Big Data in Health Care: We are not there yet

Tuan-Yen Wu
China Medical University Hospital
No.2, Yuh-Der Road, North District, Taichung, Taiwan
E-mail: tywu36@gmail.com
Phone: (+886)921568145

Words count: 1197
Keywords: Electronic Medical Records; Medical Informatics.

Abstract

Five V’s form the core concept of big data: Volume, Velocity, Variety, Veracity and Value.
However, medical application lags behind in these five aspects. The vast majority of medical databases do not meet the definition of "big data". The solution is to expand database through automated medical records and connect various databases. However, great challenges are encountered on our path. Electronic medical records follow strict rules of relational database; integration of databases employ a tremendous amount of manpower and the privacy/security concerns are of paramount importance. Only if these issues be addressed correctly that we enjoy the convenience big data give us.

The progress of artificial intelligence (AI) in the medical field has been slow.

Although Apple, Google, Amazon, Facebook, Tesla and Microsoft have provided unprecedented convenience for our daily needs, medical applications are yet to bring about
substantial changes in our life. The lack of “big data” in the field of medicine is the main reason for the lack of AI in this field.

The essence of big data lies in 5 Vs\(^1\), which are as follows.

Volume

The existing medical data is still not sufficiently large. Facebook stores more than 300 PB of data, which is growing at a rate of 600 TB a day (1 Petabyte (PB) = 1024 TB). Google’s database is around 10–15 EB (1 Exabyte (EB) = 1024 PB). On the contrary, the size of the National Health Service database in the United Kingdom or Taiwan’s National Health Insurance Database is at most a few TB. Generally, quantitative changes will only result in qualitative changes when data volume expands to the population\(^2\). This suggests that having sufficient data is important for machine learning and AI.

Velocity

Velocity considers the rate of growth and timeliness of data. When we make an online purchase from Amazon, data based on our purchasing activity is immediately acquired by them and real-time changes are implemented through algorithms. This results in them being able to promote the products that are related to the ones that the user searched. This is because personalized algorithms are based on real-time data acquisition. However, the medical data that we are currently collecting is far behind that of the medical activities.
data in medical databases are only collected after diagnosis and treatment, it is impossible to provide real-time feedback. Owing to these factors, current medical AI can only carry out diagnosis and classification and cannot make decisional adjustments according to the real-time medical scenarios. Google’s database is live and dynamic while the medical database is dead and static. A change in the intrinsic qualities is required.

Variety

Collection of uni-dimensional data also results in a lack of variety. An example is the sentence “the guy who robbed the bank lived near the bank?” Does the second instance of bank refer to a financial institution or the river bank. It is difficult to know the truth from one sentence (uni-dimensional). However, if the entire passage, article, or a short film (multi-dimensional) is included, the intended meaning becomes clearer. The more multi-faceted data is collected, the more likely the facts can be put together. The Danish Civil Registration System currently includes more than 8 million people. The studies focusing on this as a research subject are credible; however, when this database is combined with other dimensional databases, many limitations can be solved and the correlation between diseases will become clearer.

Veracity

It is needless to mention that data veracity plays an important role in implementing AI
technologies. The American College of Surgeons National Surgical Quality Improvement Program® (ACS NSQIP®) collected surgical data from hundreds of hospitals across the US. This data was based on registration data recorded and they were unable to know the quality of the data, how the data was recorded, and the recorder may not include important data, resulting in a loss of veracity. The insurance-based database, Medicare, was also questioned as the recorders tend to record more expensive codes, resulting in a tendency in misinterpreting the correlation between the two events. As mentioned above, combining with other dimensional data can make the data closer to the whole truth. Nobody can be deceived with big data.

Value

Another common misconception is mixing up data and information. Data refers to collected raw materials while information is processed data. The medical database collects information, including the preferences of researchers or recorders. Therefore, many possible variables may be missed out. Considering the flu epidemic, for example, Google did not intend to collect epidemiological data about a flu outbreak; instead, they “accidentally” found that many people were searching about the symptoms from region to region, coinciding with the spread of the disease. In future, the focus will be on the use of portable devices and sensors to collect traces of medical behavior to identify possible improvement directions.
Possible Solutions and Challenges

Throughout history, we have seen revolutionary breakthroughs: The Information Explosion, the Internet of Things, the Big Data, and the Artificial Intelligence. Now is the best time we introduce big data and the internet of things into the establishment of medical databases. First is to expand data. More raw data can be collected through portable devices, sensors, and intelligent hardware, and not intentionally collect processed data. On the other words, shift from the notion of electronic medical records to “automated” medical records.

The second step is to connect the various databases. Medical activities are not independent of the daily lives of people. Similarly, medical databases should not be independent of Google, Facebook, and Apple databases. By connecting databases, medical AI can be truly being intercalated into daily life. However, things are far beyond simple.

The first challenge is inherent to the database itself. The transaction model of electronic medical records (EHRs) must follow the ACID principle (Atomic, Consistent, Isolated, Durable). Relational database is logical, easy to use, rich integrity and avoids redundancy. However, it is very inefficient in massive data reading and writing. Since tens of thousands of units of work occur within seconds, loading capacity of the hardware limits the performance of such database. The BASE model followed by Facebook and Google (Basically Available, Soft-State, Eventually consistent) sacrifices consistency and timeliness and allows database
to expand vertically (Scale up) and horizontally (Scale out). Future medical database may balance its “pH” within such ACID-BASE spectrum.

Secondly, integration can employ a tremendous amount of manpower. One must be able to specifically link individuals across databases; merely combining data together is meaningless. Despite all the technological advances in computing and algorithms, matching of patients across systems is not fully automated. Furthermore, out of commercial concerns, different parts of demographic data can be stored in different databases; one’s health record in the hospital, social network on Facebook, and personal interest on Apple. Asking corporations to cooperate is impractical.

The third concern is the privacy and security issue. The recent scandals of Facebook have shown that data has become a valuable asset of our generation and one should be aware of data predators. Selling of information about friends and "likes" is bad enough; the potential harm with health data can be dramatically larger. Privacy and convenience lie on the two sides on the scale. Data companies should be held accountable for collecting our deidentified health information to improve health care efficiency while protecting the individual data security. Currently, efforts are underway. The Health Insurance Portability and Accountability Act (HIPAA) needs to be renegotiated⁴; Apple safeguarded information via differential privacy; and ultimately, Bitcoin technology might the problem-solver⁵.

Ma Yun of Alibaba once said that he wanted to make doctors become jobless. This does not
mean to say that he wants to develop robots to replace doctors but rather make mankind healthier and happier through artificial intelligence and prediction models. Before we reach there, we must remove the stones under our feet.

References


