Adaeze Ajoku
The Meerkats
SI 501 – Section 2
The Data Office
October 24, 2016

Background Research Report

Summary of Client's Mission

The Data Office for Clinical and Translational Research, a part of the UMMS Office of Research, aims to assist UM faculty in their research endeavors by providing confidential, and HIPAA compliant, clinical data from the UMHS. Users can access this comprehensive database, consisting of over 3.8 million patients records, either through a variety of self-service tools or by setting up a customized data pull in tandem with a Data Office staff member. The Data Office also offers consultations about data needs, guidance on maintaining security of the data, and genomic data upon request (University of Michigan Medical School, 2016).

Summary of Client's Problem

The true purpose of the Data Office is to facilitate research by serving as one-stop source for all clinical data needs. However, they struggle with effectively conveying their services to researchers across campus. The Data Office would like to better incentivize repeat users by discovering pain points in the user journey. They have identified the usability and intuitiveness of their website and tools as an area for improvement. Additionally, The Data Office would like to increase awareness of their offerings both inside and outside of the UMMS.

Three Questions

1. How does HIPAA affect the accessibility of clinical data?
2. What is the prevalence of secondary use of clinical data? What are the rewards/risks of secondary use of clinical data?
3. What type of research is currently being done? And what are some future directions?

Report Word Count: 2325

Navigating the Secondary Use of Clinical Data

The Electronic Medical Record Explained

Wishard Memorial Hospital physician Clement J. McDonald created the first at the Electronic Medical Record (EMR) in 1972. A project he anticipated would take a year, ended up spanning a quarter of a century (EMR 23). The standard medical record typically contains information on Demographics, Medical History, Family History, Social History (i.e. employment, religion), Health-influencing Habits (i.e. alcohol and tobacco use), Orders (i.e. for medications or lab tests), amongst other things (EMR 25).

Not too long after McDonald's work other healthcare providers began to digitize medical information as well. However in the early days, these databases were not interoperable. In the late 1980s efforts were made to move towards integration. By the 1990s the ability to access patient information across medical institutions had been further fleshed out resulting in a system that allowed doctors to be more efficient and cost-effective in their interactions with patients (EMR 26).

Before further discussion on EMRs, it should be understood that there are three differing terminologies with subtle differences used to refer to clinical records:

Figure 1:

Electronic Medical Record (EMR)	Electronic Health Record (EHR)	Personal Health Record (PHR)
Health- related information on an individual that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization.	Health- related information on an individual that conforms to nationally recognized interoperability standards and that can be created, managed, and consulted by authorized clinicians and staff across more than one health care organization	Health- related information on an individual that conforms to nationally recognized interoperability standards and that can be drawn from multiple sources while being managed, shared, and controlled by the individual.

(*Source: EMR 31*)

However in this context EHR will be used only when referencing interoperability of clinical information systems and EMR will be used in all other instances.

At present, the United States has no national entity governing EHR framework, although there is an elected National Coordinator for Health Information Technology (Governance 123). To rectify this, the United States aims to encourage digitization of patient health information into interoperable systems in lieu of developing a single federal EHR. The complexity of constructing a database to manage the records of over 300 million individuals nationwide is a deterrent of the latter.

Initially adoption rates amongst healthcare providers was low due to high upfront EMR

installation prices of $33,000 for a stand-alone system and up to $65,000 for an integrated system (EMR 59-60). To incentivize the switch from paper records to electronic ones, in 2010, the US passed the Health Information Technology for Economic and Clinical Health (HITECH) Act offering financial rewards to hospitals (Progress 146). By January 1, 2015, the government had paid more than $28 billion in meaningful use incentive payments dispersed amongst 426,000 hospitals (EMR 4). As a result, the percentage of physicians using EHRs in 2015 (83%) was more than double that of 2007 (35%) (EMR 7). Moreover, the economy for EMRs is also booming and is expected to rise to $35.2 billion by 2020 (EMR 9).

HIPAA and the Accessibility of Clinical Data

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) is legislation that governs the use of patient data. Its Privacy Rule established standards for data protection, which covered entities that maintain clinical data (such as hospitals or health plans) must abide by (HHS.gov). Clinical data is considered identified because it contains patient identifiers such as a patient's name, address, birthdate, or hospital ID number. This data is governed by HIPAA and cannot be accessed without approval of the Internal Review Board (IRB), the research ethics board of the National Center for Health Statistics (NCHS) that governs which entities can use patient data (without the patient's informed consent) (Protection 94) (Legislative 82).

De-identified data is data that has been stripped of these identifying elements. The HIPAA Privacy Rule designates standards for the de-identification of data involving two methods. The first is it defines eighteen data fields that are protected and must be removed during the de-identification process. The second is an expert's assessment of the risk of re-identification. This assessment involves considering the probability that a patient may be accurately linked to another less secure record. For example, a patient that is stored in another non-health dataset by name and other identifiers may be linked to their de-identified persona in a health dataset (De-identifying 148-149).

However because the de-identification process is not infallible, access to de-identified clinical data for analysis still requires IRB approval in most cases (De-identifying 148). The exception is de-identified micro-data (subset of a dataset), which is considered to have a low re-identification risk. Thus, the NCHS creates micro-data files from its population health survey data. This is considered secure because the micro-data is a sample of the entire population and there's no way of knowing which individuals comprise the sample. Also patient identifiers are removed, and any remaining risky data variables are masked. If maintaining the structural integrity of the dataset makes it impossible to mask a data variable, then those wishing to access the file must sign a data sharing agreement stipulating they will not reveal any detailed information about the dataset (De-identifying 154-155).

Additionally, the IRB also approves any data linkages. For instance, a survey administrator wishing to link a respondent's answers to other health data based on his/her submitted Social Security Number would need IRB authorization (Note: patient consent is not needed as the submission of a SSN assumes consent) (Protection 95).

However there are bodies with access to personal health data that exist outside of HIPAA

protection, such as gyms, private labs, fitness apps, and life insurance providers. These entities can- and often do- sell identified data for marketing purposes. Similarly, researchers who are not working for health insurance companies or healthcare facilities have no legal obligation to keep clinical data secure (Legislative 73).

Secondary Use of Clinical Data: The Considerations

Prevalence of Secondary Use

A good number of countries are making strides in the usage of personal data stored in EMRs. In Denmark, for example, research arising from secondary use of clinical data has supported evidence-based decisions from forecasting necessary enrollment size for medical schools to meet future health demand, to evaluating the efficacy of various substance abuse treatment programs. France has created highly secure storage areas for its clinical data and has been promoting retrospective research use of clinical data for years (Progress 132-133).

The United States lags behind in secondary use research in a number of ways. First, there isn't a national patient identifier number for healthcare encounters; different entities within the healthcare system (i.e. hospitals, insurance companies) use different unique identifiers and the process of cross-linking them is cumbersome. Second, due to many distinct EMR systems, there are a plethora of databases. Moreover, because states may act autonomous of federal regulations the laws governing use of personal data are involved and copious. Lastly, attempts to create a more longitudinal health view of a patient can be hindered by frequent insurance plan changes, resulting in an incomplete health record (Progress 133-135).

Characterizing Secondary Use

Retrospective analysis of clinical data has utility for researchers across disciplines. A two-year study at the University of Washington involving over 100 researchers request clinical data found that they represented over 25 academic departments. These researchers inputted a variety of query types, the top 20 of which are displayed in **Figure 2** (Lee 93).

Figure 2: Top 20 Research Queries

Data queried	Requests
Demographic (Age, Gender, Race, Ethnicity, Language, Vital Status, Address/Zip)	89
Visit Details (Institution, Clinic, Type, Date, Frequency, Insurance, Service, Interpreter)	85
Diagnoses (by ICD9)	67
Lab Values (Discrete)	41
Procedures by CPT or ICD9	28
Inpatient Clinical Events (Meds, Echo, Notes, Infusion, Ventilation, Gas stats)	26
Medication Orders	22
Vitals (Weight, Height, BMI, BP)	21
Radiology Information	14
Clinical Notes	14
Provider Information	13
Problem List	11
Pathology Report	8
Social History (e.g. Smoking)	7
Appointments	7
Lab Values (Textual; e.g. Microbiology)	9
Allergy Information	3
Surgery information in Anesthesia System	2
Immunization	2
Emergency Tracking Information	2

(*Source:* Lee *93*)

Providing researchers with access to aggregate data to explore research questions and determine research feasibility can lead to examination of health problems in new, beneficial ways (Lee 94-95).

Clinical Data in the Research Setting

Since EMRs are structured for the extraction of individual patient charts and not ad hoc queries, the data generally needs to be reformatted to enable secondary use (Magrabi 53). Another hurdle is that clinical data tends to be incomplete and irregular and, at times, of poor quality (Progress 131). Generally additional technical support is needed to navigate this transition and resolve any subsequent coding inconsistencies (Tolar 472).

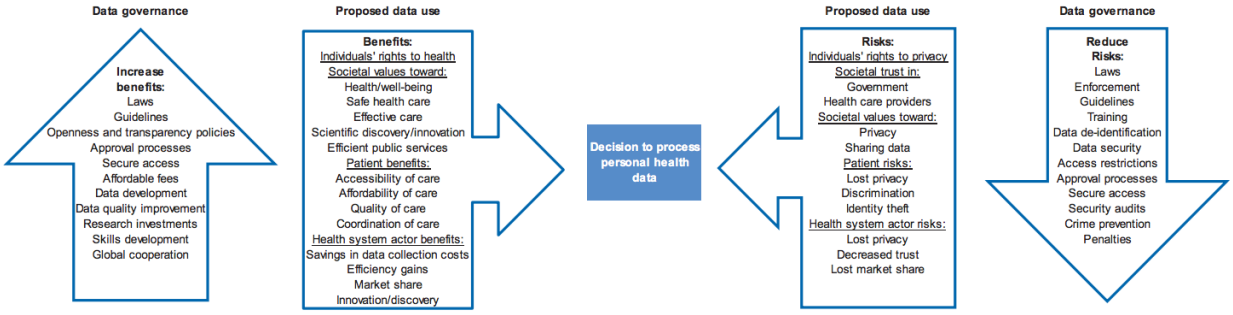
Moreover, inherent within clinical data are many biases that must be understood by the researcher. For example, clinicians introduce biases when they administer tests to certain select patients without indicating the rationale. Plus, the very data available represents another bias because of the decision made to pay for the collection of certain data over others (Magrabi 53).

Furthermore, researchers must take care to avoid p-value hacking (manipulating the data to achieve the research standard of a p-value of less than 0.05) during analysis and to generate reproducible results (Magrabi 53).

Rewards and Risks of Secondary Use of Clinical Data

The measures taken by legislation to protect personal health data seek to strike the optimal balance between mitigating risks and maximizing rewards from research.

Figure 3:



(*Source: Governing 81*)

Rewards

Previous research using pre-existing clinical data has proved very powerful in its ability to predict re-admission likelihood within 30 days for discharged patients, help stymie infection spread, and unearth relationships between pharmacological and epidemiological factors, and more (Magrabi 53). For example, it was the retrospective analysis of over 14,000 infant records that revealed the relationship between erythromycin use and infantile hypertrophic pyloric stenosis (Abhyankar 643).

It is therefore evident that a database of longitudinal clinical data would support monitoring and improvement of healthcare quality (Progress 131). Analysis of this data could make Medicare and Medicaid programs more cost-effective by identifying how members are receiving care. It can also serve as benchmark to evaluate the value of the Affordable Care Act (Progress 134-135). And the accessibility of this data for secondary research permits for innovation through retrospective analysis (Protection 88-89). It will require long-term linkage studies of health, genetics, and environmental factors to discover the agents responsible to disease and improve policy responses (Progress 135).

From an operational standpoint, the existence of an EMR improves healthcare by allowing physicians to act immediately instead of waiting for a chart and by eliminating the issue of unintelligible handwritten notes. Moreover, it reduces the frequency of adverse drug interactions and the frequency of patients receiving expensive tests and procedures unnecessarily (EMR 46-47). There are benefits for researchers as well. EMR usage can improve the quality of the types of patients recruited, as well as reduce the time spent acquiring data (Koepcke 185).

Risks

Within the EMR database itself there is the risk of programming errors, or difficulties in data extraction as the system becomes increasing complex (EMR 47) (Progress 133). Similarly, many of the end-users of EMRs do not comprehend the processes involved in cultivating good quality data. Thusly poor quality data may be utilized without an appreciation for its limitations, or viable results can have their veracity undermined (Tolar 469).

Furthermore, EMRs are susceptible to security breaches. For example, in 2014, the University of Pittsburg Medical Center had a severe data breach that totaled more than 27,000 patients (EMR 138). If personal health data is leaked, there are many potential harms to the patient such as

financial harms from discrimination from employers or health insurers and psychological harms from stigma and embarrassment (especially from the release of sensitive personal health data such as health conditions, STDs, substance abuse, and abortion). This can destroy trust between a patient and his/her healthcare providers (Protection 88) (Legislative 73). Most problematic, the data is vulnerable to tampering or exploitation by an authorized user with maleficent intent (EMR 48). In January 2012, Memorial Healthcare System in South Florida discovered improper actions by two employees scheming to use accessed patient information to process falsified tax returns (EMR 135).

Finally, failure to use EMRs is a missed opportunity for improved healthcare quality and decreased costs: "For example, when an individual is in a health emergency, their care is similar to a battlefield response because their caregivers know nothing about them, including the medications they may be taking. Emergency response could be much safer with the secure sharing of medical records" (Progress 134). Truly, the biggest risk of not collecting and analyzing clinical data is increased morbidity by preventable causes (Protection 88-89).

Conclusion

Innovative Research

There is much potential for retrospective research to generate innovative solutions to plaguing healthcare issues. Secondary analysis of clinical data, especially genomic data, has substantial potential in aiding cancer research (Ochs). At the University of Michigan some exciting research has already been conducted. From January to October 2011, researchers examined 66 consecutive patients who had visited the UMHS for peripheral vascular intervention at initial discharge and at 6 months. Based on this, they were able to isolate an effective therapy linked to a significant reduction in stroke, myocardial infarction and/or death at 6 months. They created an algorithm to gauge the appropriateness of secondary prevention measures based on best practices (Mukherjee). This exemplifies the power of secondary research because it doesn't upset current medical intervention methods, yet can have unprecedented benefits for new patients.

Future Directions

In the near future the incorporation of genetic data (of which each individual has 3 billion DNA base pairs) will explode the volume of clinical data available. Moreover, self-contributed patient data is growing. People are now using sensors and apps that continually gather health information; in example, a glucose monitor can gather over 30,000 glucose recordings in one month (Magrabi 53). This increase in data will improve the type of research questions that can be asked and answered. As secondary use of clinical data grows, it is important that the development and utilization of EMRs have this function in mind because the ease with which data can be extracted from an EMR is heavily influenced by the manner in which it was initially inputted (Levering EHR 495).

Yet as more data get stored, security becomes an even bigger concern. Several large corporations have experienced security breaches and the increased usage of smart phones and other wireless technology bring additional security risks (EMR 20). One proposal to heighten security is the implementation of a secure and verifiable digital signature. Another is tailoring the access any given healthcare worker has to a patient's EMR based on their job role (EMR 48).

Improving the quality and comprehensibility of data in the EMR is also a future goal. Although HITECH Act dictates a quality measure, the US is still working to determine what types of data are requirements for the EHR to support indicators of healthcare quality (Progress 146). And there are moves to add visualizations and built in reminders, templates, and test searches to the electronic platform (Tolar 467).

Bibliography

Abhyankar, Swapna, Dina Demner-Fushman, and Clement J. McDonald. "Standardizing Clinical Laboratory Data for Secondary Use." *Journal of Biomedical Informatics* 45.4 (2012): 642-50.*ArticlesPlus*. Web.

EMR 2016: The Market for Electronic Medical Records. Kalorama Information, 2016. MarketResearch. Web.

HHS Office of the Secretary. "Enforcement Highlights." *HHS.gov*. U.S. Department of Health & Human Services, 20 Nov. 2015. Web. 23 Oct. 2016.

Koepcke, Felix, Stefan Kraus, Axel Scholler, Carla Nau, Juergen Schuettler, Hans-Ulrich Prokosch, and Thomas Ganslandt. "Secondary Use of Routinely Collected Patient Data in a Clinical Trial: An Evaluation of the Effects on Patient Recruitment and Data Acquisition." *International Journal of Medical Informatics* 82.3 (2013): 185-92. *ArticlesPlus*. Web.

Lee, E. Sally, R. Anthony Black, Robert D. Harrington, Peter Tarczy-Hornoch, and Facmi. "Characterizing Secondary Use of Clinical Data." *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science* 2015 (2015): 92. *ArticlesPlus*. Web.

Magrabi, C., BE, Johanna I. Westbrook, MHA, FACHI, FACMI, and Michael R. Kidd et al. "Reuse Of Clinical Data." *IMIA Yearbook* (2014): 52-54. *ArticlesPlus*. Web.

Mukherjee, Debabrata, Prasanth Lingam, Stanley Chetcuti, P. Michael Grossman, Mauro Moscucci, Ann E. Luciano, and Kim A. Eagle. "Missed Opportunities to Treat Atherosclerosis in Patients Undergoing Peripheral Vascular Interventions: Insights from the University of Michigan Peripheral Vascular Disease Quality Improvement Initiative (PVD-QI2)." *American Heart Association, Inc* 106.15 (2002): 1909-912. *ArticlesPlus*. Web.

Ochs, Michael F., John T. Casagrande, and Ramana V. Davuluri. *Biomedical Informatics for Cancer Research*. N.p.: Springer-Verlag, 2010. *ArticlesPlus*. Web.

OECD. " De-identifying Personal Health Data." *Health Data Governance: Privacy, Monitoring and Research*. Paris: OECD, 2015. 147-59. *OECD ILibrary*. Web.

OECD. "Governance of National Electronic Health Record Systems Data

Collection." *Strengthening Health Information Infrastructure for Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges*. Paris: OECD, 2013. 119-30. *OECD ILibrary*. Web.

OECD. "Governing Health Data Access and Privacy: OECD Experiences." *Dementia Research and Care: Can Big Data Help?* Paris: OECD, 2015. 75-89. *OECD ILibrary*. Web.

OECD. "The Legislative Framework Governing Personal Health Data." *Health Data Governance: Privacy, Monitoring and Research*. Paris: OECD, 2015. 65-99. *OECD ILibrary*. Web.

OECD. "Progress and Challenges in Use of Personal Health Data." *Strengthening Health Information Infrastructure for Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges*. Paris: OECD, 2013. 131-48. *OECD ILibrary*. Web.

OECD. "Protection of Privacy in the Collection and Use of Personal Health Data." *Strengthening Health Information Infrastructure for Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges*. Paris: OECD, 2013. 87-102. *OECD ILibrary*. Web.

Tolar, Marianne, and Ellen Balka. "Caring for Individual Patients and Beyond: Enhancing Care through Secondary Use of Data in a General Practice Setting." *International Journal of Medical Informatics* 81.7 (2012): 461-74. *ArticlesPlus*. Web.