REPORT OF

NIH Workshop on Enabling Research Use of Clinical Data

Big Data to Knowledge (BD2K) Initiative

September 11-12, 2013

Big Data to Knowledge (BD2K) is an initiative of the National Institutes of Health (NIH) that aims to enable biomedical scientists to access, manage and utilize effectively the large, complex data sets that are becoming increasingly commonplace in NIH-funded research, and expected to grow dramatically in number and complexity as 21st century biomedical[1] science evolves. The BD2K initiative was formulated in response to recommendations presented on June 12, 2012 by the Data and Informatics Working Group (DIWG) to the Advisory Committee to the Director, NIH. The DIWG report, which can be found at <http://acd.od.nih.gov/diwg.htm>, observed that:

> *Fueled by high-throughput laboratory technologies for assessing the properties and activities of genes, proteins and other biomolecules, the "omics" era is one in which a single experiment performed in a few hours generates terabytes (trillions of bytes) of data. Moreover, this extensive amount of data requires both quantitative biostatistical analysis and semantic interpretation to fully decipher observed patterns. Translational and clinical research has experienced similar growth in data volume, in which gigabyte-scale digital images are common, and complex phenotypes derived from clinical data involve data extracted from millions of records with billions of observable attributes. (p. 8)*

The report also noted that "*Confidentiality issues, as well as fundamental differences between basic science and clinical investigation, create real challenges for the successful integration of molecular and clinical datasets*" (p. 9).

To explore the issues that are particularly (and in some cases, uniquely) associated with clinical data, this workshop was convened with an invited, multidisciplinary group of individuals representing public and private sector organizations with interest and involvement in the use of clinical data for research. The workshop agenda and roster of invited participants are included as Appendices A and B, respectively. As preparatory material for the workshop, participants provided information about ongoing relevant initiatives and publications aimed at improving research use of clinical data. For access to this information, as well as the workshop presentations and an archived videocast of the complete workshop, the reader is referred to the BD2K workshop website <http://bd2k.nih.gov/bd2k_workshop/index.html>.

[1] In this document, as in other BD2K workshop reports, the term biomedical is used in the broadest sense to include biological, biomedical, behavioral, social, environmental, and clinical studies that relate to understanding health and disease.

Workshop format

Three prominent research use cases of clinical data were chosen to serve as the focal point for discussions - Pragmatic Trials & Interventional Studies, Genome-Phenome Correlation, and Observational Studies - along with an additional session focused on the cross-cutting issues of Infrastructure, Standards and Policy. Leaders of these sessions were asked to envision, if possible, a Utopian future scenario for their research focus area, describe differences between that future state and the current state, and then facilitate a group discussion of possible actionable steps that NIH could take (alone and with others) to enable progress toward greater use of clinical data to advance the NIH mission to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.

Definitions and Scope

As discussed by workshop participants and used in this report, the term clinical data is broadly interpreted to encompass data about humans that arises from a growing number of sources and contexts. Included are sources that have been traditionally labeled clinical, such as uniform research datasets arising from interventional and observational research studies involving human participants; data recorded by healthcare providers in Electronic Health Records (EHRs) maintained by hospitals and clinics, arising from inpatient admissions and outpatient encounters; and patient registries. Also considered within scope are newer, nontraditional forms of data, such as Personal Health Records (PHRs) maintained by individuals; direct-to-consumer genetic and genomic test results; output from mHealth apps running on personal devices like smartphones and tablet computers; data from "smart" devices like scales, glucometers, and peak flow meters; environmental data; and social media data from health-related Tweets, blogs, disease-oriented discussion groups and interactive websites. The rise of 21st century consumerism, fueled by ubiquitous connectivity and interactive data technologies, and a focus on 'patient-centered' rather than 'investigator-centered' research featured prominently in discussions of the changing landscape of biomedical research.

The DIWG report defined phenotype as:

> *"the composite of an organism's observable characteristics or traits… Although the term was originally linked to the concept of genotype, it is now often more loosely construed as groupings of observations and measurements that define a biological state of interest… Unlike specific data types familiar in computer science (such as text, integers, binary large objects), phenotypes are less well-defined and usually composed of a variable number of elements closely aligned to a particular research project's aims (e.g., the phenotype of a person with type II diabetes who has received a specific drug and experienced a particular adverse event)".*

Workshop participants expanded this definition to include important health states (like pain) that are not necessarily linked to well-defined biological states, recognizing that many of the best phenotype definitions we have now are experiential rather than biological.

Topic-specific Presentations

Because the didactic content presented by the four session leaders was intended to be a provocative seed crystal for group-wide discussion, a high level summary of main points made by presenters is included here.

Michael Lauer, MD, of the NIH, National Heart, Lung and Blood Institute, addressed Pragmatic Trials and Novel Interventional Cohort Studies. Relaying the story of Matthew Fontaine Maury (Mayer-Schonberger V, Cukier K, Houghton Miflin, 2013) who used neglected ship logs, maps and charts to deepen understanding of the physical geography of the sea to transform shipping routes, Lauer described data as a disruptive technology. Data does not always have to be perfect; it can be collected from existing sources, and used for other than original intent. He described a variety of recent, novel designs for interventional studies that rely on existing health data from EHRs, insurers, and patients, and advocated for use of disruptive technologies that can generate reliable findings while reducing the cost of cohort studies from thousands of dollars per patient to tens of dollars per patient. He noted that using big data from huge numbers of participants supports robust estimates of effects and accommodates interrogation of heterogeneity. Further, the creative use of existing infrastructure such as registries and national-scale cohorts can lead to very streamlined budgets with positive patient outcomes and survival rates. He proposed a new acronym for such trials: LEVI'S, meaning: Large, Leveraged, Embedded, External (taking advantage of existing resources), Valuable (they inform practice), Inexpensive, Innovative, and producing Sound Science.

Isaac Kohane, MD, PhD, from Harvard Medical School organized his presentation on Genomically-enabled Medicine as "where we were, where we are heading now, and where we could be." Using hurricane weather prediction as an analogy for high-volume sensing data being algorithmically interpreted and generating detailed region-specific guidance, he gave examples of similar approaches with relevance to biomedical research, and opined that the real goal of biomedical big data is to "find the true name for diseases"' their etiological basis using all available data and multiple scales of resolution. In this context, personal genomic variation is only a subset of the relevant data, which also prominently includes an individual's environment, broadly interpreted. Kohane advocated for being universal letting all potentially relevant data sources into the mix used for research and then weighting each data source for use and quality. He provided examples of novel data sources, such as data mined from public media reports for international epidemiology, and the crowdsourced data that sparked genomic discovery via the Clarity Challenge competition of Boston Children's Hospital <http://genes.childrenshospital.org/>. He envisions a future of patients as partners (peer-players) in healthcare research, with neither patients nor providers blinded to results, and of societal changes where healthcare systems are no longer the major source of clinical trial participants, rather "FlashTrials.gov" (analogous to flashmobs) where participants can be assembled in near-real time. Genomic data will grow explosively as whole genome sequencing will be no more expensive than an electrolyte panel, there will be ubiquitous physiologic monitoring, and lifetime health records maintained by third parties. He opined that in ten years each home will be capable of generating more health-related data than an ICU does now.

Greg Simon, MD, of Group Health in Seattle opened the session on Observational Research by giving historical examples of observational studies that were well designed but whose results were available too late to change clinical practice. Noting that current approaches are often analogous to "predicting yesterday's weather and delivering the prediction tomorrow", he advocated for making a reality of the Institute of Medicine's vision of a Learning Healthcare

System, where all experience contributes to the evolving evidence base for healthcare, continuously in real time. This is in contrast to a current environment that incentivizes secrecy (hide good ideas until the study is finished), stasis (do not answer questions too quickly), and research inefficiency (maximize grant dollars). He outlined three cultural challenges to achieving a more effective clinical research enterprise: 1) improving quality of the data (if the data is not good enough for research ("the tail"), it certainly is not good enough for patient care ("the dog"); 2) building a culture of transparency and trust; and 3) reforming a problematic business model of research. He proposed new approaches emphasizing transparency of many clinical and research activities as an approach to overcoming these challenges. If successful, research would become a beneficial contributor to higher healthcare quality, rather than just the "tail on the dog" of healthcare.

Brad Malin, PhD, of Vanderbilt University led the session on the cross-cutting issues of Infrastructure, Standards and Policy. In pursuit of a future clinical research environment that is characterized by transparency, trust, and timeliness he articulated a set of needs for technologies that mitigate risk for patients and facilitate research workflow, informed by acceptable use policies with accountability and effectiveness. These policies would engage patients and leverage 3rd party big data managers that have limited trust. He cited EHR access logs as a novel, emerging form of clinical big data, and reviewed the effects on research of the Health Insurance Portability and Accountability Act (HIPAA) as currently implemented. Characterizing the current approaches to obtaining research consent as non-scalable in an era of ubiquitous clinical big data, he discussed how the changing information and communications environment may provide pathways to solving problems that historically have been intractable. Overlapping and sometimes conflicting layers of policy have proliferated, such as those attached to HIPAA limited data sets, arising from local organizations, sponsors, state, and federal sources. In some cases the cumulative effect of multiple policies is effectively unknowable by a researcher, and hence, policies that are codified in computer manageable languages may be needed. He pointed out that many organizations are concerned about HIPAA's Safe Harbor definition of de-identified data, but that more quantitative de-identification methods that preserve the scientific utility of data can be developed. He noted that the availability of these emerging methods and their variable acceptance by IRB's (particularly those with limited understanding of de-identification techniques) is problematic and argued for more centralized/shared expert review of research involving novel de-identification methods. Dr. Malin also argued for risk-based privacy policies where risks of identification can be quantified, and the need for training of IRBs in information risk. The policy implications of cloud computing architectures, where the physical location of the data is unknown, and the unclear liability of cloud computing providers for unintended disclosure, was discussed. Encryption of sensitive data can help mitigate many re-identification risks in such settings, and emerging computational methods that permit analysis and mining of encrypted data (e.g., homomorphic cryptosystems) are technically feasible but have unknown acceptance by users, who cannot directly see the data they are analyzing.

Workshop Findings

The provocative future scenarios envisioned by the session leaders spurred a vigorous and wide-ranging discussion of issues among the participants. Many of the issues were not specific to a single type of research involving clinical data, and so the Workshop Findings section of this report is organized by cross-cutting issues rather than by topical sequence of the workshop agenda, with the most general findings listed first.

At the highest level, workshop participants uniformly expressed the following:

- Clinical data, taken broadly as the health states experienced by individuals and populations and observable by them and others, particularly healthcare providers and provider organizations, is essential to understanding human health and disease. However, the focus of clinical data as a byproduct of a clinical encounter or a hospital admission is shifting to longitudinal recording of health events in a lifelong continuum, only a part of which constitutes traditional clinical data, i.e., signs and symptoms, diagnostic findings, and records of treatment recorded by providers and provider organizations.

- Clinical data as big data is growing in both volume and variety as information and communications technologies become ubiquitous. Major sea changes are evident in the adoption of Electronic Health Record (EHR) systems by providers spurred by federal incentives, the availability of mHealth apps on smart phones and other personal electronic devices, the emergence of low-cost, network-enabled physiologic monitoring technologies, the availability of low cost high throughput laboratory technologies such as whole genome sequencing, and 21st century consumer empowerment (e.g., patient-reported outcomes).

- Clinical data differs from other forms of biomedical research data in a number of ways. Most importantly; clinical data is patient-level data with variable levels of sensitivity and a strong need for protection of confidentiality. Additionally, most clinical data, including EHR data, is not collected initially for research purposes, but for patient care and must be repurposed for research. In this regard, the overlapping and sometimes conflicting policies of research organizations, sponsors, states, and the federal government create a formidable set of barriers to the effective conduct of research involving clinical data.

- Access to clinical data by researchers is currently far less than optimal for the progress of science and is by far the most difficult and pervasive barrier to research productivity for these types of data. Some access barriers are technical (e.g., proprietary data systems that cannot be queried in a usable or cost effective manner for research), but most are related to culture and policy. In this regard, the complexity and cost of acquiring consent for use of clinical data are often rate-limiting and cost-prohibitive impediments to research, especially interventional research.

- Clinical data acquired via traditional clinical trials or observational studies is generally of high quality but is prohibitively expensive and takes too long to acquire to be the sole means for answering relevant health-related questions.

- Clinical data acquired as a byproduct of care delivery is of variable quality and consistency, and suffers from gratuitous heterogeneity in the naming and coding of clinical content. Improving the quality of clinical data at the time of initial capture will not only improve its utility for research purposes, but will simultaneously enable improvements in patient-specific care and population health. Improvements in the quality and consistency of clinical care data are expected to result from federal incentives for Meaningful Use included within programs of the Office of the National Coordinator (ONC) for Health Information Technology and the Centers for Medicare and Medicaid Services (CMS), but much work remains to be done. ONC invites researchers to engage in the process of defining future stages of Meaningful Use to facilitate research, as well as improve the quality of care.

- Patient registries provide an intermediate cost step between opportunistically acquired clinical data from routine operations and formal clinical trials. An increasing number of professional organizations and government agencies are developing registries as a means to document quality of care and provide a mechanism for continuous quality improvement. However, the quality of the data within registries can vary, and integration across registries can be difficult because of issues related to the use of standards, common data elements and definitions. Additionally, permissions may not be in place for the use of registry data for research.

- Routinely acquired clinical data can support large, low cost observational studies, but its utility is limited by issues of data completeness and quality, selection, lead time and other biases that may exist in EHRs, and unmeasured confounders. Many of these issues can be addressed through randomization, completeness of follow up, and larger sample size, but a better understanding is needed of issues such as biased assignment and confounding by indication, where larger sample size will not help. While findings from observational studies may, in some cases, be persistently incorrect, and resolvable only by prospective randomized studies, many types of scientific inquiry do not need to await improvements in the quality and consistency of clinical data acquisition. For example, when the effect is very large, or the question is not an effort to infer causal effect of an intervention, strong patterns and signals detectable in the inherently non-standardized and noisy clinical data that is already available may be sufficient. Better interpretation of clinical data and its biases requires an understanding of patient preferences, as well as provider and practice biases (e.g., training, past experience, cost). All of these factors influence the diagnostic and therapeutic options that are subsequently chosen and recorded. Patient preferences are an element of phenotype that are generally missing from clinical data, and clinical systems are not designed to capture them

- There is a need to identify successful strategies for engaging patients to better understand their views on sharing data, returning results, etc. and encourage their active participation in the research process

- Standards for coding of clinical data, such as ICD9 and ICD10 are essentially undefined, as they simply assign a numerical code to a text term without providing an associated definition of that term. Although coding systems are improving (e.g., ICD11 is better in this regard), more robust approaches are being developed for extracting clinical phenotypes from EHR data. The experience of the NIH-supported eMERGE (electronic Medical Records and Genomics) consortium, the FDA-funded Sentinel project and the NIH Health Care Systems Research Collaboratory is that a combination of structured and unstructured elements available within the clinical data (codes, laboratory values, medications, and concepts identified in narrative text by natural language processing) is optimal for both research and clinical care quality improvement

- Current centralized models of curation of clinical data combined with molecular variation data for research can be expensive and difficult to scale as the number of health-related data sources increases.

- The generation of knowledge will accelerate as clinical and basic research data become more available. However, the effective translation of new knowledge to improved care is inadequate and could remain a significant impediment to the timely implementation of best practices identified by NIH-supported research. The gap between what we know and what

we do grows daily. Research and development to create systems infrastructures that are not dependent upon human beings reading and remembering the published literature is a national, indeed, global need that is unmet. Improving the way that EHRs capture and record clinician decision logic would greatly advance EHR-based research and provide an even greater advance for improving care. This makes implementation science an emerging and highly leveraged area of research.

Recommendations

In light of the workshop findings as listed above, participants offered the following actionable recommendations for NIH with respect to clinical big data, organized into five conceptual areas. Recommendations should be implemented with knowledge of and in conjunction with ongoing efforts of other organizations, such as those currently underway by PCORI and ONC.

1. Improve access by researchers to routinely-acquired clinical data

Workshop participants were unanimous in their assessment that clinical data, as a research resource, is a *sine qua non* of progress in understanding human health and disease, and that the volume and variety of it will escalate dramatically in coming decades. But even now there are large volumes of clinical data in electronic form that are fundamentally inaccessible for purposes of research due primarily to policy constraints and the practice of healthcare organizations to treat their data as proprietary, and not to technical impediments. Workshop discussions highlighted the irony that the inability to efficiently and effectively study human health responses to diagnostic and therapeutic interventions made in the course of routine care, condemns healthcare to the status quo of a "trillion dollar cottage industry" that is incapable of evidence-based improvement at scale. Although the principles of standardization and large-scale learning are gaining some traction in the area of improving safety, these principles need to be expanded and applied to improving the clinical decisions that are made intentionally. Since policy-based impediments to research uses of clinical data are the greatest problem, workshop participants focused many of their recommendations for progress in this area. The NIH has a particular role to play in using research to enable policies based on evidence rather than opinion, and to evaluate the intended and unintended consequences of policies using rigorous research methods.

a. Address issues related to consent.

Consent by individuals for prospective use of their person-identifiable clinical data is a fundamental tenet of the ethics of human subjects' research in cases where there is more than minimal risk involved in the use of the data. In this context, participants recommended that NIH invest in empirical research to provide evidence-based approaches for informing the public about learning activities conducted in the course of biomedical research or quality improvement and engage them in an active process of consent and participation in these activities. This research would involve the public, health care systems, researchers, regulators and other federal agencies. Ultimately, this research can inform improvements in federal regulations and agency policies, including simplification and standardization of consent procedures and documents.

b. Streamline mechanisms to grant access to clinical data for research

In this area, workshop participants recommended both policy and technology innovations:

i. Support alternative (centralized/shared) models of IRB review that avoid duplicative effort and optimize use of specialized review expertise e.g., quantitative de-identification and risk assessment, for projects that are not well served by the highly decentralized IRB review model currently in place.

The current model for IRB review, requiring each of thousands of IRBs to have or acquire by consultation technical expertise that is scarce, results in a predictable tendency towards over-protectionism for research that involves data transformations such as de-identification and encryption that are technically powerful but understood in depth only by experts. A referral and/or delegation for specialized review conducted centrally would maximize scarce human capital in rapidly advancing areas of science, including information science relevant to clinical data.

Additionally, clearer guidance and dissemination of best practices is needed for uniform quality improvement across local IRBs. The oversight of research must occur at the local level and extension of pioneering work by the NCI and the CTSA's (IRBshare) should be expanded with empirical evaluation to improve human research protections systems while reducing ineffective bureaucracy.

ii. Support research to define alternative models (beyond HIPAA) for risk-based assessment of research data and its uses.

Statutory requirements such as HIPAA currently dominate the policy framework for research availability of clinical data. But HIPAA's simple definition of three classes of clinically-derived data (PHI, limited datasets and de-identified) contrast with the reality of a continuum of risk and benefit that derives from the nature of the data, the scenarios of research usage, and the types of threats that are envisioned as risks to confidentiality and data integrity. This is an area that is worthy of NIH support for both technology and policy research.

iii. Support the development and evaluation of strategies to enhance public trust in biomedical and clinical research, and strategies to enhance public education on how research using clinical data is done, and its public benefits. In addition, support the development of technologies to track participant consent, authenticate and authorize researcher access, and codify local, state and federal policies affecting data access so that they are computer manageable.

iv. Bolster the ethical framework for quality improvement (QI) studies to resolve tension across the spectrum of learning activities and to improve the knowledge of individuals about the use of their data by health systems and researchers. Models of consent to participation in Learning Health System should be evaluated, including notification using electronic means to improve participation and reduce incentives for work-arounds of human subjects research rules. Promote transparency in learning, whether conducted for QI or research purposes, with the aim of generalizability when appropriate.

Methods used by healthcare organizations for internal QI are often identical to those used by health services and comparative effectiveness researchers. The exemption from IRB review of activities using person-identifiable clinical data for QI has led to a widely known loophole in human subjects protections, where investigators will declare a project to be quality improvement, and then apply post hoc to an IRB to get permission to use existing data for publication. In these settings, where the intent at the start is to contribute, if possible, to generalizable knowledge by publication and/or presentation, there should be uniform rules for review that do not incentivize avoidance of ethical review. To do otherwise threatens public confidence in the clinical research enterprise. It was requested that we work with OHRP to address current distinctions between QI and research so that contributing to the pool of generalizable knowledge, the use of empirical information to inform care, and transparency, become de rigueur.

v. Develop an infrastructure for NIH trusted researchers and trusted data providers for access to clinical data.

One model envisioned, as described originally by Barbara Wold of the National Cancer Institute, is of a "researcher's library card" for access to research data by qualified researchers. Like driver's licenses and other forms of documentation of privileges, the researcher's library card would be a set of credentials based on education, training, certification of organizational affiliation, and agreement with an enforceable set of acceptable use policies governing the ethical use of clinical research data, that would be time limited but renewable at an appropriate interval. Elements of this model already exist in the time-limited certification of human subjects protections training required at most research institutions.

A researcher could then use their researcher's library card to get access to sets of clinical data resources that match their privileges, which might range from currently defined categories of sensitivity such as HIPAA Limited Datasets, as well as emerging classes of data with nonzero re-identification potential (e.g., genomic sequence data with or without associated de-identified clinical data).

Workshop participants recommended that NIH convene groups to identify the needed elements of this infrastructure, and survey current best practices in industry for similar approaches to credentialing-based rather than continuing with a project-at-a-time based access to data.

c. Study policy gaps that cause impediments to access, and identify areas where new policy or incentives are needed.

Knowledge of compliance with and outcomes of NIH's current research data sharing policies is anecdotal. A more formal and systematic approach to understanding how current policy affects the nation's scientific enterprise, and where there are gaps and opportunities to modify existing policies or create new ones that will have a favorable effect on improving knowledge of human health and disease for the public good would be a timely and valuable effort by NIH. This should include support for analysis of incentives and disincentives for healthcare organizations to share clinical data for research purposes.

2. Increase the quality of clinical data available for research

Workshop participants agreed that the quality and consistency of clinical data acquisition needs substantial improvement and that such improvements will immediately benefit both the care provided to individuals and the quality of research done using the data generated in the course of that care.

a. Foster creation of shared public library of phenotype elements and algorithms related to clinical data.

Finding a cohort of individuals whose clinical data meets a specification of a disease or physiological condition is fundamental to many forms of NIH-sponsored research. Current methods of cohort identification using clinical data often rely upon simple coded elements, such as diagnostic billing codes, that have high error rates and biases. More robust methods of cohort selection using combinations of multiple types of data (e.g., codes plus laboratory results, medications, and concepts derived from natural language processing) will enhance research productivity and reproducibility for projects using clinically-derived data from EHRs. NIH-supported resources, such as the Phenotype KnowledgeBase (PheKB.org) and the computable phenotypes being developed under the Health Care Systems Research Collaboratory, are important early starts that need to be scaled up from current libraries of dozens to potentially thousands of standardized definitions.

b. Partner with organizations that develop and promote the use of standardized clinical quality measures (e.g., NCQA, HEDIS measures) to enhance their utility for research.

Process and outcome measures related to clinical data that are developed for quality assessment report cards and quality improvement can have positive effects on research as well. NIH should take advantage of opportunities to partner with other organizations that develop standardized quality measures so that use of the measures will be appropriate for both quality improvement and research, and organizations will have multiple incentives to use them.

c. Foster the development and use of standards for including new classes of data, as well as existing classes of data, in clinical systems,

NIH should use its convening authority and resources to help develop consensus standards for emerging classes of clinical data that will have important benefits to biomedical research. For example standards are needed to include genomic sequence variation data in clinical systems in a manner that supports both human readable interpretation and clinical decision support logic. The predictable entropy of ad hoc data formats for molecular variation (‘omics) data can be reduced substantially by NIH taking the initiative to support standards development in this area now. Standards are also needed to capture other new classes of data, including, e.g., patient preferences and patient-generated data, microbiome, and environmental data.

d. Support research and development to improve the validity of clinical data. Research is needed to better understand clinician bias in diagnosis, treatment decisions, and the way such information is entered in an EHR. Technology development is needed to assist in validating data at the time it is entered, with positive benefits for clinical care and research. New procedures and systems are needed so that when inconsistencies in clinical data are discovered in the course of research, corrected information can be included in research systems and fed back into clinical data systems to improve data quality.

e. Strengthen NIH grant and review procedures to improve the quality of research data resulting from NIH sponsored research.

Workshop participants observed that the problem of variable quality and nonstandard naming and coding afflicts not only routinely acquired clinical data, but also the data from NIH-supported research. NIH has the opportunity to improve this situation at a national scale by encouraging researchers requesting NIH funding to address data management and representation in their grant proposals and by implementing changes to the grant review process to evaluate the same. Recommendations in this area include:

i. Provide guidance to investigators submitting research proposals on the importance of using standardized data elements and measures, along with educational resources that help them to know about and effectively use national and international standards for their research variables and variable values, wherever feasible.

ii. Enhance the grant review process to promote high quality data outputs. This can be done by adding relevant informatics expertise to standing and *ad hoc* review groups, as well as the use of specialized consultants to review submitted data sharing and management plans for use of standardized clinical data elements, data privacy and data security best practices. NIH should consider explicitly scoring grants on their adoption of standardized measures and data elements, the use of existing data resources, and the strengths and weaknesses of the data sharing plan. In order to accomplish this, NIH will need to expand current guidance on what constitutes a high quality data sharing plan. These improvements in the quality of submitted research plans and the effectiveness of the review of those plans is consistent with current OSTP guidance on achieving maximum return on the public investment in research.

3. Increase the quantity of clinical data available for research

Workshop participants observed that the tsunami of potentially valuable clinical data that will result from large-scale adoption of computerized clinical systems with increasing interoperability is at an early stage. In addition to improving quality, workshop participants recognized opportunities for NIH to improve the quantity of clinical data available now and in the future. Recommendations in this area include:

a. Support the public availability of key clinical data resources

A public investment on behalf of researchers in a relatively small number of data resources could have a highly leveraged effect. An example of this is the licensing of SNOMED CT terminology on behalf of researchers by the National Library of Medicine, for inclusion in the Metathesaurus of the Unified Medical Language System <http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/>. Workshop participants recommended minimizing adoption of proprietary resources, but recognized there may be special opportunities to benefit large numbers of researchers by such site licensing on behalf of the entire biomedical research community. Candidate resources include national scale claims data, and licensable terminology databases. NIH should establish an ongoing process, led by NLM, to evaluate evolving needs of researchers for such data resources that would be catalytic for research productivity.

b. Fund infrastructure that contributes to use and interoperability of big data in healthcare.

Behind-the-scenes activities such as data standards and terminology development are components of an essential infrastructure for creating research data that is interoperable and re-usable. Ongoing efforts to create and validate terminologies, and common data elements which change and grow as the knowledge base of science changes and grows, are particularly important in the area of research use of clinical data. A variety of support mechanisms (contracts, cooperative agreements and investigator initiated grants) can be employed to provide adequate resources for this critical infrastructure activity. A special opportunity exists for NIH to better align its support for and coordination of research data element definitions (e.g., those contained in the NIH Common Data Element resource <http://cde.nih.gov>, PhenX <https://www.phenx.org/>, and PheKB <http://phekb.org/>) with other clinical and research data standards being implemented by developing healthcare research consortia.

4. Spur innovation in analytical methods and tools for research involving clinical data

a. Support investigator-initiated research to characterize clinical data with respect to biases, strengths and weaknesses for various study types, and to identify study designs that are inherently more effective and efficient for the various study types.

A number of novel trial designs incorporating the strengths of big clinical data were presented and discussed at the workshop. Also noted were as yet not well-defined issues that confound the use of routinely acquired clinical data for research, such as confounding by indication and other forms of bias, and the potential lack of external validity. NIH support for research that leads to better understanding of the strengths and weaknesses of clinical data, and study designs that can compensate for bias, will catalyze improvements in the quality and consistency of research.

b. Support research into the validity of predictive models for individualized care.

Statistically based predictive models have been effectively incorporated into many industries, and predictive models based on clinical data have been developed for

many disease states. Uptake and use of these for patient-specific and population-relevant clinical decisions has been sparse and uneven in healthcare, where the highest valued mechanism of decision-making is often the intuition of a clinician. Understanding the implementation science aspects of predictive models, as well as their strengths and limitations, is a worthy area for additional NIH support.

c. Support development of shared resources of data management and analytical tools, and clearinghouses of best practices.

Knowing which software to use for a specific big data research problem is an important and ongoing challenge. NIH support for investigator-initiated computational tools for both the pre-processing (e.g., data normalization of data from multiple heterogeneous sources) and specialized forms of analysis that are needed for clinically-derived data, will be an ongoing need, if we hope to maximize the return on public investment in research using clinical data. In a manner analogous to the creation of a Data Discovery Index of research datasets, there needs to be a clearinghouse for researchers to find software tools, and learn how to use the tools correctly and for the appropriate research use cases. Of note, relevant efforts are under way in NIH and PCORI, and the concept of a software catalog is being explored as part of the BD2K program. NIH should take stock of such efforts and determine what additional infrastructure is needed.

d. Convene public conversations about conducting science in new ways.

Many of the issues that came up in discussion by workshop participants pose significant challenges to the way academic institutions currently do research: tenure decisions optimize novel, independent discoveries, where increasingly a team science approach is optimal; determination of research questions is considered the exclusive purview of the scientific community, but there is growing interest in involving patients in setting research priorities; research remains largely disease-focused vs. concentrating on health-promoting phenomena; research data are typically held from public view, but technology allows the possibility of documenting researchers' databases queries for purposes of audit and scientific integrity. These pressures call for a shift in incentive structure, benefits and rewards, and increased transparency in the conduct of research.

These trends have many educational implications. Clinician scientists will need to be better equipped to understand how to curate, use and analyze big data. Training should be expanded to improve clinicians' ability to gather data that can contribute to knowledge building as well for use indecision support. Enhanced education and training of the clinical and clinical research workforces were acknowledged to be the focus of a different BD2K workshop, but participants in this workshop reiterated their importance in improving the data competency of the biomedical research workforce. Beyond the clinical and research workforce, public acceptance of research involving clinical data will also require citizens to become more sophisticated about the use of clinical data in learning. At a national level, if citizens are to be involved in science, the K-12 curriculum must ensure a scientifically-literate citizenship. Efforts to improve education will be a powerful adjunct to all of the technical and policy recommendations noted in this workshop report.

In this regard, participants recommended that NIH use advisory boards to develop guidelines for direct-to-citizen and direct-to-participant communications and public education. For example, this could provide a forum for understanding the desire and means for providing consumers with information regarding how their data was used and contributed to learning, as well as discussions concerning the return of results and incidental findings. It will be important to learn from relevant PCORI initiatives currently addressing participant engagement.

e. Support novel funding mechanisms (e.g., prize competitions, bake-offs) for novel approaches to improving access and effective research use of clinical data.

Traditional investigator-initiated research funded by R01 and similar grant mechanisms remains the foundation of the NIH-supported research portfolio. Workshop participants noted the utility of novel approaches to engaging the creative energies of a growing, technically sophisticated segment of software tool and resource developers. mHealth apps for smartphones that improve the quality and timeliness of clinical research data capture is a contemporary example of opportunities where such competitive approaches could be applied, and where public recognition of success complements and may outweigh financial incentives.

5. Facilitate effective uptake and use of resulting research findings to improve health and health care

a. Support efforts to represent new clinical knowledge in computable form, amenable to decision support systems use. Support research and development to create approaches to effectively use those computable forms in operational settings.

Workshop participants observed that genomics is the poster child for escalating complexity in healthcare, but such complexity is extending to other areas as well, including behavioral, social, and environmental factors. Twentieth century models of decision making based on professionals reading and remembering the published literature is fundamentally inadequate for healthcare based on current evidence. The disarmingly simple goal of doing the right thing, and only the right thing, and doing it every time for every patient, demands a systems infrastructure that exists in other high-risk industries but is notably absent in healthcare currently.

Since clinical research frequently results in findings that improve clinical operations, research and development are needed to create knowledge representation and systems infrastructures for decision support for providers, patients and their families. NIH can expect an immediate and long-term benefit from such infrastructure, which will provide a path for implementing best practices based on the results of NIH-sponsored research, combined with continuous quality improvement and other elements of the Learning Healthcare System. Stated otherwise, the Learning Healthcare System will learn new knowledge faster if it is in computer-interpretable as well has human-interpretable formats.

b. Foster implementation science.

Advances in knowledge representation will create an expanded body of best practices guidance. Knowing how to insert that guidance effectively into care is a

principle focus of Implementation Science, which addresses the growing gap between what we know and what we do in the 21st century. NIH should foster implementation science research, particularly in areas where big data needs to inform care delivery (e.g., effective incorporation and use of 'omics data for individualized care). For some classes of NIH-sponsored research, it may be feasible and appropriate to require an explicit implementation plan, as other federal agencies (e.g., VA) are doing. Partnerships with other federal organizations, particularly AHRQ and CMS Innovation Center, regarding dissemination and implementation plans, can leverage NIH research investments in this area.

Summary

Clinical data as a big data research resource is growing in both volume and variety, and is an essential resource for many types of NIH-supported research. The recommendations offered by workshop participants on the basis of an intensive and rich discussion offer actionable steps and a research agenda that NIH can take to improve the quality of clinical data used for care and research, access to that data by researchers, availability of tools and resources to facilitate its analysis, and an infrastructure for using the new knowledge generated by that research to improve health and health care.

Appendix A

NIH BD2K Workshop on Enabling Research Use of Clinical Data
September 11 – 12, 2013

Agenda

Workshop Objectives: Identify actionable steps that NIH can take (alone and with others) to enable greater use of clinical data from electronic health records, patient-reported outcomes, and other clinical sources in biomedical research. By examining research use cases in pragmatic clinical trials, observational studies, and genome-phenome studies, the workshop will identify needs for: 1) research and development of new technologies and methods; 2) common infrastructure to enable future research scenarios; and 3) policy changes necessary to facilitate progress.

Day 1 – Wednesday, September 11th

9:00 – 9:20	Welcome and Introductions ▪ **Leslie Derr, Ph.D.**, NIH, Office of the Director ▪ **Jerry Sheehan**, National Library of Medicine ▪ **Rob Califf, M.D.**, Duke University (*Workshop Co-chair*) ▪ **Dan Masys, M.D.** University of Washington (*Workshop Co-chair*)
9:20 – 10:00	Keynote Presentation **Eric Green, M.D., Ph.D.**, Director, National Human Genome Research Institute *and* Acting Associate Director for Data Science, NIH
10:00 – 10:15	**BREAK**
10:15 – 10:30	Setting the Stage and Methodology for topic sessions ▪ **Dan Masys, M.D.** ▪ **Rob Califf, M.D.**
10:30 – 12:00	**Research Use Case 1:** Pragmatic Trials & Interventional Studies ▪ *Presenter:* **Mike Lauer, M.D.**, National Heart, Lung, and Blood Institute ▪ *Roundtable discussion*
12:00 – 1:00	**LUNCH**
1:00 – 2:30	**Research Use Case 2:** Genome-Phenome Correlation ▪ *Presenter:* **Zak Kohane, M.D., Ph.D.**, Harvard Medical School ▪ *Roundtable discussion*
2:30 – 3:00	**BREAK**
3:00 – 4:30	**Research Use Case 3:** Observational Studies ▪ *Presenter:* **Greg Simon, M.D.**, Group Health Cooperative ▪ *Roundtable discussion*

4:30 – 4:45 pm First day summary and wrap-up

Day 2 – Thursday, September 12th

8:30 – 10:00 **Cross-Cutting Needs**: Infrastructure, Standards, and Policy
- *Presenter:* **Brad Malin, Ph.D.**, Vanderbilt University
- *Roundtable discussion*

10:00 – 10:30 **BREAK**

10:30 – 12:00 Group review and synthesis of recommendations
- *Rapporteur:* **Dan Masys, M.D.**
- *Moderator:* **Rob Califf, M.D.**

12:00 Adjourn

Workshop Planning Committee

Leslie Derr (NIH-OD) and Jerry Sheehan (NLM), *co-leaders*; Denise Bonds (NHLBI), Dana Casciotti (NLM), Jim Cimino (CC), Elaine Collier (NCATS), Valery Gordon (NIH-OD), Lyn Hardy (NINR), Lucia Hindorff (NHGRI), Lisa Lang (NLM), Catherine Myers (NCCAM), Nancy Miller (NIH-OD), Rick Moser (NCI), Dina Paltoo (NIH-OD), David Patton (NCI), Laura Rodriguez (NHGRI), Robert Star (NIDDK), Barbara Wells (NHLBI).

Webcast: http://videocast.nih.gov
Discussion forum: http://clinicaldata.prophpbb.com/forum3.html

Appendix B

NIH BD2K Workshop on Enabling Research Use of Clinical Data

Invited Participants

Patricia Flatley Brennan, RN, PhD
University of Wisconsin-Madison

Ralph Brindis, MD, MPH
University of California, San Francisco

William Chin, MD
PhRMA

Chris Chute, MD, DrPH
Mayo Clinic-

Gregory Downing, DO, PhD
HHS, Office of the Secretary

Sharam Ebadolahi, PhD
IBM Research

Lynn Etheredge
George Washington University

Rachael Fleurence, PhD
Patient Centered Outcomes Research Institute

Doug Fridsma, MD, PhD
HHS Office of the National Coordinator for Health IT

J. Michael Gaziano, MD, MPH
Veterans Administration

Alan Go, MD
Kaiser Permanente Northern California,

Eric Green, MD, PhD
National Human Genome Research Institute

William (Ed) Hammond, PhD
Duke University

Mark Hoffman, PhD
University of Missouri-Kansas City

George Hripcsak, MD, MS
Columbia University

Anil Jain, MD
Explorys, Inc.

Taha A. Kass-Hout, MD, MS
Food and Drug Administration

Isaac Kohane, MD, PhD
Harvard Medical School

Rick Kuntz, MS, MSc
Medtronic, Inc.

Michael Lauer, MD
National Heart, Lung, and Blood Institute

Bradley Malin, PhD
Vanderbilt University

Clement McDonald, MD
National Library of Medicine

Deven McGraw, JD, MPH, LLM
Center for Democracy and Technology

Jerry Menikoff, MD, JD
HHS Office for Human Research Protections

Jim Nemecek
McKesson Corp.

Lucila Ohno-Machado, MD, PhD
University of California, San Diego

Véronique Roger, MD, MPH
Mayo Clinic

Andrew Shatto
Centers for Medicare & Medicaid Services

Gregory Simon, MD, MPH
Group Health Cooperative

Brennan Spiegel, MD, MSHS
UCLA School of Medicine.

Anne Trontell, MD, MPH
Agency for Healthcare Research & Quality

Marc S. Williams, MD
Geisinger Health System