INFORMATICS PROFESSIONALS. LEADING THE WAY.

5 December 2014

Cindy P. Lawler, Ph.D.
The National Institute of Environmental Health Sciences (NIEHS)
BD2K_CBS_RFI@niehs.nih.gov

Request for Information on BD2K - NOT-ES-15-002 – "Making Data Usable – A Framework for Community-Based Data and Metadata Standards Efforts for NIH-Relevant Research"

Dear Dr. Lawler:

On behalf of the American Medical Informatics Association (AMIA), I am pleased to submit these comments in response to the above-referenced request for information. AMIA is the professional home for over 5,000 members who are focused on biomedical and health informatics and who work throughout the health system in a broad spectrum of clinical care, research, academic, government, and commercial organizations. AMIA is dedicated to the development and application of informatics in support of patient care, public health, population health, consumer health, professional development, research, and administration. We also seek to advance the development of effective policies and regulations to support our members and mission. AMIA ultimately works to enhance human health and health care delivery through the transformative use of information and communications technology.

Please note that the following comments were developed by a small group of AMIA members with expertise and insights related to the NIH's Big Data to Knowledge (BD2K) initiative. The comments do not necessarily reflect official AMIA positions or policies, though we do believe that the comments provided will be helpful to the NIH as they facilitate the development of data-related standards in support of the biomedical research community.

The main topics covered by members of our Clinical Research Informatics Working Group in these comments include:

- Approaches to developing a comprehensive and extensible framework for data and metadata standards
- Approaches to data and metadata standards related to EHR data and biomedical research
- Opportunities for community-driven data and metadata standards
- Recommended features and framework for developing standards
- Approaches to developing an extensible and inclusive biomedical "Big Data" framework

AMIA welcomes the opportunity to participate further in your efforts to advance standards for data sharing and data management to support biomedical research. We can also serve as a convener for continuing dialogue on the subject through our working groups and conferences. Please let us know how we can be of further assistance.

Ross D. Martin, MD, MHA
Vice President, Policy and Development
ross@amia.org

Response to NOT-ES-15-002 ("Making Data Usable – A Framework for Community-Based Data and Metadata Standards Efforts for NIH-Relevant Research") from members of the American Medical Informatics Association (AMIA) Clinical Research Informatics Working Group*

Philip R.O. Payne, PhD, FACMI[1]; Rachel L. Richesson PhD, MPH, FACMI[2]

[1]The Ohio State University, Department of Biomedical Informatics

[2]Duke University, School of Nursing

Executive Summary

Ensuring that the biomedical research community can ask and answer meaningful questions in the context of emerging big data resources will require investments in and support of wide-ranging efforts to create and disseminate the theories and methods that are foundational to an extensible and inclusive biomedical big data framework and associated methods/standards. This framework will require at a minimum:

1) Creating tailored and targeted workforce and knowledge development programs;
2) Empowering knowledge workers to be integral parts of the data discovery, integration, and analysis "pipeline"; and
3) Establishing and sustaining community-wide data analytics "commons" and data standards-setting initiatives.

Further, we believe it is imperative for efforts that hope to address the preceding issues to employ and leverage trusted and impartial entities with community-wide convening authority, such as the American Medical Informatics Association (AMIA), so as to ensure that resulting frameworks are not only extensible, but also inclusive of the full spectrum of stakeholder needs and expertise.

1. Introduction

The healthcare and life science communities have seen a dramatic increase in the collective focus on the role of big data and its potential impact on research, healthcare delivery, and population health. Despite such vigorous dialogue, there remains much confusion as to what precisely big data is and how it will

*These comments in response to the referenced RFI were developed by AMIA members at our request and represent the perspectives of the authors listed. They are provided for informational purposes but may not necessarily reflect official AMIA policies or positions.

generate demonstrable value in these domains. As an example, in a recent editorial, Dr. Phil Bourne (NIH Associate Director for Data Science) notes, "Interestingly, no two stake-holders would be likely to define Big Data in exactly the same way" (1). In that same commentary, he goes on to describe a variety of definitional aspects of what may constitute big data, such as size, the speed with which data is generated, its heterogeneity, or more broadly, the fact that conventional data analysis tools cannot generate useful insights in the context of such big data (1). These definitions are emblematic of many others that have been posed, all of which revolve around a sense that big data is both complex and that you will likely "know it when you see it" (2-6). At the same time, a variety of reports have made strong arguments concerning the ways in which big data will generate new and actionable insights that will advance the state of human health. For example, claims have been made that the appropriate use of big data will lead to improvements in: 1) the conduct of collaborative scientific endeavors leading to advances in foundational basic and clinical science knowledge (2-4, 7-9); 2) our ability to identify and apply bio-markers in order to enable personalized medicine (5, 10); and 3) the delivery of contextual information at the point of care in the form of highly targeted clinical decision support (3, 6, 11).

When one reads such reports and commentaries, it becomes clear that big data represents an emergent and potentially highly valuable complement to current approaches to computational, basic science and clinical research, as well as healthcare delivery. ***However, a review of the current state of the art in biomedical big data science surfaces a number of important and unanswered questions that are pertinent to the RFI which this document is being submitted in response to, namely:***

1) How can we arrive at a consensus as to what data types or sources (e.g., genomic, physiologic monitoring, clinical data, etc.) qualify as big data and how can biomedical researchers access and use these data in a reproducible and systematic manner?
2) How can we encourage the broad and consistent adoption of highly usable data, metadata, and methodological standards capable of ensuring the reproducibility and transportability of health-relevant data analytics that employ big data resources?
3) Can we quantify the quality, uptake, and reproducibility of those standards and their impact on biomedical research?
4) Once such big data and corresponding methods and standards are identified and applied, are there ways to ensure we are adequately leveraging them so as to ask and answer important questions that will generate value and that would not otherwise be feasible?

5) What existing communities can be engaged to address the questions above and advance the use of big data to support advances in biomedical healthcare and improvements in healthcare delivery and outcomes?

When taken as a whole, these outstanding questions are related to the fundamental ways in which we find, manage, and analyze big data such that results have value and impact for NIH-relevant research, and that the methods can be reproduced and extended by others in various stakeholder communities. We define NIH-relevant research to include all studies that broadly fall under the spectrum of basic and clinical research as defined by the NIH, including investigation of the mechanisms of human disease, therapeutic interventions, clinical trials, development of new technologies, epidemiology, behavioral studies, and outcomes and health services research (12). The broad scale of such research, as well as the widespread adoption of EHRs, and emerging data capture or generation technologies, have led to the aforementioned increasing amounts of data; the scientific community needs to identify strategies to share it in meaningful ways. Further, the NIH policy on the sharing of research data (13) has raised questions about how data should be represented for effective data sharing, making the need for data standards critical and immediate. The use of other data types – such as EHR derived and/or physiological monitoring data that are generated for purposes other than research – are particularly challenging area to standardize. These standards are important for NIH-relevant research using these data sources, but the standardization will require cooperation and incentives for a variety of stakeholders and communities outside of the NIH.

1) Developing a Comprehensive and Extensible Framework for Data and Metadata Standards – Key Components and Possible Approaches

Background, Dimensions, and Current Challenges for Data and Metadata Standards

The term 'standards' is widely used and garners almost universal support but is rarely defined. For purposes of this RFI, we define data standards as being consensual specifications for the representation of data from different sources or settings (14). Further, we broadly characterize such 'standards' to include 3 different types: 1) data (representation) standards, which includes both the names of variables or fields and their associated answers or value sets, which can include entire coding systems; 2) metadata standards that are the 'data about data,' and describe data element definitions, data type, units, provenance, and other data about the element itself; and 3) methodological standards that guide how data are accessed, safeguarded, managed, manipulated, and interpreted to support biomedical

discovery. Each of these types of standards encompass a wide range of methods and examples and is buttressed by specialists and tools from multiple domains including informatics, health professions, research, computer science, library and information science, statistics, and others including the newly emergent field of data science. Below we provide a brief description of each; examples to illustrate the breadth and scope of each standard type; a sample of what we perceive as the current challenges that need attention and leadership; and potential resources that could address those challenges.

Data representation standards can include a range of formalisms, from standard languages (e.g., XML, RDF) to robust ontologies (e.g., Gene Ontology) and standardized coding systems (e.g., ICD, CPT, SNOMED CT, LOINC). Because so many options exist, there is still uncertainty and debate as to how to use which standards and in what contexts. And even when standard coding systems are used in clinical information systems, the issues of whether the codes are used consistently (as shown in inter-rater reliability studies) and if they have validity for the condition or state that researchers assume are an unremitting concern. In addition to the need for standards and standardized approaches for clinical data as we have noted, the standardization of *de novo* research data is important to NIH-relevant research and presents unique challenges. For example, the management of standards for questions on case report forms is a challenge because of the vast number of items needed and the protocol-specific nature of research. Solutions reported for this challenge include the idea of standardizing items rather than forms (15). A number of Common Data Element (CDE) projects have now been developed, including a number of NIH-sponsored projects such as those by the NINDS and the NCATS/ORDR (16-18). However, a great challenge – if not the primary challenge – for re-using CDE items is finding them. The NCI CDE Browser, which has been around for more than a decade, is a valuable resource for research (19), although a commonly reported limitation is the presence of multiple data elements for the same construct, making searching and selection of "standards" problematic(20). The NIH CDE browser (21) is a notable and commendable step for assimilating different collections of CDEs across NIH and a great resource for researchers to attempt to reuse existing items – moving toward standardization by the reducing the number of item variations used. Now that the NIH CDEs are assimilated in one place, controlled terminologies and other efforts can be better directed and duplicate or similar elements can be deleted or harmonized.

Metadata standards can represent contextual information about the data – such as the definition of the data element, the author, and implementation details –a and are usually described in the form of data attributes or qualifiers. In EHR systems or clinical data warehouses, the metadata give important context

such as description of a medication as prescribed, or the type of diagnosis. This context is often critical for the sensible aggregation and interpretation of data from heterogeneous EHR systems or clinical data warehouses to support biomedical research. Metadata can also provide specific labels for actions related to clinical research processes and analyses. While there are some standards for metadata that seem to be well accepted, such as the ISO 11179 standard for representing data around data elements and associated values, there is actually some variability in how that might be interpreted (22, 23). Regardless, in addition to the need for metadata to represent very specific data element attributes for management purposes, there is a substantial need for shared metadata standards that describe common attributes of the data related to their source (e.g., electronic health record) or intended research context of use (e.g., baseline assessment or reported adverse event). Standard reference information models – such as the HL7 RIM, the CDISC ODM, and the BRIDGE – show promise and have potential for broad relevance, but widespread use of detailed instantiations for specific clinical domains and studies are lacking, as is research that can evaluate the consistency and interoperability of different applications driven by developers' interpretation of these abstract and multifarious information model specifications.

Methodological standards include approaches for accessing, coding, analyzing, and interpreting data; a framework for such standards is presented later this document. For example, standardized coding systems and metadata can support consistent understanding (computable context) and appropriate application of methods as well as appropriate interpretation of results – particularly when data are missing. There are several summaries that inventory such issues and describe the implications for observational and interventional research (24, 25).

Approaches to Data and Metadata Standards related to EHR data and biomedical research

Most researchers and informatics professionals agree that standards are necessary for data exchange and knowledge aggregation; certainly this has been a claim since the first introduction of electronic information systems (26). Yet the broad and consistent adoption of data collection and representation standards have eluded us for a number of reasons, including their enormous scope and complexity, lack of easy access, and lack of incentive for healthcare organizations using electronic information systems to support immediate business needs. Indeed, as standard coding systems increase in size and complexity, the investment to use them increases, exemplified by the extensive resource investment and pervasive controversy surrounding the ICD-10 adoption (27-30), delays around which are impeding progress toward uniform data standards and national health information infrastructure. In addition to the

growing size and complexity (e.g., for all of the aforementioned coding systems, each unique code often includes a number of component terms and concepts, some required, some optional, and some constrained), the intricacy and variability with how these terminologies are used within local database structures cannot be overstated. While sophisticated terminology models with multiple component concepts (e.g., LOINC) and capabilities for post-coordination for new or refined terms (e.g., SNOMED CT and ICD-10) offer flexibility of coding and streamline terminology management, they create inestimable variability in how complex concepts and clinical phenomena are represented in local systems. For example, the use of units can be included in the LOINC code itself or encoded as a discrete data field called 'units'. The notion of who reported a symptom could be represented in SNOMED CT expression for the symptom, a LOINC code for the assessment, or in the HL7 exchange message or structured document. The concept of negation can be included as part of a SNOMED CT term (e.g., "no family history of heart disease") or in the database itself. This issue of terminology-information ("term-info") model interactions has been known for decades (31, 32) but is generally under-appreciated outside of informatics venues and data standards organizations. There are thousands of these examples, and they get more complex as we consider the types of detailed and domain-specific data that need to be standardized for research cohorts or disease-specific analysis and comparative studies.

The ***standards for EHRs*** are still evolving, but the few standards that are broadly adopted only came as a result of significant financial incentives. Further, their acceptance and adoption as standards came when they were framed in the context of explicit and relevant use cases. (The CHI standards (33) were a start but did not tie into concrete business exchanges or an explicit use case). As such, we encourage NIH leaders to consider this common misstep and identify relevant business cases for clinical research data standards. Further, the vast and growing number of clinical information system products (more than 2,000 certified products as registered with the ONC as of the time of this submission[†]) and variety of associated information needs and documentation preferences make it a challenge to control and standardize, yet we argue that the use of standards are absolutely essential for the use of big data in biomedicine – particularly research interests such as the NIH. Standards for data representation, exchange, and analysis can support consistent – and hopefully prudent – evidence-based approaches to using data from EHRs, labs, devices, etc. to support NIH interests in biomedical research.

There are a number of data standards, with different physical structures, scope of coverage, and governance (14, 15, 34). In fact, a great part of the challenge is to simply identify all potentially relevant

[†] http://oncchpl.force.com/ehrcert/

standards for a particular data collection or data transformation context and to understand the different intents, structures, scope and application for existing standards. There is no single resource for researchers, research sponsors, or policy makers to get an inventory of various standards with updated and objective information about where and how they should be used.

There are at least two broad and different approaches for achieving standardized representation of clinical data from various EHRs: 1) to promote the standardization of the data *as they are collected*, aided by mandates or incentives, or 2) accept that variation in EHR and CDW systems will exist and standardize approaches for using the heterogeneous data that exists now. There are several informatics activity areas that make the best use of heterogeneous data:

- One option is to take data in any form, including unstructured narrative data, and ***map to a reference terminology***, such as SNOMED CT. Indeed, this is what is done in NLP: real language concepts are identified in unstructured text and then codified using some standardized *reference terminology* or ontology as a reference. The use of structured data is more problematic because of the great variety of terminology-information model interactions across thousands of different local systems.
- Extending the idea of a reference terminology, another method for achieving standardized data or interoperability across inherently different systems is to identify a ***common reference information model*** that healthcare EHR vendors could map to. There would, of course, be work for each vendor to map their products (and changing data collection) to these reference standards but in theory this could allow infinite number of applications to be written against the reference standards, hence promoting faster innovation and dissemination. This was the underlying idea for the HL7 Reference Information Model, but a number of experts from the AMIA community have called for a need for shared semantics for both biomedical knowledge and the business of healthcare (35-37). A common set of reference standards would allow heterogeneous local EHR systems and vendor products to persist and much-needed applications that use the data to proliferate. Furthermore, a common set of reference standards would allow developers of new products, apps, and decision-support algorithms (including research-related decision support relevant to NIH- research) to rapidly develop new products. These products, apps and algorithms could be built once to a common reference standard, making development (and testing) faster and potentially enhancing maintenance of the tools and patient safety (38, 39). The path to achieving a set of reference information-models with embedded terminology is not clear. For example, the detailed clinical models (36, 40) developed by Dr. Stan Huff and

colleagues from Intermountain Health Care for years, are the proposed starting point for this effort and are now under continued development in and international collaboration called the Clinical Information Modeling Initiative (CIMI)(41) and used in the Health Care Services Platform.

Opportunities for Community-Driven Data and Metadata Standards

In this age of big data (and availability of existing data sources, such as EHR and personal health information as noted above), it is increasingly important to recognize a broader range of stakeholders for input and consensus of data standards. The concept of multiple uses (secondary use) of clinical data necessarily entails the active involvement of a broad range of stakeholders with a broader range of motivations for implementing systems and collecting data. The scale and complexity of the conversations that need to happen to move toward standardized data are enormous and those required to achieve ("research quality" data) are greater still.

For this RFI, we define "communities" as distributed or centralized groups of individuals that work in a common domain and have some established means of communication and interaction. Some of the most successful standards are developed by communities include grassroots efforts – such as the Gene Ontology (GO) – and more heavily organized efforts such as Logical Observation Identifiers Names and Codes (LOINC) and the Digital Imaging and Communication in Medicine (DICOM) standard for radiological image transfer[‡].

Two broad communities that are well developed and relevant to the notion of research and EHR data standards are CDISC and HL7. Although there is a MOU between the two, they represent very different groups of individuals and organizations with little overlap. The approach to standards for research and clinical have historically been different amongst such communities and represent the different approaches of their constituencies. HL7 has extended significant effort on clinical modeling (though the most adopted standards are the ICD and CPT coding systems used for billing.) The research world thinks in terms of case report forms, component items ("Questions and answers"), and data analysis variables. These concepts can fit into complex models, but the models have yet to be developed. But NIH and other researcher groups and professionals (including informaticians) rarely define standards of practice for how these data can be used. HL7 has tried to engage medical professional societies in the data standards effort but has had limited success. Professional societies and patient groups can enable and enhance the development of meaningful standards and promote their adoption in EHR systems and

[‡] http://grants2.nih.gov/grants/guide/notice-files/NOT-ES-15-002.html#sthash.IU8YRHkO.dpuf

research. All play a part and the need for communication and coordination is unprecedented. NIH can play a role in providing information and facilitating coordination and communication, creating a culture change around standards development and implementation that ensures multi-stakeholder engagement and endorsement of standards.

3. Developing an Extensible and Inclusive Biomedical Big Data Framework

In summarizing all of the themes and recommendations noted in this document, we have identified five critical needs that we believe are relevant to community-level consensus building and standardization efforts (as are called for in the RFI to which we are responding) as follows:

1) the need for ongoing research and development that can advance the foundational computational and informatics theories and methods and their state-of-the-art relative to the collection, storage and analysis of big data – such as the specific standards-setting initiatives and requirements enumerated in Section 2 of this document;
2) the corresponding need to design, implement, and evaluate mechanisms of integrating and harmonizing large amounts of diverse and differentially codified biomedical data spanning the bio-molecular, clinical, and population domains;
3) the further study of the role of big data analytics in terms of improving the analysis of high-value and currently under-utilized data types such as those derived from EHR-based phenotyping algorithms, imaging modalities, and quantitative pathology studies;
4) the expansion of efforts to create community-based and accepted definitions relative to the operational, policy and cultural issues that serve to either promote or prevent the use of big data to improve healthcare related decision-making; and
5) the advancement of efforts to enable community-based data standards at all levels for the purposes of facilitating the use and re-use of big data.

To this end, we recognize the need for standards and also the lack thereof. Of note, the AMIA community has been intimately involved in the evolution of data standards and a national health information infrastructure and fully appreciates the complexity and number of stakeholders with interests and concerns related to national data and metadata standards. AMIA also includes members from a number of communities that have the potential to drive the development, adoption, and refinement of data and metadata standard. We applaud the NIH for its initiative around the topic of big data and standards and encourage NIH to support the adoption of standards in a very proactive way. A number of specific suggestions in this area include the following:

1) ***Provide detailed, current, and relevant information on various standards and guidance on appropriate context of use for researchers and healthcare organizations engaging in research.*** A significant obstacle in the adoption of data standards is the lack of tools to help researchers and research sponsors and policy makers identify all potentially relevant standards for a particular data collection or data transformation context and to understand the different intents, structures, scope and application for existing standards. In the interest of supporting NIH researchers and NIH program staff, the NIH should facilitate the curation of an inventory of standards with updated and objective information about where and how they should be used. This resource could be an inventory of standards and use cases developed in conjunction with other standards stakeholders such as FDA and CMS and CDC. The existence of such information could support other stakeholders and research sponsors – such as private foundations and patient advocacy groups – to endorse and advocate for the permeation of standards, including those designed to support *de novo* research and the secondary analysis of clinical data. Given the number of different stakeholders and communities, it is reasonable to assume that there would be a distributed collaboration component. Further, this information could be a valuable and central resource for program staff and NIH Institutes and Centers, who can incorporate requirements for their development and use into RFPs and sponsored research activities.
2) ***Endorse and/or support standards with good governance, responsive submission processes, and clearly defined scope.*** It has been widely recognized that the number of data standards are proliferating, not merging (40). NIH should recognize that, as capacities grow for research and the number of electronic sources of data increase, the need for NIH to require standards grows. We urge NIH to be proactive in the endorsement and support of standards efforts – particularly those that have relevance to biomedical research – such as HL7 and CDISC. We also urge support other domain-specific efforts to the extent that they fulfill genuine gaps and can assure that they have done the "due diligence" of identifying and communicating with existing standards. ***The NIH should not necessarily dictate standards (because context and research needs are so broad) but should provide information, supportive tools, and free and easy access to important standards.*** NIH should recognize two important issues in this regard: 1) standards are dynamic; and 2) the breadth of standards and number of codes to represented biomedical research information are tremendous. This dictates a need for NIH to recognize and support defined processes for data standards and/or organizations that are recognized as standards development organizations, which are accredited organizations that have processes in place to ensure inclusion, fairness and consensus.

3) ***Encourage domain-specific groups to get together to agree and move from a proliferation of data standards to assimilation. .*** Our goal should be to allow new data to be quickly standardized, but such processes must include sharing standards and communicating them to broader communities. This requires us to encourage the development and sharing of a wide range of use cases (public health to research) to facilitate the broadest relevance of data, accompanied by said standards.
4) ***Provide incentives for researchers to use specific data standards – including approaches to data extraction (e.g., logical specifications of computable phenotype definitions) – data elements, and value sets.*** The first step to advancing standards is to require that investigators report the codes or logic used in their data sets. There are efforts underway to standardize such reporting, but perhaps the best solution would be to require such metadata submission in a comprehensive manner for NIH-funded research. This could likely be tied to publication and clinicaltrials.gov requirements. This goal likely requires that we consider both mandates and incentives to ensure that any published or NIH-funded research studies using EHR data for research purposes (sampling, etc.) also report upon accompanying definitions (phenotypes) and data about the distribution of features and quality of the data.
5) ***Consider the need for common, detailed clinical data reference models and associated terminology for EHR data and allow research applications to pull from them.*** The lack of common semantics – plus the fact that the information systems and the coding systems therein are developed and maximized to support local business needs – create enormous challenges for data completeness and quality relative to any kind of scientific inquiry. The issue of heterogeneous data collection will not go away. Mandates will take years. The process is not agile enough for new research needs. As such, we should consider better supporting the CIMI effort and HL7 and AMIA communities in such capacities.
6) ***Coordinate NIH efforts, particularly distributed research networks and registries.*** Though a challenge, the need for trans-NIH coordination across major networks is unprecedented. The perspective or perhaps the drivers of such coordination are the health care and research organizations (academic medical centers) that are required to report to all these entities. Coordination and harmonization of such efforts in a meaningful and comprehensive way is critical to eliminating barriers to adoption and participation therein.

Beyond the preceding tactical issues, during the course of the preparation of this response, a broad and reoccurring theme was voiced by the involved domain experts concerning the current and future state

of big data in healthcare, namely: ***how can we as a community ensure that we are both asking and answering the right questions in the context of such big data resources, particularly given the high costs associated with their assembly, analysis and dissemination?*** A conceptual framework for generating value from big data emerged in this regard, as is illustrated in **Figure 1** and described below.

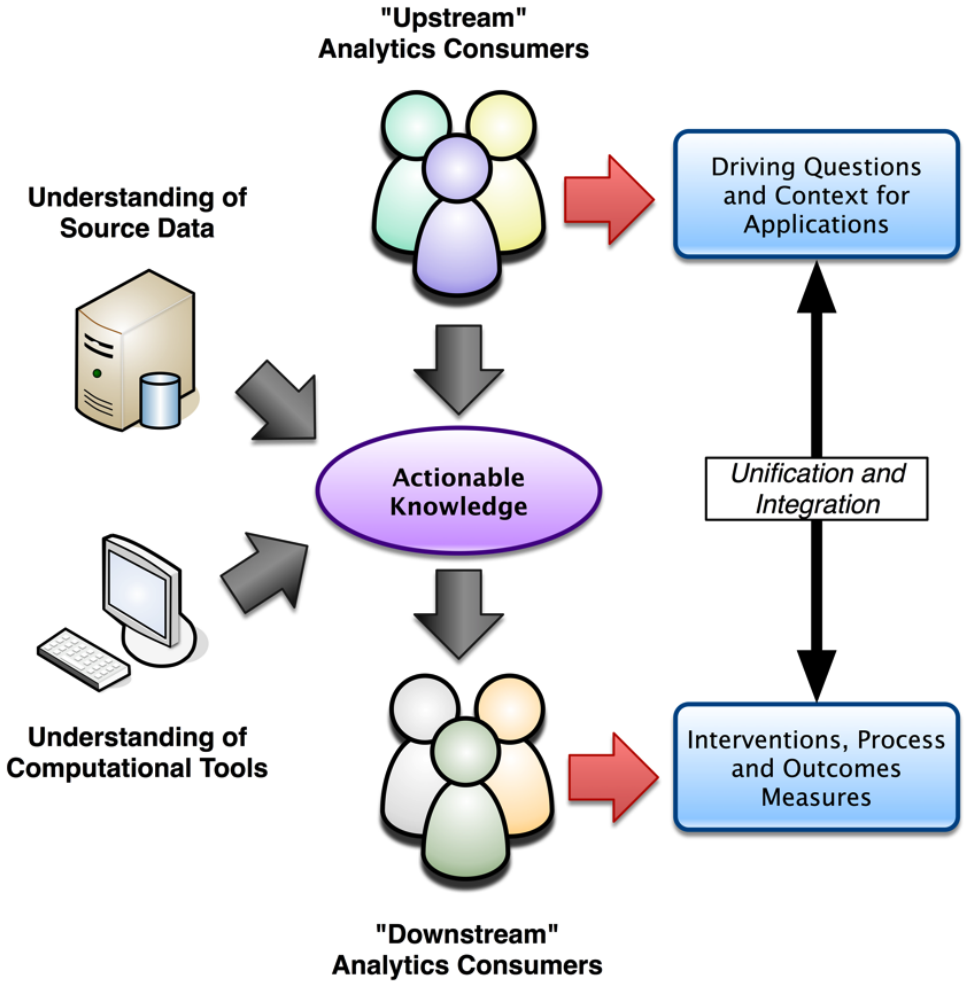

Figure 1: Conceptual framework for a solution-oriented framework describing the high-value use of big data in healthcare-relevant research and practice.

The aforementioned conceptual framework for value generation in the context of big data involves four major components, one primary outcome, and an overarching and unifying framework, as follows:

- At the core of the big data value formation model is the interplay between **four major components**, namely: 1) "upstream" analytics consumers (e.g., researchers and/or strategic decision makers) who serve to define driving problems that require data-centric solutions, which

can only be satisfied via the integrative analyses of integrated collections of large-scale and heterogeneous data; 2) "downstream" analytics consumers who may not necessarily define driving problems, but have a research or clinical need to apply information products generated via big data analytics in order to advance hypothesis generation, testing, clinical decision making and/or population health management; 3) processes and tools that enable systematic and tractable understanding of source data, where such understanding may have to be inferred using numerous computational techniques given the scarcity of well-characterized data in the biomedical domain; and 4) processes and approaches to ensure that the preceding stakeholders understand what big data analytics tools exist, how they are appropriately used, and how to interpret their outputs.

- The **primary outcome** of the big data value formation model is that of "actionable knowledge." This term is used purposely to indicate that such knowledge should both contextualize data products and deliver them in the correct format to the correct individual(s) at the correct time, so as to optimally support or influence decision-making or other data-driven workflows.
- Finally, the **overarching and unifying framework** for this big data value formation model is to "connect the dots" between: 1) the driving question and contextualizing factors articulated by "upstream" data analytics consumers; and 2) the interventions, processes and outcome measures associated with the decision-making or other data-driven workflows of "downstream" data analytics consumers (e.g., creating a virtuous and self-supporting cycle of both asking and answering big data questions).

It was the consensus of the individuals who were engaged in the preparation of this response that it is only possible to ask and answer questions that matter in a healthcare research and practice context – and thus generate value from big data resources – if the relationships between the preceding components and outcomes are uniformly strong and well understood. It is imperative that such questions be informed and motivated by the overarching and unifying framework that serves to link the information needs of "upstream" and "downstream" consumers in synergistic and quantifiably measurable ways. Further, the notion of big data generally presupposes a need for standards – that the difference in data can be thought of as noise and one can easily be overcome by the sheer volume of data. This may be true, but the need for reproducibility and quality (provenance) in research might require that the rigor of both coding of EHRs and EHR data management increase dramatically in order to keep up with research-level standards. While some analytic approaches and models can accommodate lack of standards by reducing data to triples, it is possible that they only postpone the

complexity. For some purposes – such as data mining – it is possible that we could have true knowledge discovery and data mining without defining things. But for true research – including health services research and scientific research –explicit and reproducible data standards are absolutely essential. Achieving such an ideal end state – in which big data is leveraged to create clear and demonstrable value based on the use of data standards and harmonization methods – will require a number of critical initiatives, including the following:

1) Create and deliver tailored workforce and knowledge development programs to ensure that all stakeholders depicted in our proposed model (**Figure 1**), are knowledgeable and informed relative to identifying, interacting with and generating value from big data.
2) Develop and apply appropriate methods to ensure that big data platforms are capable of empowering these knowledge workers and key stakeholders to leverage their domain knowledge and support the discovery, integration, harmonization and analysis of "big data."
3) Create a collaborative research analytics commons for big data resources, tools and best practice documentation in order to promote an innovation and knowledge ecosystem surrounding healthcare-relevant big data analytics.

4. Conclusions

It has been argued that using big data to improve or advance health relevant research and practice represents the future of both healthcare and data science fields. Current approaches to the use of big data in the healthcare and life science communities have been very focused on the "how" of such information management, such as how we can collect, store, and transact in increasingly voluminous data sets. This has led to an appropriate emphasis on technical and associated policy issues that must be addressed in order to create a "fluid" economy of such "big data." These types of issues are certainly timely and will serve as the foundations of an emerging big data era in science and medicine. However, as was identified and conceptualized during the course of the preparation of this response on behalf of AMIA, it is also equally important to identify strategies to ensure that we are in fact asking and answering the right questions given the availability of increasingly large and complex data sets – particularly given their significant costs and resource demands therein. The consensus of the authors of this document was that doing so requires a broad understanding of big data and how it can be made relevant to addressing biological or clinical problems. This level of understanding requires big data analytics drivers and consumers alike to achieve: 1) an appreciation of the intrinsic nature of the data sets themselves; 2) the information needs of "downstream" consumers of analytical products; and 3)

familiarity with computational tools that can coexist with human expertise and enable the discovery of actionable knowledge.

Achieving such ends requires continued investments in and support of wide-ranging efforts to create and disseminate the theories and methods that are foundational to an extensible and inclusive biomedical big data framework, as we have described in this document. Doing so will require us to address current and persistent gaps in such theoretical and applied knowledge, with a particular emphasis on the areas noted previously, such as: 1) creating tailored and targeted workforce and knowledge development programs; 2) empowering knowledge workers to be integral parts of the data discovery, integration, and analysis "pipeline"; and 3) establishing and sustaining community-wide data analytics "commons" and data standards setting initiatives. We believe it is imperative for efforts that hope to address the preceding issues to employ and leverage entities with community-wide convening authority, such as AMIA, to ensure that resulting frameworks are not only extensible, but are also inclusive of the full spectrum of stakeholder needs and expertise.

References:

1. Bourne PE. What Big Data means to me. J Am Med Inform Assoc. 2014;21(2):194.
2. Jain SH, Rosenblatt M, Duke J. Is Big Data the New Frontier for Academic-Industry Collaboration? JAMA. 2014;311(21):2171-2.
3. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. JAMA. 2014;309(13):1351-2.
4. Schneeweiss S. Learning from Big Health Care Data. NEJM. 2014;370(23):2161-3.
5. Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury MJ, Hesse BW. Collaborative Biomedicine in the Age of Big Data: The Case of Cancer. J Med Internet Res. 2014;16(4):e101.
6. Wang W, Krishnan E. Big Data and Clinicians: A Review on the State of the Science. J Med Internet Res Med Inform. 2014;2(1):1.
7. Kush R, Goldman M. Fostering Responsible Data Sharing through Standards. NEJM. 2014;370(23):2163-5.
8. Halamka JD. Early Experiences With Big Data At An Academic Medical Center. Health Affairs. 2014;33(7):1132-8.
9. Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Affairs. 2014;33(7):1163-70.
10. Phillips KA, Trosman JR, Kelley RK, Pletcher MJ, Douglas MP, Weldon CB. Genomic Sequencing: Assessing The Health Care System, Policy, And Big-Data Implications. Health Affairs. 2014;33(7):1246-53.
11. Longhurst CA, Harrington RA, Shah NH. A 'Green Button'For Using Aggregate Patient Data At The Point Of Care. Health Affairs. 2014;33(7):1229-35.
12. NIH. The NIH Director's Panel on Clinical Research Report to the Advisory Committee to the NIH Director, December, 1997: NIH Director's Panel on Clinical Research (CRP); 1997 [cited 2011 May 15]. Available from: http://www.oenb.at/de/img/executive_summary--nih_directors_panel_on_clinical_research_report_12_97_tcm14-48582.pdf.
13. NIH. Final NIH Statement on Sharing Research Data, February 26, 2003. National Institutes of Health, 2003 NOTICE: NOT-OD-03-032.
14. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007 Nov-Dec;14(6):687-96. PubMed PMID: 17712081. Pubmed Central PMCID: 2213488. Epub 2007/08/23. eng.
15. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc. 2011 May 1;18(3):341-6. PubMed PMID: 21486890. Pubmed Central PMCID: 3078665.
16. Loring DW, Lowenstein DH, Barbaro NM, Fureman BE, Odenkirchen J, Jacobs MP, et al. Common data elements in epilepsy research: development and implementation of the NINDS epilepsy CDE project. Epilepsia. 2011 Jun;52(6):1186-91. PubMed PMID: 21426327. Pubmed Central PMCID: 3535455. Epub 2011/03/24. eng.
17. Biering-Sorensen F, Charlifue S, Devivo MJ, Grinnon ST, Kleitman N, Lu Y, et al. Using the Spinal Cord Injury Common Data Elements. Topics in spinal cord injury rehabilitation. 2012 Winter;18(1):23-7. PubMed PMID: 22408366. Pubmed Central PMCID: 3298358.
18. Grinnon ST, Miller K, Marler JR, Lu Y, Stout A, Odenkirchen J, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project - approach and methods. Clin Trials. 2012 Jun;9(3):322-9. PubMed PMID: 22371630. Pubmed Central PMCID: 3513359.
19. Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc. 2003:1048. PubMed PMID: 14728551. Pubmed Central PMCID: 1480162.
20. Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. Methods Inf Med. 2006;45(6):594-601. PubMed PMID: 17149500. Pubmed Central PMCID: 2980785.

21. NLM. The NIH Common Data Element (CDE) Resource Portal: National Library of Medicine; 2013 [cited 2013 March 6]. Available from: http://www.nlm.nih.gov/cde/
22. Park YR, Yoon YJ, Kim HH, Kim JH. Establishing semantic interoperability of biomedical metadata registries using extended semantic relationships. Stud Health Technol Inform. 2013;192:618-21. PubMed PMID: 23920630.
23. Ngouongo SM, Lobe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? J Biomed Inform. 2013 Apr;46(2):318-27. PubMed PMID: 23246614.
24. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013 Aug;51(8 Suppl 3):S22-9. PubMed PMID: 23793049.
25. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. Yearb Med Inform. 2014;9(1):215-23. PubMed PMID: 25123746.
26. Hammond WE. The making and adoption of health data standards. Health affairs. 2005;24(5):1205-13 %@ 0278-2715.
27. Boyd AD, Li JJ, Burton MD, Jonen M, Gardeux V, Achour I, et al. The discriminatory cost of ICD-10-CM transition between clinical specialties: metrics, case study, and mitigating tools. J Am Med Inform Assoc. 2013 Jul-Aug;20(4):708-17. PubMed PMID: 23645552. Pubmed Central PMCID: 3721160.
28. Averill RF, Bowman SE. Don't delay implementation of ICD-10. Health Aff (Millwood). 2012 Jul;31(7):1650. PubMed PMID: 22778363.
29. Chute CG, Huff SM, Ferguson JA, Walker JM, Halamka JD. There are important reasons for delaying implementation of the new ICD-10 coding system. Health Aff (Millwood). 2012 Apr;31(4):836-42. PubMed PMID: 22442180.
30. Meyer H. Coding complexity: US Health Care gets ready for the coming Of ICD-10. Health Aff (Millwood). 2011 May;30(5):968-74. PubMed PMID: 21555481.
31. Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An event model of medical information representation. J Am Med Inform Assoc. 1995 Mar-Apr;2(2):116-34. PubMed PMID: 7743315. Pubmed Central PMCID: 116245.
32. Rocha RA, Huff SM. Coupling vocabularies and data structures: lessons from LOINC. Proc AMIA Annu Fall Symp. 1996:90-4. PubMed PMID: 8947634. Pubmed Central PMCID: 2233073.
33. CHI. CHI Executive Summaries. Consolidated Health Informatics, 2004 Contract No.: May 12.
34. Tenenbaum JD, Sansone SA, Haendel M. A sea of standards for omics data: sink or swim? J Am Med Inform Assoc. 2014 Mar-Apr;21(2):200-3. PubMed PMID: 24076747. Pubmed Central PMCID: 3932466.
35. Bakken S, Campbell KE, Cimino JJ, Huff SM, Hammond WE. Toward vocabulary domain specifications for health level 7-coded data elements. J Am Med Inform Assoc. 2000 Jul-Aug;7(4):333-42. PubMed PMID: 10887162. Pubmed Central PMCID: 61438.
36. Oniki TA, Coyle JF, Parker CG, Huff SM. Lessons learned in detailed clinical modeling at Intermountain Healthcare. J Am Med Inform Assoc. 2014 Nov;21(6):1076-81. PubMed PMID: 24993546. Pubmed Central PMCID: 4215059.
37. Rector AL, editor The Interface Between Information, Terminology, and Inference Models. Tenth World Conference on Medical and Health Informatics: MedInfo-2001; 2001; London.
38. Goossen W, Goossen-Baremans A, van der Zel M. Detailed clinical models: a review. Healthcare informatics research. 2010 Dec;16(4):201-14. PubMed PMID: 21818440. Pubmed Central PMCID: 3092133. Epub 2011/08/06. eng.
39. Goossen WT, Goossen-Baremans A. Bridging the HL7 template - 13606 archetype gap with detailed clinical models. Stud Health Technol Inform. 2010;160(Pt 2):932-6. PubMed PMID: 20841821. Epub 2010/09/16. eng.

40. Jiang G, Evans J, Oniki TA, Coyle JF, Bain L, Huff SM, et al. Harmonization of Detailed Clinical Models with Clinical Study Data Standards. Methods Inf Med. 2014 Nov 26;54(1). PubMed PMID: 25426730.
41. CIMI. The Clinical Information Modeling Initiative (CIMI) Wiki: The Mayo Clinic; 2013 [cited 2013 April 8]. Available from: http://informatics.mayo.edu/CIMI/index.php/Main_Page.