Finding the Missing Link for Big Biomedical Data

It has been argued that big data will enable efficiencies and accountability in health care.1,2 However, to date, other industries have been far more successful at obtaining value from large-scale integration and analysis of heterogeneous data sources. What these industries have figured out is that big data becomes transformative when disparate data sets can be linked at the individual person level. In contrast, big biomedical data are scattered across institutions and intentionally isolated to protect patient privacy. Both technical and social challenges to linking these data must be addressed before big biomedical data can have their full influence on health care. It is this linkage challenge that we address in this Viewpoint.

Political campaigns, government, and businesses use big data to learn everything possible about their constituents or customers, and then apply advanced computation to hone strategy. The 2012 Obama campaign identified, approached, and influenced swing voters using data fused from Facebook, census, voter lists, and active outreach. The National Security Agency employs massive data on individuals from phone and Internet companies to identify terrorists. Google personalizes search results with the user's web history and geographic context. In all these examples, the key has been to go beyond aggregate data and link information to individual people. Knowing that there are many swing voters in a zip code is helpful, but contacting those specific individuals may help to win an election.

Linking big data will enable physicians and researchers to test new hypotheses and identify areas of possible intervention. For example, do grocery shopping patterns obtained from stores in various areas predict rates of obesity and type 2 diabetes in public health data bases? Does level of exercise recorded by home monitoring devices correlate with response rates of cholesterol-lowering drugs, as measured by continued refills at the pharmacy? Does increased physical distance from patients' homes to hospitals and pharmacies affect utilization of health care and result in distinct patterns in claims data? To what extent do patients' Facebook friends influence lifestyle choices and compliance with medical treatments? It is unknown whether these types of correlative inferences will really be found in big data and how physicians would use that information. However, being able to link data at the patient level is a prerequisite to exploring the possibilities.3

The first challenge in using big biomedical data effectively is to identify what the potential sources of health care information are and to determine the value of linking these together. The Figure presents a potential way of approaching this problem by organizing data sets along different dimensions of “bigness.” Although some big data, such as electronic health records (EHRs), provide depth by including multiple types of data (eg, images, notes, etc) about individual patient encounters, others, such as claims data provide longitudinality—a view of a patient's medical history over an extended period for a narrow range of categories. Linking data adds value when they help fill in the gaps. With this in mind, it becomes easier to see how nontraditional sources of biomedical data outside of the health care system fit into the picture. Social media, credit card purchases, census records, and numerous other types of data, despite varying degrees of quality, can help assemble a holistic view of a patient, and, in particular, shed light on social and environmental factors that may be influencing health.

The lack of a national unique patient identifier (UPI) in the United States introduces another technical obstacle in linking big biomedical data. However, driven by the absence of a UPI to enable precise linkage, hospitals and clinics have developed sophisticated probabilistic linkage algorithms based on other information, such as demographics.2 Although 2 different patients may share the same name, age, zip code, or other characteristics, by requiring enough variables to match, hospitals and clinics are able to reduce the risk of linkage errors to an acceptable level. An advantage of probabilistic linkage is that the same techniques used to match patients across different EHRs can be extended to data sources outside of health care. However, as indicated in the Figure, some data have fewer variables available for linkage, either because they do not contain the information or because policies restrict their use. These data may still be linked to patients, but the likelihood of errors is greater. Uncertainty due to possible linkage errors may be balanced by the advantages of having access to data about millions of patients. Therefore, future tools that use big data for health research or clinical decision making will need to use statistical techniques that correctly model these trade-offs.

Privacy and security concerns present a social challenge in linking big biomedical data. As more data are linked, they become increasingly more difficult to deidentify.4,5 The consequences of this in health care, particularly for mental health records and genetic markers, have been extensively studied and discussed.6,7 However, given that data linkage is already happening in other industries and is increasingly being thought of as an informational asset for health care delivery, monitoring, and marketing, it would behoove the medical establishment to guide societal and legislative standards in this regard. One constructive response would be to regulate what is legal and ethical, to ensure that benefits outweigh risks, and to include patients in the decision-making process.8 An alternative approach would simply be to put the onus entirely on the patients and give them control over their data. However, as has been seen for far less private data, individuals are likely to share
Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care

**TYPES OF DATA**

- **Medication**
  - OTC medication
  - Medication filled
  - Dose
  - Route
  - NDC
  - RxNorm

- **Demographics**
  - Employee sick days
  - Visit type and time
  - HL7

- **Encounters**
  - Death records
  - SNOMED
  - ICD-9

- **Diagnostics (ordered)**
  - CPT
  - ICD-9
  - LOINC
  - Pathology, histology
  - ECG
  - Radiology

- **Diagnostics (results)**
  - Lab values, vital signs
  - SNPs, arrays

- **Genetics**
  - 23andMe.com
  - SNPs, arrays

- **Social history**
  - Police records
  - Tobacco/alcohol use

- **Family history**
  - Ancestry.com

- **Symptoms**
  - Indirect from OTC purchases
  - HealthMap.org, GIS maps, EPA, phone GPS

- **Lifestyle**
  - Fitness club memberships, grocery store purchases
  - Facebook friends, Twitter hashtags

- **Socioeconomic**
  - Census records, Zillow, LinkedIn

- **Social network**
  - Facebook friends, Twitter hashtags

- **Environment**
  - Climate, weather, public health databases, HealthMap.org, GIS maps, EPA, phone GPS

**STRUCTURED DATA**

- **UNSTRUCTURED DATA**

**Probabilistic linkage to validate existing data or fill in missing data**

**Examples of biomedical data**

- Pharmacy data
- Health care center (electronic health record) data
- Claims data
- Registry or clinical trial data
- Data outside of health care system

**Ability to link data to an individual**

- Easier to link to individuals
- Harder to link to individuals
- Only aggregate data exists

**Data quantity**

- More
- Less

**Observation Identifiers Names and Codes**

- NDC, National Drug Code; OTC, over-the-counter; SNOMED, Systematized Nomenclature of Medicine; SNP, single-nucleotide polymorphism.

**CPT indicates current procedural terminology; ECG, electrocardiography; EPA, US Environmental Protection Agency; GIS, geographic information systems; GPS, global positioning system; HL7, Health Level 7 coding standard; ICD-9, Institutional Classification of Diseases, Ninth Revision; LOINC, Logical Observation Identifiers Names and Codes.**

their data publicly only to regret it later when those data were used in unanticipated circumstances. To avoid paternalism, is there an effective and affordable mechanism, analogous to consent for participation in a trial, to enable patients to decide how and when their data can be shared with or “mashed up” against other databases? It may therefore be timely to convene a public forum whereby the relevant stakeholders, including citizens, the health care community, and commercial data vendors could meet to frame the policy from which legislation and ultimately technical protections for big biomedical data linkage will devolve.

**ARTICLE INFORMATION**

**Published Online:** May 22, 2014. doi:10.1001/jama.2014.4228.

**Conflict of Interest Disclosures:** All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

**REFERENCES**


