A Survey of Data Sources Available for Research in Healthcare and the Underlying Technologies Used for Data Exchange

Doug Foster* and Nitin Karandikar*
*Advanced Data Sciences LLC, San Francisco, California

Background

Healthcare is a data-rich industry. The US healthcare system is estimated to have created a total of 2,314 exabytes of data in 2020.[1] This translates to approximately 30% of the world's data volume. By 2025, the compound annual growth rate of data for healthcare will reach 36%. For context, that's 6% faster than manufacturing, 10% faster than financial services, and 11% faster than media & entertainment.[2]

Stakeholders across the research and medical communities recognize that data - including clinical, lab, claims and other types - provide material advantages for a diverse range of applications including research, care delivery, clinical trials, commercial operations, and new technology development.

This survey is intended to inform researchers about where, and how, to access data for research. In addition to identifying the data sources available for these applications, it also categorizes sources by their 'data sharing model'.

'Data sharing model' refers to the means by which a researcher can access the data both technically and financially if the data is purchased.

Objectives

The survey has two objectives:
(a) quantify the number of healthcare data sources available, and
(b) categorize these data sources by the data sharing model used to access the data.

Methods

Data came from online and database research of publicly available information about organizations offering, or selling, access to healthcare data. Details regarding the data sources were further investigated with direct interviews with organizations hosting data or their vendors.

A categorization system was created to differentiate between data sharing models. Three categories were defined as:

1. Patient registries aggregate data from many different sources, usually EHRs, to create a specialized database for a particular specialty, therapeutic area, indication or patient profile.
2. Data vendors are organizations that aggregate data from various healthcare settings for commercial purposes.
3. Data marketplaces are brokers connecting buyers and sellers of data.

These data were then aggregated into an independent database that was used for analysis.

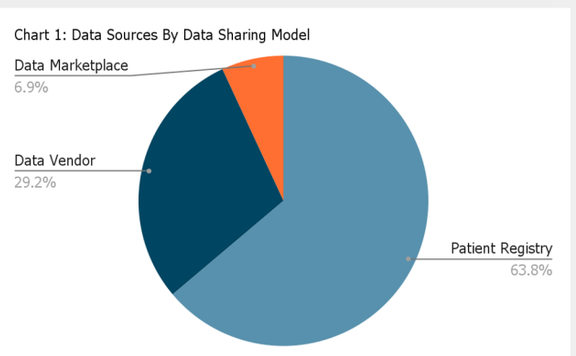
Certain types of data sources are excluded from the scope of this analysis. These include government sponsored databases (eg CMS, FDA, etc.), private databases at academic institutions, and databases of synthetic data that can be used for running simulations and/or training algorithms.

The scope of the exploration is limited to data in the United States.

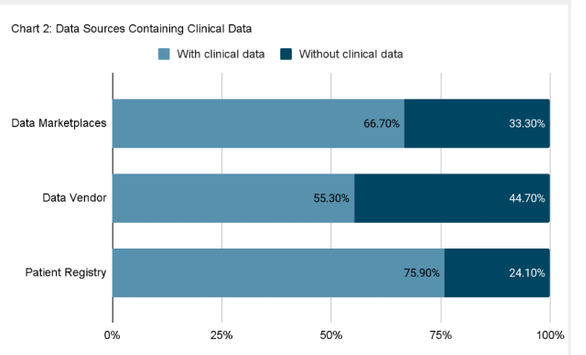
Results

130 data sources were identified during the course of this research. These databases were placed into three categories: (1) patient registries, (2) data vendors, and (3) data marketplaces.

Sixty-four percent (64%) of total organizations were patient registries (n=83). Specialty societies and patient advocacy groups created the majority of patient registries for specific therapeutic areas. Twenty-nine percent (29%) were data vendors (n=38). Seven percent (7%) were data marketplaces (n=9).



69% of these sources contained some elements of clinical data from electronic health records (EHRs). When looking at the percentage of data sources that contained clinical data, data vendors contained the smallest percentage (55%) and patient registries had the highest percentage (75%). Additional data types included claims, lab data, hospital chargemaster data, genetic data, pharmacy, and medical images.



Conclusions

A significant number of healthcare data sources are available for research through a variety of mechanisms to access these data.

The most common data sharing model for research is the patient registry. The patient registries identified in this survey are mostly from specialty societies. Alternative sources from commercial vendors and marketplaces are readily available.

Additional research should be conducted to assess data quality and completeness of these data sources to further investigate the capabilities of these data to be used for research.

References

[1] Nick Culbertson, The Skyrocketing Volume Of Healthcare Data Makes Privacy Imperative, Forbes, August 6, 2021.
[2] Greg Wiederrecht, Ph.D., The Convergence of Healthcare and Technology, RBC Capital Markets.