University of Louisville Journal of Respiratory Infections

PERSPECTIVE PIECE

Considerations for a COVID-19 Research Data Warehouse in the Time of COVID

William A. Mattingly[1]*, *PhD*

[1]Center of Excellence for Research in Infectious Diseases, Division of Infectious Diseases, School of Medicine, University of Louisville, Louisville, KY, USA

*bill.mattingly@louisville.edu

Recommended Citation: Mattingly WA. Considerations for a COVID-19 research data warehouse in the time of COVID. Univ Louisville J Respir Infect **2020**; 4(1):Article 64. doi: 10.18297/jri/vol4/iss1/64.

Introduction

The recent COVID-19 pandemic has created an immediate community need for the benefits of medical research. These include decision support systems for at-risk patients, knowledge of how COVID-19 interacts with comorbidities, and above all vaccine development and research. Clinical research centers around the activities of hypothesis generation, patient recruitment, data collection, statistical analysis, peer review of results, and finally, publication. These activities depend heavily on modern technology to facilitate research at a large scale. Networked computers are necessary for efficient and accurate data collection and sophisticated programming languages are the core of advanced statistical analysis packages. What then, is a data warehouse [1], and how does it fit into the clinical research paradigm?

Approach

As the name suggests, a data warehouse is intended to provide long term storage and easy access to data. For example, suppose an organization is interested in streamlining its operations. It might gather and record data from activities like transactions, salaries, contacts, etc. Once this data enters the data warehouse, it can be accessed to provide reports and statistical summaries. Over time, reports which provide greater operational value become part of the workflow itself, improving the way an organization accomplishes its mission. This creates a loop whereby an organization becomes data driven, constantly improving as it generates and learns from new data.

In the case of a clinical data warehouse, the creation of new knowledge is the process that is being improved. Rather than publication being the ultimate goal of research, study data sets will form the foundation of a data warehouse. Over time, data sets can be standardized and added. This allows the warehouse to grow and ultimately answer many more clinical questions than just the original hypotheses of the data sets.

Data ownership greatly affects how these data warehouses are established. Because owners of the data warehouse control access, contributors need to make sure they are not sharing data that may become inaccessible at a later date. Clear and transparent terms of service and use agreements need to be established early on, to encourage contributors and maintain faith in the platform. Medical facilities and points of care will have the largest number of source records to contribute to a data warehouse, and many hospital systems will have a proprietary data warehouse available for their researchers and affiliates to use. Data ownership is very clear in these cases as the warehouse is part of an organization's other research assets.

While ownership is clear in the case of an institutional data warehouse, such warehouses are limited in their scope, containing data only for patients treated at their facilities. Data that is diverse is usually more generalizable to different situations. Development of an effective data warehouse should include data from more than one source whenever possible, but there are challenges with doing this. In addition to data ownership concerns and negotiation, time must be spent standardizing the data sets which make up the warehouse. If clinical values do not have uniform names, units, and constraints across the warehouse, it will not be useful for answering clinical questions.

Discussion

Starting with a case report form from one of the studies is the best way to develop a standard for the data warehouse. For a COVID-19 data warehouse, using

Table 1. Case report form.

Variable	Description
Record ID	Unique identifier
Demographics	
Age	
Race	
Ethnicity	
Gender	
Lab Values	
Weight	
Height	
BMI	Body mass index
Laboratory and physical findings	
COPD	History of chronic obstructive pulmonary disease
Asthma	History of asthma
CHF	History of congestive heart failure
CAD	History of coronary artery disease
CKD	History of chronic kidney disease
Diabetes	History of diabetes
Liver Disease	History of liver disease
IDP	
Autoimmune disorder	
Immunodeficiency	
HIV	
AIDS	
Cancer/malignancy solid tumor	
Cancer/malignancy hematologic	
Organ transplantation	

a similar respiratory disease study CRF, such as one for pneumonia, helps develop the data points that would be necessary. **Table 1** shows fields that are commonly found in pneumonia and other infectious disease studies. Records need a unique identifier to distinguish from other records, but no identifying information. New records can be considered for addition to the warehouse so long as they have a minimum threshold of completeness for the above variables and are sufficiently de-identified.

Once a starting data set has been established, and the founders have agreed on both the use agreements and a common place for sharing the data, the life of the data warehouse can begin. Although the initial setup can be time consuming and costly, there are legal and technical resources available to help with the process.

Longitudinal Data

A cross-sectional data set only includes information from one point in time. A data warehouse supporting the growth of cross-sectional records is straightforward to design because there is no matching that needs to be done between records. The data can even be stored in a spreadsheet. This is advantageous in many scenarios as data can be exported from the warehouse as a spreadsheet with no advanced processing. But a more effective data warehouse should be capable of holding multiple time points of information related to every record. Recording multiple lab tests, new measurements of weight and height, and the onset of new symptoms provides a more complete view of the progress of a disease. These types of data sets are usually called longitudinal.

The added benefits of storing longitudinal data comes at the cost of some ease of use. Data exports no longer conform to a spreadsheet format, and relational database management is necessary to get any structured information from the data warehouse in the form of a report. Making a cross-sectional database longitudinal involves creating additional tables which hold the list of time dependent information like lab values and vital measurements. Each row in this new table has the same unique identifier, allowing information at that time point to be linked to other record information.

The computer programming for this process is complex and can require significant planning but is necessary to achieve acceptable performance for large data sets. Data queries to a poorly designed relational database with tens of thousands of records and multiple lab measurements per record can take days instead of the seconds we expect. Despite these challenges, an open data warehouse for COVID-19 is essential in preparing for future pandemics and worth the effort involved.

Received: September 14, 2020

Accepted: September 17, 2020

Published: October 15, 2020

Funding Source: The author(s) received no specific funding for this work.

Conflict of Interest: All authors declared no conflict of interest in relation to the main objective of this work.

References

1. Kimball R, Ross M. The data warehouse toolkit: The complete guide to dimensional modeling: Wiley, **2011**.