Models for Data Storage & Exchange, Aggregate Data, De-identification/Re-identification

Practical, real world example: Information exchange between health encounter and research infrastructure: Kenneth H. Buetow Ph.D., National Cancer Institute, National Institutes of Health

The National Cancer Institute (NCI) was charged with providing a concrete example of information exchange between health and research settings to facilitate discussion of the topic of models. To this end, we describe below the technical capability NCI has developed that demonstrates the transition and use of health information with various levels of aggregation, de-identification, and return to care delivery setting.

Scenario. A patient in a standard care setting is offered access to experimental treatment and chooses to participate in a clinical trial or share data for clinical research. (The standard practice in pediatric oncology is for children to be treated through clinical trials protocols.)

- The individual is enrolled in a clinical trial through a research registration system that tracks consent and status of institutional review board (IRB) approval. A research Clinical Data Management System (CDMS) connects and receives authorization from the registration system; the clinical trial is executed in the care setting and the research data pertaining to the individual is entered into the CDMS. Upon appropriate authentication and authorization, the treating physician has full access to the patient's individually identifiable health information through the research CDMS.

- Individually identifiable health information generated through the CDMS is returned to the primary care setting as part of the continuity of care record. (This capability was shown in NHIN-caBIG Connect demonstration)

- The individual level health information is "de-identified" and transferred from the CDMS to a clinical trials data management system (CTMS) for analysis (but not de-identified in the strict HIPAA sense in that it retains dates). The "de-identified" data are aggregated for study analysis.

- The "de-identified" data generated from this individual's participation, along with individual level health information for multiple studies are reported from the CTMS, by the institution conducting the clinical trial to a central resource for monitoring safety and comparative evaluation. Because the data contain some HIPAA identifiers, they are encrypted prior to transmission and transmitted under a standardized, HIPAA-compliant data use agreement and other required regulatory controls.

SPECIFIC COMMENTS:

Under this scenario, the patient consents to participate in a specific research study. Although the data are not de-identified from the standpoint of HIPAA, they may be considered a Limited Data Set as defined by HIPAA, and thus shareable without patient authorization under a HIPAA-compliant data use agreement and other appropriate data

security controls. However, for complete transparency, the patient would be told what will happen to the data, including the impossibility of being able to guarantee 100% confidentiality. The provider that is consenting the patient for this particular study should be explicit concerning the 1) importance of health/biomedical data for research, 2) who collects data, 3) how it is managed and shared, 4) the data security policies that are in place (in detail, if people want that information), and 5) what steps will be taken if there are security breaches. In consenting the patient, the provider should explain the intent to have the research data concerning the patient returned data to the healthcare setting; the patient should have the opportunity to "opt-out."

Should the providers want to include any additional HIPAA identifiers to make data more useful when reported to a centralized research resource, they would need to obtain either (1) patient consent/authorization or (2) IRB waiver of consent/authorization based on a determination of minimal risk or impracticability considerations.

Explicit consent/authorization. This is the safest approach from the standpoint of HIPAA compliance to transmit data to centralized resource. However, there are costs/risks:

- High (financial) cost for questionable or moderate gain
- Possibility and even likelihood of introduction of bias into results (this could exacerbate existing bias due to insurance status, geographic access, educational and socioeconomic status). If providers have the option of convincing their patients to opt out, then measurement and generalizability of results for quality, appropriateness, and outcomes of care becomes difficult, yet these are critically important research activities.
- Appropriate data security controls must be in place through policies, practices, and data access agreements; data security measures should constantly be improved as technology advances.

IRB waiver of consent/authorization. Current IRB and HIPAA requirements would permit use and disclosure of the data without permission to a centralized research repository when appropriate ethical and data security controls are in place, such as:

- Transparency of policies and practices to subjects/patients
- Accountability of clinical data custodians (providers/their staffs) to assure data are protected and released only to bona fide researchers, etc. - proper certification and authentication of systems for such custodians/users
- Accountability of researchers through appropriate data security controls to use the data only to advance the research and as appropriate to benefit individual participants (e.g., via return of results). This accountability can be achieved through multiple mechanisms, including ethical/professional standards, contractual agreements, and regulatory/legal standards.

GENERAL COMMENTS

Current strategies to protect individual's privacy while supporting clinical research using data from an EHR:

- IRB oversight – even where not required by regulation in some cases (e.g., Vanderbilt synthetic derivative database)
- Transparent policies and data use agreements: NIH dbGaP database is an example (GWAS policy); access to Medicare data via CMS data use agreements
- Enhanced regulatory protections and enforcement (GINA, HITECH, etc.).

Pitfalls to above strategies:

- IRB oversight may not be the most efficient mechanism to address these issues:
  - Often lack expertise in data security and integrity standards and policies
  - Activity takes biomedical IRBs away from their core focus on interventional clinical trials
  - Perceived lack of accountability and enforcement (not clear this has been demonstrated)
  - Increased reliance/over-emphasis on informed consent … with cost and bias challenges referenced above as a result
  - Often encourage anonymization, which can undermine utility of the research, ability to validate results, and ability to benefit individual subjects if they can't be told of research findings due to inability to re-identify data
  - Often encourage separation of medical records from research, public health, and QI, when all activities are a critical part of the health care ecosystem
- Current EHR technology is relatively immature with respect to research as this does not represent a primary goal. For example, as compared to clinical trials settings, audit trails are poor. Additionally, role-based and context-based access limitations are often more theoretical than actual. Data mapping and standardization of data elements from care to research setting is challenging.

Recommendations to facilitate clinical research using EHRs:

- Educate consumers and patients about the risks and benefits of information-based research. Discussion of risks of misuse of information should be clear, factual and easy to understand, and when applicable should include the fact that discrimination and misuse is prohibited by law.

- The public needs to know about the importance and integration of research, quality improvement, and public health and the impact of bias that is introduced when individuals exercise their "autonomy" (or in many cases uninformed refusal) to decline participation in records research. Recognition of research as a social good is critical; research results are more reliable when larger populations are studied, and in certain cases, inability to capture data on the entire population can either diminish the value or, in the worst cases, lead to inaccurate or incomplete results. For example, only through access to large amounts of data can the promise of personalized medicine be achieved.

- Punish the wrongdoers rather than refusing access to bona fide researchers, quality improvement (QI) professionals, and public health authorities.

- Modify the regulatory structure to reflect the risk of the activity rather than an arbitrary designation. The fact that data is gathered for purposes of quality improvement does not necessarily minimize or remove the risks of misuse of the data; the fact that the goal is to develop generalizable information does not make the risk higher. The focus should be on what the real risks are to participants, both in terms of use and non-use. Participants should also be told of potential benefits of use

- Opt-out, rather than opt-in, approaches would be helpful in those situations where waiver is not approvable.

- When determining that obtaining informed consent or HIPAA authorization in research is impracticable, IRBs need to include situations where the validity of the research would be adversely impacted by selection bias, where large populations are needed to identify (or disprove) trends, and similar situations. While some IRBs and Privacy Boards already embrace this concept, others do not. Explicit guidance from regulators regarding this issue would be helpful.

- Aggregation of data frequently decreases the likelihood that a particular patient can be identified. However, this approach does not work well for some projects, including rare diseases. While aggregation may be a useful approach, it should not be mandated to the point of removing the flexibility that exists under current regulations. Otherwise we will undermine the ultimate viability and utility of a great deal of research, QI, and public health activities. While regulations and standards can encourage reliance on these mechanisms, they should not require it across the board.

- Current IRB and HIPAA requirements permit use and disclosure of even identifiable data without permission when appropriate ethical and data security controls are in place. For example, healthcare providers were permitted to provide identifiable health data to the National Oncologic PET Registry (NOPR), a registry for a CMS coverage with evidence development project.

- While not covered by the scenario above, many patients are treated without participating in clinical trials. If their health information is collected in EHR systems that utilize structured common data elements, those information assets can be de-identified, aggregated and shared, subject to appropriate ethical and data protection measures.