Testimony of Paul Harris, JASON Task Force

Dear Ms. Wilson, July 28, 2014

Thank you very much for the invitation to attend the JASON Task Force Hearing on July 31st. I apologize that scheduling conflicts prevent my attendance, but please find responses below to your proposed research-domain questions. If I can be of further assistance, please don't hesitate to contact me.

Best regards,

Paul A. Harris, PhD
Vanderbilt University

1. **Could you address the tension in the JASON report around consumer control of data sharing versus unfiltered data?**

   "The tension for data access regarding the balance between patient privacy and the potential societal benefit of access to patient data" is real and the importance of patient preference and perception should not be underestimated. The architecture proposed in the JASON report is ambitious and the scale of aggregated data will need to be large in order to justify the resources and time required for implementation. We have seen in the research community that models providing consumers revocable, fine-grain control over individual data used in aggregate for research purposes (e.g. discovery, feasibility, retrospective analyses) can be overly burdensome. In addition, when patients have been asked their desires for this model, studies have shown only a small proportion of individuals would want to exercise a high level of control.

2. **What do you need in terms of unbiased participation?**

   Large data collections are needed for data-driven hypothesis generation and discovery (e.g. record counters, creation/testing of robust predictive modeling and machine learning algorithms). Assembling these collections will require sizable technological efforts and appropriate policy and administrative infrastructure. The JASON report does a nice job articulating issues and a potential architecture to support access, collation, storage, transport, maintenance, and security of data – from a national perspective. Starting from this foundation, more work may be needed to define participation and implementation incentives at local sites for research use cases. While large datasets increase the ability to detect signal from noise, it has been our experience that even large dataset work requires investment of content and knowledge experts at the local level.

3. How they feel about opt in or opt out option (for patients)?

Vanderbilt has extensive experience implementing a patient opt-out model through BioVU, a biorepository which contains DNA extracted from discarded patient blood specimens as part of routine clinical care. Currently, the resource contains >180,000 DNA samples from adult and pediatric patients, linked to the Synthetic Derivative database (containing de-identified electronic medical records). Rapid and sustaining accumulation of DNA samples at this capacity has been possible by the election to utilize an opt-out model, whereby patients are presented the opportunity for sample exclusion. Vanderbilt patients encounter the BioVU opt-out language during the registration process, subsequent to standard consent to treat forms. The BioVU opt-out language includes a description informing the use of blood for uses such as in DNA research followed by a check box signifying an official opt-out, and patients retain the ability to actively opt-out at any time. The election of opt-out is permanent, and overall the rate of opt-out is approximately 20%.

Patient or legal guardian perspectives and attitudes surrounding the opt-out model were surveyed to inform biorepository development using a series of community studies (2008-2011) involving both adult and pediatric phlebotomy clinics. Our studies highlight the need for public transparency and information dissemination regarding these programs. Success of the opt-out model is highly dependent upon patient awareness and continued education related to the societal value of contributing to the biorepository program.

References – Research Question 3

Brothers KB, Morrison DR, Clayton EW. Two Large-Scale Surveys on Community Attitudes Toward an Opt-Out Biobank. American Journal of Medical Genetics Part A. 2011

Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84:362-369.

Pulley JM, Brace MM, Bernard GR, Masys DR. Attitudes and perceptions of patients towards methods of establishing a DNA biobank. Cell Tissue Bank. 2008 Mar;9(1):55-65.

Pulley JM, Brace M, Bernard GR, Masys D. Evaluation of the effectiveness of posters to provide information to patients about a DNA database and their opportunity to opt out. Cell Tissue Bank 2007;8:233-241.

4. **What is your position on consent? JASON report argues that de-identification techniques are not robust for ensuring privacy so they advocate the need for informed consent. Federal research rules does include a form of consent via professional committees- IRBs Consent may cause relevant data to be filtered.**

We have seen in our own experience that privacy, confidentially and security methods are predominantly technical, as are the methods for minimizing risk of reidentifiability. However, we do not underestimate the issues associated with patient preference and perception, including fears of disclosure resulting from breaches of confidentiality, or the policy aspects of maintaining confidentiality. These aspects are equally critical to augment technical methods and include ongoing patient preference assessment, stakeholder engagement, and rigid policy enforcement.

Our current, local de-identification procedures enable data sharing. The de-identification methodology applied to our EHR results in what we call the Synthetic Derivative (SD), and is based primarily on the systematic removal of the fields that are specified in Section 164.514 of the HIPAA privacy rule. The creation of a centralized de-identified resource composed of the entire EHR linked to other important data types and used for research promotes privacy for patients and streamlines sharing of data across institutions. This process also allows efficient pooling of data, as the data within the SD are already de-identified and deemed non-human subjects by the Vanderbilt IRB. Recognizing that technical solutions are not complete, we combine both policy and technology to create a more robust approach. Access to data within our databases requires the user to sign a data use agreement, which outlines the approved uses of the data and requires, among other things, that the user does not attempt to re-identify any records and safeguards the data appropriately. The institution enforces the data use agreement with penalties for non-compliance up to and including termination. While our local model serves the Vanderbilt research enterprise well, we recognize and agree with the Jason report concerning the fact that privacy protection at a national level will require new, rather than simply scaled, approaches.

With regard to sharing data for collaborative studies, the expected terms of usage are already approved and in place to allow qualified investigators from other institutions who collaborate with a Vanderbilt researcher access to the SD and/or BioVU resource. Our research has shown that risks to re-identification can be formally quantified and mitigated, but that solutions need to be contextual and tailored to specific studies or types of data.

References – Research Question 4

Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. Journal of investigative medicine : the official publication of the American Federation for Clinical Research 2010;58:11-8.

Malin BA. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. Journal of the American Medical Informatics Association : JAMIA 2005;12:28-34.

Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. Human genetics 2011;130:383-92.

5. What challenges and successes have you had to date collecting and utilizing data from EHRs and other health IT systems? Would a JASON like architecture help address the challenges you encountered?

Our team has significant experience managing issues associated with multi-site, collaborative national scale programs, including our participation as a site and leadership as a coordinating center for the NHGRI-funded eMERGE network. This network combines EHR systems with biorepositories across the nation for large-scale, high-throughput genetic research with the ultimate goal of implementing genomic testing in a clinical care setting to improve care. We have also recently been awarded a PCORI grant with focus on developing a Mid-South Clinical Data Research Network. In this initiative, we have partnered with several organizations (including Greenway Medical Technologies) to provide a large data sharing and research implementation network covering a wide geographic area around the United States. While our work in eMERGE, PCORI and other national programs provides evidence that managing and leveraging data from multiple IT systems is achievable and useful for the advancement of scientific research, several generalizable challenges related to the JASON report and proposed architecture are provided below:

The data integrity problem

- **Incompleteness, inconsistency, and inaccuracy** - The primary role of clinical users is caring for patients, and technology must support and complement this mission. Subsequently, the resulting data might be incomplete, in different formats or missing altogether, and need to undergo a careful cleanup and transformation process before they can be used for research.

- **Role of Standards** - While there are many standards and each standard has benefits and shortcomings, no single standard (syntactic or semantic) will meet data exchange needs. Furthermore, we have found that there are no existing standard(s) for some data, e.g. Patient Reported Outcomes.

- **Unstructured data** – All clinical information cannot directly be structured and standardized, which makes data interchange standards difficult to obtain. Not having standard inputs makes data consolidation efforts extremely difficult. In addition, time is an important factor in understanding clinical outcomes, but many temporal relations are not *explicitly* stated in the clinical narratives, but rather needs to be inferred.

The JASON report states, that "In the case of the HIT software architecture, the APIs will need to be negotiated by the stakeholders and codified through an open process." While this would be helpful in addressing above, it is our experience that such efforts only marginally address this problem, this will always be an issue and would need to be part of assumptions in building the architecture.

The access and curation problem

- **Master Patient Indexing across multiple systems for record linkage** – A patient's data residing in more than one EHR is an identified problem.

The JASON report solution, "A straightforward way to prevent mis-ID problems is to associate a unique identifier (UID) with each EHR. There is currently substantial opposition to implementing a national system for assigning UIDs for health care."

While we agree that a UID for each EHR would be an asset, our experience with our Medical Record Number locally and Research Identifiers in our collaborations, is that person disambiguation is an issue that continues to arise, and a technical solution for Master Patient Indexing is greatly needed.

The data security problem

- **Re-identification risks** - One of the challenges with assessing and mitigating privacy risks when sharing patient data in clinical research environments is that the system is ever changing. This is problematic because standard re-identification risk and data anonymization methodologies have been developed under the assumption that the data is static and unchanging (i.e., one-time data collection and release). Yet, datasets will be extracted from clinical data warehouses and will subsequently contain different amounts of information on overlapping cohorts.

Despite JASON's consent model, creating a large national database would still require an assessment of possible security threats. Understanding the re-identification risk as patients are added to the system and the quantity of their information grows over time has yet to be investigated when the underlying data are being revised over time with new types of information, such as that which may be contributed by patients.

The data sharing and use problem

- **Streamlining agreements for data sharing and use** – It is challenging to create simple, easy to use documentation for sharing and using EMR data for research given the diversity in legal and compliance environments facing partner institutions. Streamlined, global data use and sharing agreements would be necessary to both provide additional protection for data security and

to avoid research inefficiency. One such model has been implemented within the eMERGE Network, a process and model we assisted in developing.

Other integration issues

- **Mapping Old data to New standards** – While Meaningful Use I and II bring the Medical community closer to creating a standard electronic perspective, the years of historical data is not necessarily translated to these standards. When data sharing, this data discrepancy would need to be further addressed.

- **Big Data and Search Capacity** – As this resource grows, building a search and indexing capability that returns accurate and timely results is a large effort that will only grow as the resource itself grows. If growth occurs faster than the technical solution, users might find this resource untenable.

- **Genomic Data** – As the Jason report acknowledges, genomic data will continue to be a valued resource for clinical information. As the report states, "The current standard for individual genomes is to sequence to approximately 30-fold coverage, or approximately 1011 bases of sequence data. Although these data can be compressed by denoting only the difference with respect to the reference human genome sequence, there is clearly a rapidly growing need to incorporate vast amounts of genome sequence information into individual EHRs." These data are vast and require experts to translate this information into usable information for both the patient and the researcher. Without education and resources for this, data will be confusing, misleading, or not useful.

References – Research Question 5

Danciu I, Cowan JD, Basford M et al. Secondary use of clinical data: The Vanderbilt approach. J Biomed Inform Published Online First: 14 February 2014. Doi:10.1016/j.jbi.2014.02.003

Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association : JAMIA 2010;17:169-77.

El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PloS one 2011;6:e28071.

Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association : JAMIA 2010;17:322-7.

Malin B, Airoldi E. Confidentiality preserving audits of electronic medical record access. Studies in health technology and informatics 2007;129:320-4.

Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Journal of the American Medical Informatics Association : JAMIA 2011;18:3-10.

Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. Journal of investigative medicine : the official publication of the American Federation for Clinical Research 2010;58:11-8.

Benitez K LG, Malin BA. Beyond safe harbor: automatic discovery of health information deidentification policy alternatives.

Loukides G G-DA, Malin BA. Privacy-preserving publication of diagnosis codes for effective biomedical analysis.