Journal of Biomedical Informatics 54 (2015) 337–344

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Visual grids for managing data completeness in clinical research datasets

Robert R. Kelley [a], William A. Mattingly [a,*], Timothy L. Wiemken [a], Mohammad Khan [a], Daniel Coats [a], Daniel Curran [a], Julia H. Chariker [b], Julio Ramirez [a]

[a] *Division of Infectious Diseases, Department of Medicine, University of Louisville, Louisville, KY, USA*
[b] *Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY, USA*

ARTICLE INFO

Article history:
Received 12 May 2014
Accepted 1 December 2014
Available online 30 December 2014

Keywords:
Missing data
Data visualization
Clinical trial data
Data completeness

ABSTRACT

Missing data arise in clinical research datasets for reasons ranging from incomplete electronic health records to incorrect trial data collection. This has an adverse effect on analysis performed with the data, but it can also affect the management of a clinical trial itself. We propose two graphical visualization schemes to aid in managing the completeness of a clinical research dataset: the binary completeness grid (BCG) for single patient observation, and the gradient completeness grid (GCG) for an entire dataset. We use these tools to manage three clinical trials. Two are ongoing observational trials, while the other is a cohort study that is complete. The completeness grids revealed unexpected patterns in our data and enabled us to identify records that should have been purged and identify missing follow-up data from sets of observations thought to be complete. Binary and gradient completeness grids provide a rapid, convenient way to visualize missing data in clinical datasets.

1. Introduction

The initial product of a clinical trial is the dataset collected during its execution. Missing data in such a dataset complicates the statistical analysis process. In fact, missing data may preclude certain types of analysis because there is not enough data to analyze; in the extreme case, the entire dataset may not be useful. However, discovering that data are incomplete after the data collection period is often not sufficient, particularly in the case of a study in which follow ups at various time points are collected. Therefore, managing missing data during the execution of a clinical trial is a critical aspect of the clinical trial process.

Data for medical research studies, such as retrospective chart reviews or prospective randomized clinical trials (RCT), are typically collected using paper or electronic case report forms (CRFs) [1]. An investigator, or a designee, examines the electronic health record (EHR) or paper chart for data required by the study, and then transfers these data to a CRF by hand or by retyping it into an electronic system. In the case of a study in which follow ups are required, the process usually involves the investigator contacting the subject by telephone or through a follow up clinical visit to collect further data on their condition and progress.

Study data collection policies and practices and differences in electronic health records systems (EHRs) have the most effect on the quality of data collected for a clinical trial. First, if data are missing or not captured in the EHR, it will also be missing in the CRF and subsequently the study database. Moreover, data required for a clinical trial is often stored as unstructured free text in an EHR; data collection personnel are required to locate these data, which are often located in clinician's notes, which may necessitate a clinician's experience to successfully parse. Furthermore, in a facility still using paper charts, potential legibility problems may interfere with transcribing data to the CRF that, in turn, leads to missing data in the study database.

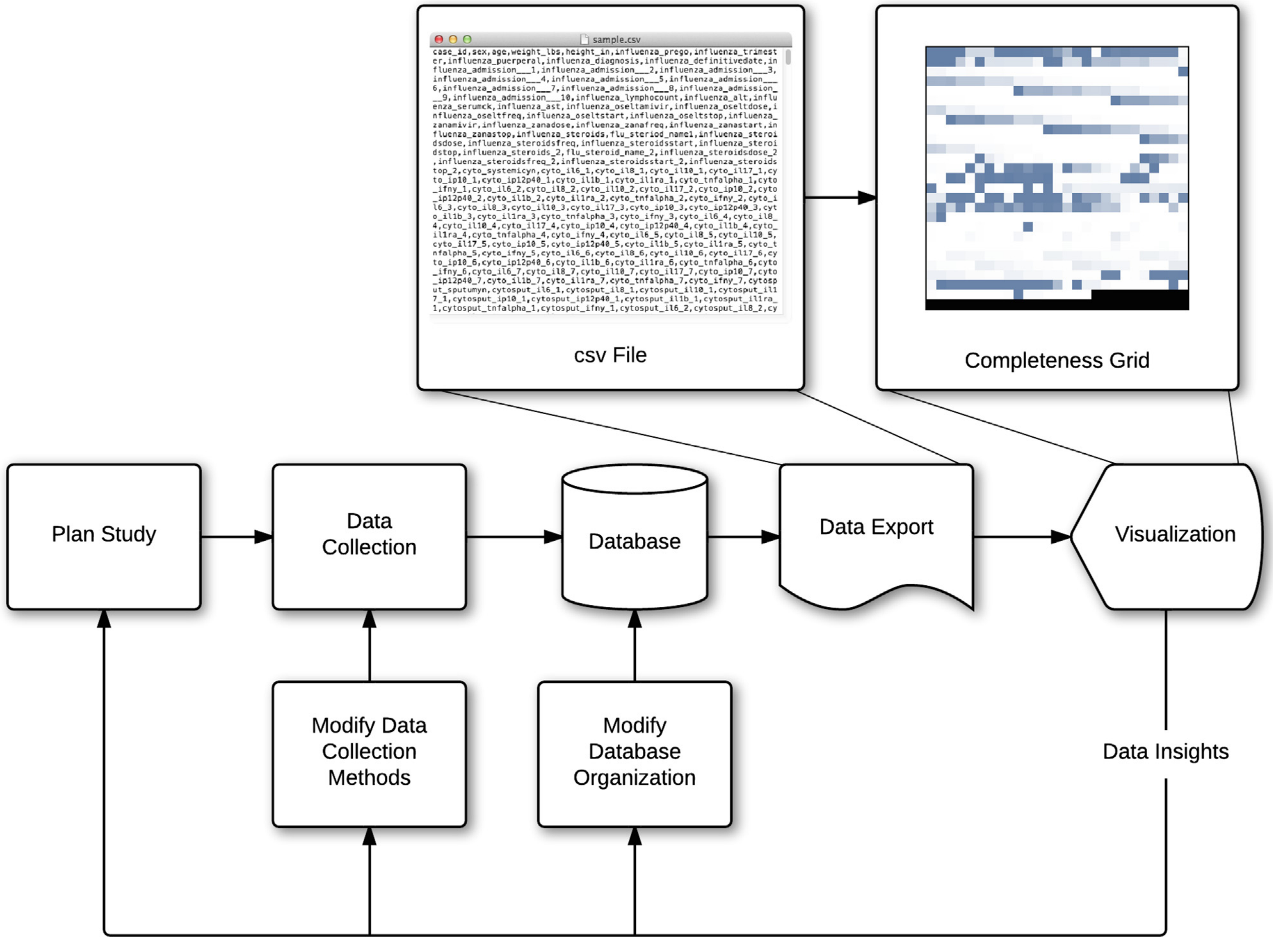
Outside of the EHR, typographical errors and misunderstanding of data entry policy by those entering data may also lead to data quality issues. Several preventative measures can be used to address these problems including: documenting the data collection process (e.g. manual of operations), training data collection staff on proper data collection procedures, performing double-data entry, fostering frequent communication between investigators and data collectors regarding potential data issues, and using monitoring reports to track the state of the dataset [2].

The data quality literature frequently categorizes missing data as "completeness", a characteristic of the Contextual Information

* Corresponding author at: Division of Infectious Diseases, Department of Medicine, University of Louisville, 501 East Broadway, Suite 140B, Louisville, KY 40202, USA. Fax: +1 502 852 1555.

E-mail address: wamatt02@louisville.edu (W.A. Mattingly).

http://dx.doi.org/10.1016/j.jbi.2014.12.002
1532-0464/© 2015 Elsevier Inc. All rights reserved.



```
case_id,sex,age,weight_lbs,height_in,influenza_prego,influenza_trimest
er,influenza_puerperal,influenza_diagnosis,influenza_definitivedate,in
fluenza_admission___1,influenza_admission___2,influenza_admission___3,
influenza_admission___4,influenza_admission___5,influenza_admission___
6,influenza_admission___7,influenza_admission___8,influenza_admission_
__9,influenza_admission___10,influenza_lymphocount,influenza_alt,influ
enza_serumck,influenza_ast,influenza_oseltamivir,influenza_oseltdose,i
nfluenza_oseltfreq,influenza_oseltstart,influenza_oseltstop,influenza_
zanamivir,influenza_zanadose,influenza_zanafreq,influenza_zanastart,in
fluenza_zanastop,influenza_steroids,flu_steriod_name1,influenza_steroi
dsdose,influenza_steroidsfreq,influenza_steroidsstart,influenza_steroi
dstop,influenza_steroids_2,flu_steroid_name_2,influenza_steroidsdose_2
,influenza_steroidsfreq_2,influenza_steroidsstart_2,influenza_steroids
top_2,cyto_systemicyn,cyto_il6_1,cyto_il8_1,cyto_il10_1,cyto_il17_1,cy
to_ip10_1,cyto_ip12p40_1,cyto_il1b_1,cyto_il1ra_1,cyto_tnfalpha_1,cyto
_ifny_1,cyto_il6_2,cyto_il8_2,cyto_il10_2,cyto_il17_2,cyto_ip10_2,cyto
_ip12p40_2,cyto_il1b_2,cyto_il1ra_2,cyto_tnfalpha_2,cyto_ifny_2,cyto_i
l6_3,cyto_il8_3,cyto_il10_3,cyto_il17_3,cyto_ip10_3,cyto_ip12p40_3,cyt
o_il1b_3,cyto_il1ra_3,cyto_tnfalpha_3,cyto_ifny_3,cyto_il6_4,cyto_il8_
4,cyto_il10_4,cyto_il17_4,cyto_ip10_4,cyto_ip12p40_4,cyto_il1b_4,cyto_
il1ra_4,cyto_tnfalpha_4,cyto_ifny_4,cyto_il6_5,cyto_il8_5,cyto_il10_5,
cyto_il17_5,cyto_ip10_5,cyto_ip12p40_5,cyto_il1b_5,cyto_il1ra_5,cyto_t
nfalpha_5,cyto_ifny_5,cyto_il6_6,cyto_il8_6,cyto_il10_6,cyto_il17_6,cy
to_ip10_6,cyto_ip12p40_6,cyto_il1b_6,cyto_il1ra_6,cyto_tnfalpha_6,cyto
_ifny_6,cyto_il6_7,cyto_il8_7,cyto_il10_7,cyto_il17_7,cyto_ip10_7,cyto
_ip12p40_7,cyto_il1b_7,cyto_il1ra_7,cyto_tnfalpha_7,cyto_ifny_7,cytosp
ut_sputumyn,cytosput_il6_1,cytosput_il8_1,cytosput_il10_1,cytosput_il1
7_1,cytosput_ip10_1,cytosput_ip12p40_1,cytosput_il1b_1,cytosput_il1ra_
1,cytosput_tnfalpha_1,cytosput_ifny_1,cytosput_il6_2,cytosput_il8_2,cy
```

Fig. 1. Clinical study workflow using completeness grid for visualization. After data has been collected, completeness grids can be created which inform the processes of plan study, data collection, and database organization.

Quality dimension [3–6]. Completeness is specifically defined as "the extent to which data are not missing" [7]. The statistical analysis literature is replete with warnings about the effects of missing data on research results, as well as approaches to mitigate the problems it causes [8–11]. The two most significant issues with missing data are that they can introduce bias into the statistical results and they reduce the power of the study. Common techniques for handling missing data include complete case analysis, analysis of observations available, using data imputation to "fill-in" data with statistical methods, and intent to treat [12].

Missing data can be more effectively recognized using data visualization techniques. Data visualization is a means of representing data in a graphical form allowing the recognition of details and patterns that might otherwise remain obscure. Ways to represent data while taking into account missing or incomplete datasets is an active area of research in data visualization and data quality. Many visualization methods for missing data depend on knowing the uncertainty associated with data [13]. Different types of visual attributes, like color hue and texture can be used in a dataset visualization to draw attention to uncertainty values [14,15]. However, in the case of many sparse datasets, missing data is the norm, and estimation and imputation should be avoided in favor of alerting the viewer as to which data fields are empty [16].

There are also many statistical software packages for visualizing and interacting with missing data. Missing Are Now Equally Treated (MANET), XGobi, GGobi and Mondrian, were all designed to allow users to explore and visualize data and missing data [17–19]. The R programming language contains many packages and functions for data analysis and visualization, and the VIM package has been designed specifically to facilitate the analysis and interactive exploration of missing data [20].

One drawback of previous missing data visualizations is their interdependence with complex data analysis. Many stakeholders in the clinical data collection process are interested in quickly and easily viewing the completeness of a dataset, but do not have the background to engage in complex analysis. To facilitate this we have developed two visualization methods, the binary completeness grid and the gradient completeness grid, tools for clinical trial data managers to monitor and evaluate missing data and effectively visualize large two-dimensional datasets like those collected for clinical trials.

Fig. 1 shows a typical workflow for using completeness grids. As data is collected for a study and entered into a database, the entire dataset or subsets can be exported in an easy-to-process format such as a comma-separated value (csv) file. Completeness grids can then be generated providing a visual representation of the completeness of the dataset. Using this information, data monitors can investigate unexpected patterns and resolve potential problems with the database or data collection methodology.

2. Materials and methods

2.1. Completeness grids

Completeness grids account for the visualization needs of individual users in a clinical study. Investigators are generally involved