Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/survophthal

Major review

Big data and ophthalmic research

Antony Clark, PhD, FRANZCO[a,b,*], Jonathon Q. Ng, PhD, FRANZCO[a,b], Nigel Morlet, FRACS, FRANZCO[a,b], James B. Semmens, MSc, PhD[a,b]

[a] Centre for Population Health Research, Faculty of Health Sciences, Curtin University, Bentley, Western Australia
[b] Eye and Vision Epidemiology Research (EVER) Group, Perth, Western Australia

ARTICLE INFO

Article history:
Received 5 July 2015
Received in revised form 16 January 2016
Accepted 25 January 2016
Available online 1 February 2016

Keywords:
data linkage
clinical registry
health services research
ophthalmic epidemiology
big data

ABSTRACT

Large population-based health administrative databases, clinical registries, and data linkage systems are a rapidly expanding resource for health research. Ophthalmic research has benefited from the use of these databases in expanding the breadth of knowledge in areas such as disease surveillance, disease etiology, health services utilization, and health outcomes. Furthermore, the quantity of data available for research has increased exponentially in recent times, particularly as e-health initiatives come online in health systems across the globe. We review some big data concepts, the databases and data linkage systems used in eye research—including their advantages and limitations, the types of studies previously undertaken, and the future direction for big data in eye research.

1. Introduction

"Big data" is a relatively new concept that describes data so large and complex that it exceeds the storage or computing capacity of most systems ability to perform timely and accurate analyses.[B,244] Health generates huge amounts of data from a wide array of sources such as electronic health records, health insurance claims, and even smart phone health applications. It is the subject of intense interest as industry and researchers realize the potential for extracting value from existing data systems. This big data revolution is being increasingly supported by national governments who have invested significant funds into initiatives designed to develop and capitalize on big data.[A,B]

Health researchers recognized the value in large health administrative databases long before the big data revolution.[267] These databases contain a wealth of information that can now be accessed in a timely and cost-efficient manner owing to advances in computing power and analytical methodologies.[104] These advances have also facilitated data integration processes that afford greater utility over using individual databases for health research, for example, linking pharmaceutical claims data to hospital discharge data allows the study of health outcomes associated with medication use. It promises the potential for more, almost limitless, data available for research.

The last decade has seen a dramatic increase in the use of large databases for ophthalmic research.[33,228] We explore

* Corresponding author: Antony Clark, PhD, FRANZCO, Centre for Population Health Research, Faculty of Health Sciences, Curtin University, Bentley, Western Australia.
E-mail address: a.clark@curtin.edu.au (A. Clark).
0039-6257/$ – see front matter
http://dx.doi.org/10.1016/j.survophthal.2016.01.003

some database concepts, the administrative and linked databases used in ophthalmic research, how they compare with clinical registries, their benefits and limitations, and outline some of the new developments in big data analytics for the future.

2. Information storage: why databases?

The vast amount of information on personal computers is essentially unstructured. Information is typically stored as files within folders that are sorted into a hierarchical structure of menus according to user assigned categories. Although data within a file may have a distinct structure, its extraction and interpretation requires the appropriate software program. Finding information stored this way is analogous to the Librarian's Problem of locating any individual book on a library shelf (essentially the first "big data" issue).

A huge amount of data is also generated from our interactions with computerized systems. These are metadata, machine logs, websites visited, pings to and from a telecommunications tower, and so forth—the "data exhaust" of modern day activity. Some of this information is needed for commercial purposes and kept as part of transactional databases used to generate bills for users. Governments and other organizations have found such information useful in investigating individuals' behavior as they interact with today's world.

Transactional databases are the most commonly used formal computer database and serve a different purpose to analytical databases, which are used for understanding the collected information as broader knowledge (Fig. 1). Transactional and analytical databases are useful for research, but first their information is typically extracted, transformed, and loaded into a specific relational database to enable formal queries—the data warehouse type model.

The advent of the World Wide Web (Internet) saw computer databases evolve into newer network and navigational models. Around that time, object-oriented databases were also developed, principally to deal with information problems that did not readily fit a table and relational model. Information such as documents, images, audio, and video can be stored in object databases along with the methods to retrieve and use those files so all the necessary information is stored fully "encapsulated."

The other main group of modern database models is collectively known as the "NoSQL" databases (not only structured query language). These are not relational in the formal sense and not really tabular in nature as well, so they are truly a modern innovation. Graph databases in this group may be considered a new paradigm in information structuring. They reflect the information structures bought about by the Internet and allow an understanding of information content and its contextual structure. In particular, graph databases depict the type of interactions between content structures, a powerful hitherto not easily achievable task.

Columnar and key-value databases are often used for their speed and ease for computing. Data from primary sources can be quickly extracted, transformed, and loaded into these databases. These databases are often used temporarily during the computing processes to extract, transform, and load information into analytical databases. Their speed is the basis of today's "Big Data" analytics.

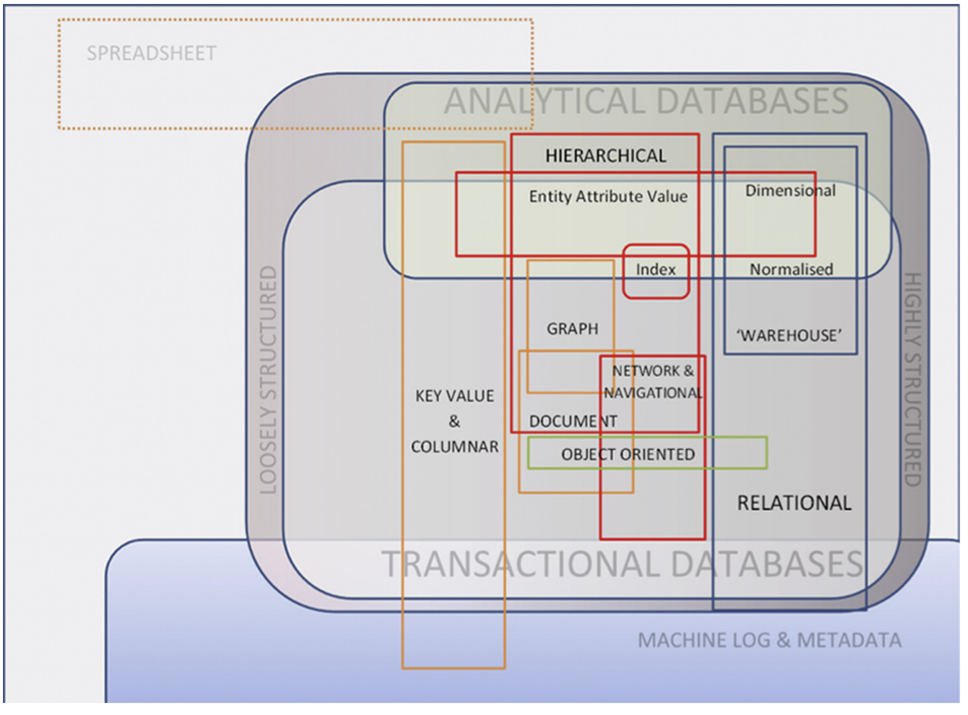

Fig. 1 – Computerized systems for structuring information as databases. The various database models used to store information in computers may overlap depending on the category of use and degree of structure. The type of structure used also means there are distinct relationships between database models.