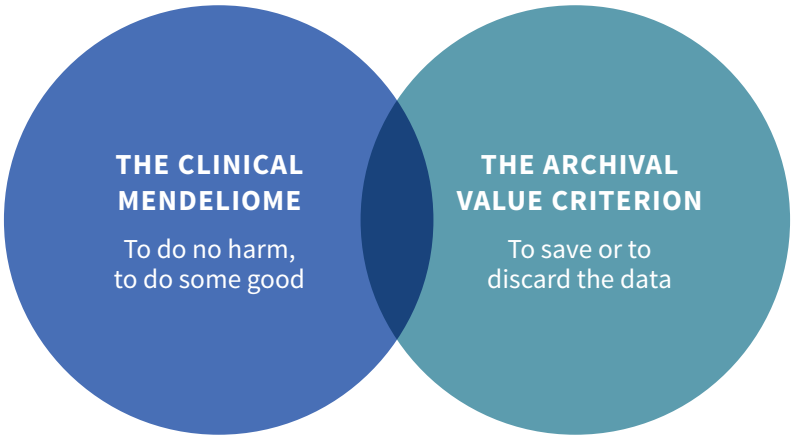
data should be incorporated into the EMR and associated Genomic CDSS. The contrast between our ability to cheaply collect and store massive amounts of genomic data and our inability to effectively use most of those same data forces us to navigate a balance between the *Scylla* of too little information and the *Charybdis* of littering the EMR with so much data that meaningful information is diluted to the point of irrelevance and concerns regarding provenance, governance, security, privacy, and storage costs become paramount.

A New Vision for a Genomic CDSS

We argue for a two-pronged solution to this challenge (Figure 2). For the first prong, we advocate the concept of the EMR retaining information focused upon the *Clinical Mendeliome*,[1] a patient-centric list of variants that potentially affect clinical care, guided by the principals of "do no harm" and "do some good." This approach addresses the dilemma of too much information and too little by first prioritizing variants found in genes known to be implicated in human disease (currently ~3,000 of ~22,000 genes). The variants to be included would be based upon evidence from published literature and public databases and emphasize a minimal set of genetic variants of known status with regard to utility, pathogenicity, and actionable treatment options. Importantly, these variants would be linked in the Genomic CDSS to pertinent aspects of the patient's demographics and history. Such a panel of variants in a select number of genes, the *Clinical Mendeliome*, could be automatically generated through a rules-based search engine or selected on the basis of a defined number of parameters chosen by clinicians (i.e., a clinical "order").

FIGURE 2.

A Proposed Model for a Genomic CDSS. *(Image courtesy of RENCI.)*



[1] We use the term "Clinical Mendeliome" not to suggest that all genetic diseases are Mendelian, but rather to capture the essence and simplicity of our approach, which targets only those variants with established clinical validity and utility.

The relevance of genomic variation for any given individual is critically affected by context. Thus, an individual patient's clinical status and demographics must guide the construction of variant lists in the Genomic CDSS such that those variants most relevant to a patient's care, in the appropriate context, would be presented for analysis. In healthy individuals (i.e., those without obvious disease), for whom genomic analysis is pursued as a screening modality for prevention (Evans, et al. 2013), a highly curated gene list based upon general demographics would be generated, and only those variants meeting a very high bar with regard to pathogenicity, predictive value, clinical utility, or pharmacogenomic relevance would be displayed in the Genomic CDSS. In each context, the Genomic CDSS must include a simple, dashboard-style interface that would display the mendeliome results for individual patients, along with complete annotation on each variant (including a measure of reliability of the variant) and an algorithm of actionable clinical measures that should be taken, again defined using evidence from the published literature and publically available databases. One could envision an extension of this interface to include a visual representation of variant data and treatment outcomes for a comparative population of patients with a similar genetic profile (Mane, et al. 2012). The mendeliome results could also be incorporated into a patient's Personal Health Record (Tang, et al. 2006), in line with the Meaningful Use incentives afforded by the Health Information Technology for Economic and Clinical Health Act of 2009 (Blumenthal & Tavenner 2010), as well as the transition from volume- to value-driven accountable care, and satisfying the long-term interests of patients, clinicians, and insurance providers.

For the second prong, we introduce the concept of the *Archival Value Criterion*, a decision-making formalism that addresses the dilemma of "to save" or "to discard" genomic data. We envision that WGS/WES data will be incorporated into the EMR as an underlying data layer (perhaps as vcf files) that includes extensive annotation on how the data were collected/stored/curated and a reliability measure(s) on the quality of the data. These data would not be proximal to the provider, so as to eliminate excessive and irrelevant data, but the data would be accessible and available for both future clinical needs (as new genomic information arises and the mendeliome is updated accordingly) and research purposes (with appropriate consent or de-identification; Gymrek, et al. 2013). The data would also be available for upgrade to the provider dashboard, allowing proactive monitoring and alerting for cases in which new annotation or pharmacogenomic indicators arise. The data could also be combined with (de-identified) data from the Genomic CDSS, thereby enabling clinical care to directly stimulate and guide new research.

The cost-effectiveness of capturing, storing, and curating WGS/WES-derived data is simple to calculate to a first approximation. We suggest the following formalism as the *Archival Value Criterion (AVC)* for determining cost-effectiveness: $AVC = (P_{reuse} \times S')/S$, where S is the total cost for the storage and curation of sequencing data, P_{reuse} is the estimated probability of reuse, and S' is the cost of regeneration. We propose that an AVC metric $>10^2$ suggests data archiving rather than data regeneration. After data have been archived, the AVC metric for removal should be much lower, perhaps $<10^{-6}$. We note that the AVC is influenced by numerous factors, including compression, secondary factors such as staff time and resource reallocation, and time-dependent parameters (see Wilhelmsen et al. [2013] for a detailed discussion), but the AVC as presented herein provides a useful formalism to estimate the cost-effectiveness of archiving genomic data.

The Upshot

Our vision of a two-pronged Genomic CDSS encompasses the concept of the *Clinical Mendeliome* and introduces the concept of the *Archival Value Criterion*. Our model embraces the perspectives and facilitates the goals of stakeholders from proximal clinical encounter to distal research endeavors. Such an approach enables genomic medicine to facilitate genomics research and *vice-versa* and provides incentives to promote industry commoditization of the stored genomic data.

The Big Picture

As a first step toward realizing our vision and implementing our Genomic CDSS model, we recommend the creation of a *Task Force on the Clinical Mendeliome*. The task force would be charged with defining context-specific Clinical Mendeliomes and creating an initial set of Clinical Practice Guidelines for their use. The task force would be comprised of various stakeholders with expertise in clinical genomics, genetics, information technology, computer science, and bioethics and include, for example, members of the American College of Medical Genetics and the National Consortium for Data Science, as well as members representing the Clinically Relevant Genetic Variants Resource initiative and the Health Level Seven Clinical Genomics working group. By bringing together stakeholders with different interests and areas of expertise, the task force would be well positioned to identify solutions to the inevitable challenges that will arise when addressing the central question of "how much data is enough" for the care of patients and furtherance of research.

ACKNOWLEDGMENTS

Karen Green, Director of Communications & Outreach for RENCI, provided editorial support for the preparation of this white paper. Design was provided by UNC Creative.

REFERENCES*

Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science*, 130 (3366), 9-21.

de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., & Horrocks, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal*, 2 (5804), 9-13.

Mitchell, J. A., Gerdin, U., Lindberg, D. A., Lovis, C., Martin-Sanchez, F. J., Miller, R. A., Shortliffe, E. H., & Leong, T.-Y. (2011). 50 years of informatics research on decision support: what's next. *Methods of Information in Medicine*, 50, 525-535.

Charles, D., King, J., Patel, V., & Furukawa, M. F. (2013). Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008–2012. ONC Data Brief, No. 9, March 2013. www.healthit.gov/sites/default/files/oncdatabrief9final.pdf.

Hoffman, M. A. (2007). The genome-enabled electronic medical record. *Journal of Biomedical Informatics*, 40 (1), 44-46.

Douali, N., & Jaulent, M. C. (2012). Genomic and personalized medicine decision support system [conference paper]. IEEE 2012 International Conference on Complex Systems (ICCS), pp. 1-4. doi: 10.1109/ICoCS.2012.6458611.

Masys, D. R., Jarvik, G. P., Abernethy, N. F., Anderson, N. R., Papanicolaou, G. J., Paltoo, D. N., Hoffman, M. A., Kohane, I. S., & Levy, H. P. (2012). Technical desiderata for the integration of genomic data into electronic health records. *Journal of Biomedical Informatics*, 45 (3), 419-422.

Neri, P. M., Pollard, S.E., Volk, L. A., Newmark, L. P., Varugheese, M., Baxter, S., Aronson, S. J., Rehm, H. L., & Bates, D. W. (2012). Usability of a novel clinician interface for genetic results. *Journal of Biomedical Informatics*, 45, 950-957.

Foreman, A. K., Berg, J., Roche, M., Weck, K., Wilhelmsen, K., & Evans, J. (2013). A year of NCGENES: North Carolina Clincial Genomic Evaluation by NextGen Exome Sequencing [abstract]. Presented at the American College of Medical Genetics Annual Clinical Genetics Meeting, Phoenix, AZ, March 21, 2013. ww2.aievolution.com/acm1301/index.cfm?do=abs.viewAbs&abs=1570.

Kawamoto, K., Lobach, D. F., Willard, H. F., & Ginsburg, G. S. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. *BMC Medical Informatics and Decision Making*, 9, 17. www.biomedcentral.com/1472-6947/9/17.

Ahalt, S., Bizon, C., Evans, J., Erlich, Y., Ginsberg, G., Krishnamurthy, A., Lange, L., Maltbie, D., Masys, D., Schmitt, C., & Wilhelmsen, K. (2014). Data to discovery: Genomes to health. A White Paper from the National Consortium for Data Science. RENCI, University of North Carolina at Chapel Hill. dx.doi.org/10.7921/G03X84K4. data2discovery.org/publications.

Jacob, H. J., Abrams, K., Bick, D. P., Brodie, K., Dimmock, D. P., Farrell, M., Geurts, J., Harris, J., Helbling, D., Joers, B. J., Kliegman, R., Kowalski, G., Lazar, J., Margolis, D. A., North, P., Northup, J., Roquemore-Goins, A., Scharer, G., Shimoyama, M., Strong, K., Taylor, B., Tsaih, S.-W., Tschannen, M. R., Veith, R. L., Wendt-Andrae, J., Wilk, B., & Worthey, E. A. (2013). Genomics in clinical practice: lessons from the front lines. *Science Translational Medicine*, 5 (194), 194cm5.

Marsolo, K., & Spooner, S. A. (2013). Clinical genomics in the world of the electronic health record. *Genet Med*, 15 (10), 786-791.

Evans, J. P., Berg, J. S., Olshan, A. F., Magnuson, T., & Rimer, B. K. (2013). We screen newborns, don't we? Realizing the promise of public health genomics. *Genetics in Medicine*, 15 (5), 332-334.

Mane, K. K., Bizon, C., Schmitt, C., Owen, P., Burchett, B., Pietrobon, R., & Gersing, K. (2012). VisualDecisionLinc: a visual analytics approach for comparative effectiveness–based clinical decision support in psychiatry. *Journal of Biomedical Informatics*, 45 (1), 101-106.

Tang, P. C., Ash, J. S., Bates, D. W., Overhage, J. M., & Sands, D. Z. (2006). Personal health records: Definitions, benefits, and strategies for overcoming barriers to adoption. *Journal of the American Medical Informatics Association*, 13 (2), 121-126.

Blumenthal, D., & Tavenner, M. (2010). The "meaningful use" regulation for electronic health records. *New England Journal of Medicine*, 363 (6), 501-504.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339 (6117), 321-324.

Wilhelmsen, K., Schmitt, C. P., Fecho, K. (2013) Factors influencing data archival of large-scale genomic data sets. RENCI Technical Report TR-13-03. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/GOMW2F25. www.renci.org/publications/technical-reports.

*All hyperlinks were last accessed on February 4, 2015.

ABOUT RENCI:

RENCI, an institute of the University of North Carolina at Chapel Hill, develops and deploys advanced technologies to enable research, innovation, and economic development. For more information, see www.renci.org.

HOW TO REFERENCE THIS PAPER:

Evans, J.*, Wilhelmsen, K.*, Berg, J., Schmitt, C., Krishnamurthy, A., Fecho, K., & Ahalt, S. (2015): Clinical Genomics: How Much Data is Enough?. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0F769G9

*These authors contributed equally to this work.

renci

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL