The Practice of Informatics

Technical Brief ■

A Data Accounting System for Clinical Investigators

T. MICHAEL KASHNER, PhD, JD, MPH, ROBERT HINSON, MA, GLORIA J. HOLLAND, PhD, DON D. MICKEY, PhD, KEITH HOFFMAN, LISA LIND, PhD, LINDA D. JOHNSON, PhD, RN, BARBARA K. CHANG, MD, MA, RICHARD M. GOLDEN, PhD, STEVEN S. HENLEY, MS

Abstract Clinical investigators often preprocess, process, and analyze their data without benefit of formally organized research centers to oversee data management. This article outlines a practical three-file structure to help guide these investigators track and document their data through processing and analyses. The proposed process can be implemented without additional training or specialized software. Thus, it is particularly well suited for research projects with small budgets or limited access to viable research/data coordinating centers.

■ J Am Med Inform Assoc. 2007;14:394–396. DOI 10.1197/jamia.M2218.

Introduction

Investigators conducting clinical trials often produce small, defined research databases that are complex and contain sensitive information. Unlike large on-line health information systems, a research database is governed by an approved scientific protocol, is self-contained, is time-limited, and is accessible only to a small number of individuals. Much has been written about clinical trial data quality¹, management,² and capture³ that typify coordinating centers where research data are efficiently managed. In practice, however, many clinical investigators in private offices or at academic institutions often attempt studies without support from qualified biomedical informatics consultants or access to viable data coordinating centers. This article briefly describes a Data Accounting System designed to help these investigators account for their research files during processing and analyses. Our model is based on the practices of the Office of Academic Affiliations within the Veterans Health Administration.

By exercising greater accountability over their research data, we believe clinical investigators will: (1) enhance human subject privacy and data security, (2) extend data understanding during analyses and report writing, and (3) pro-

vide paper and electronic audit trails that improve file handling and documentation. Our experience suggests that a data accounting system will help decrease downtime during staff turnover and facilitate appropriate data sharing among investigators. These objectives can be achieved without significantly higher administrative costs or additional software.

Data Accounting System (DAS)

The DAS is a set of file structures designed to help clinical investigators track the processing and analyses of their data. The trail extends from raw collected files through written reports. The file types include original raw files (R-files) that are preprocessed into permanent research-ready files (RR-files). The latter are then processed into temporary analytic files (A-files) that are subject to statistical analyses with software that creates final tables and graphs. All processing is accomplished through written programs stored for review and future use. Field names, value codes, data structures, and programs are also written using formats and naming conventions that are agreed on by the investigators.

R-files

Most research databases begin as downloaded electronic files, scanned forms, abstracted records, keyed text, or information entered by respondents through Web-based applications. These data are collected and uploaded into a common statistical package and stored in its original context as R-files. The purpose of R-files is to validate, document, and monitor data collection activities. We recommend that field names and response codes point directly to the original data capture form. This allows investigators easy access to compare values entered into the data file with the original source materials. For example, a raw field may be named for the number of the corresponding response box on a web-based form or data collection sheet. Pointing to the precise location where data were obtained helps facilitate comparing double entries, allows spot checking for data accuracy,

Affiliations of the authors: Office of Academic Affiliations, Department of Veterans Affairs, (TMK, RH, BKC, DDM, GJH, LDJ) Washington, DC; Allocation Resource Center, Department of Veterans Affairs, Braintree, MA (KH); Private Practice, (LL) Dallas, TX; University of Texas Southwestern Medical Center at Dallas, (TMK) Dallas, TX; University of New Mexico School of Medicine, (BKC) Albuquerque, NM; University of Texas at Dallas, (RMG) Dallas, TX; Martingale Research Corporation, (SSH) Plano, TX.

Correspondence and reprints: T. Michael Kashner, PhD, JD, MPH, Department of Clinical Sciences, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75390-9066; e-mail: <michael.kashner@va.gov>.

Received for review: 7/19/2006; accepted for publication: 4/10/2007.

and informs investigators of exactly where the information was recorded.

Data dictionaries and code books that list field names, locations, descriptions, and definitions for response codes are included in the documentation, along with descriptive statistics (e.g., counts, means, and frequencies) that help investigators verify data content. Data collectors meet with clinical investigators after data collection has been completed to review, certify, and formally close R-files. Data collection is always governed by institutionally approved protocols that set limits on what, whom, and when information can be collected. Once closed, R-files are locked, encrypted, and password protected to ensure data integrity for long-term storage. We recommend that R-files be placed off-line on encrypted and password-protected disks that are stored in locked cabinets located in secured areas. They also should be stored in secondary locations for backup. Access is granted only by the principal investigator or an authorized agent, by following a formal and written procedure. Access should be limited to creating and updating RR-files as described below. Access also may be required for administrative oversight, such as data security and review of human subjects. Locking R-files to preserve original data integrity is necessary because: (1) the data collection team is often disbanded or inaccessible, (2) investigators are unblinded as to subject protocol assignments, and (3) investigators have begun discovering research findings.

R-files are sometimes created from data sources in which the investigators may have time-limited access to specific fields that can help identify patients (e.g., subject date of birth). Such fields are sometimes necessary for study purposes (compute patient age, merge data files). To account for these time-limited fields, two sets of R-files are created: one containing all the data (R-all fields), and a second containing all data except fields set to be purged after the agreed time limits have expired (R-purged fields). At the appointed date, the R-all files are deleted and the R-purged files are left open for further use.

RR-files

R-files are preprocessed into RR-files based on written computer scripts called Research Ready programs (RR-programs). Preprocessing includes the usual data scrubbing, missing value recoding, data formatting, and variable labeling, as well as instrument scoring, index computations, and data merges. To ensure uniform preprocessing across users, analysts are allowed access only to RR-files. This procedure helps investigators enforce data use agreements by retiring R-all files while allowing analyst-continued access to RR-files. In terms of data security, R-files contain all identified data needed for preprocessing. The preprocessed RR-files (containing subject age, merged data) are trimmed of any unnecessary identifying information (e.g., patient date of birth) before it is disseminated to analysts to compute study findings.

Preprocessing under written and retained RR-programs can be invaluable in helping investigators understand how each research field was computed from original source materials. We recommend that RR-programs be included as part of the research documentation. Thus, each RR-program has a title, date prepared or modified, name of programmer, and

written code that follows a logical order with explanatory notes and title headings. A quick listing of all RR fields is placed at the beginning for reference. When scoring specific instruments, the program should cite applicable references to the literature, including applicable version numbers. RR-files are formatted to enhance viewing displays (formatting "1" or "0" instead of 1.00 or 0.00 to indicate "yes" or "no"). This strategy allows clinical investigators to see how RR-fields were constructed from raw variables.

Unlike closed R-files, RR-files may be changed to reflect additional demands for the original R-files. That is, as investigators gain experience during analyses, RR-files can be expanded to correct data errors discovered during analyses, compute alternative scoring algorithms, and construct additional variables. Such modifications, however, are accomplished by a designated programmer who accesses R-files under investigator oversight to make the preapproved changes. Creating new versions of the RR-files should be considered additional data preprocessing and not data analyses. Thus, investigators must re-review and recertify these updates, disseminate to data analysts, and retire older versions. New RR-files should preserve prior variables, if appropriate, for continuity of data analyses.

Postcollection corrections to data are rare, but do occur. Instead of correcting the closed R-files, postcollection corrections are handled by modifying the RR-program and regenerating RR-files. Such corrections are reversible and documented (RR-program), and the impacts are computable. This special attention recognizes that the original data collection team may be unavailable for review, investigators are now unblinded, and data analyses have begun generating study results.

New RR-files (with a version number and creation date) should be disseminated to data analysts. For purposes of security, uniformity of preprocessing, and accuracy, investigators must ensure that only study-approved RR-files are used to produce study findings and results. This is accomplished by limiting access to R-files to a designated study programmer for the sole purpose of regenerating RR-files. This is done under the direction and oversight of an executive committee representing study investigators.

R-fields are named to point to the location on data capture instruments. RR-fields are named to reflect data content. Because clinical research files stand alone, investigators should consider their programming and representational needs when naming RR-fields. We believe that properly designed field names simplify programming code, reduce errors, and facilitate data understanding. Maintaining the same naming rules within an investigator group will also facilitate programming across different studies and reduce downtime caused by staff turnover.

Although conventions do exist, we suggest at a minimum that variables be named so that those with similar functions are listed together when arranged alphabetically. For example, the first letter denotes class (D for demographic, O for outcome evaluation, M for health service use), and the next three denote type (DAGE is age, DEDU is education), and subsequent positions denote subtypes (DEDUyr indicates years of formal education, DEDUhs indicates yes/no if high school graduate). There are cases

Table 1 ■ Data Accounting System (DAS) Files for Study Databases

File Category	Purpose	Actions	File Characterization
Raw files (R-files)	Collection	Data entered, validated, and documented	Certified, locked after data collection, limited access, fields named to reflect data capture
Research ready files (RR-files)	Dissemination	Created from R-files with RR-programs, under investigator supervision	Long-term, secured but shared among analysts, controlled by investigators, fields named to reflect data content
Analytic files (A-files)	Analyses	Created from RR-files with A-programs	Temporary, task specific, controlled by data analysts, fields named at discretion of each analyst

in which analysts need personal identifying data. To indicate a field containing personal information, we recommend using the initial letter P (e.g., subject name, PNAM; street address, PADDst). These P-variables require special handling and come under the review by local institutional review boards. With this format, protected fields easily can be identified. To protect patient privacy, we recommend collating all P-fields into separate files that are linked to subjects in all other RR files by an arbitrary number (PID). In this way, user access to files containing P-fields can be restricted without limiting access to other de-identified data. Ontological conventions are available that can help clinical investigators systematically name fields according to national and international standards.

In addition, file structures are selected to offer clear displays of the data. For instance, we recommend that repeated assessments of the same patient at different times appear as new records rather than as new fields to the same patient-level record. Again, RR-files are structured to service functions of storage, display, transfer, and programming statistical analyses.

A-files

RR-files are processed into temporary A-files that are used directly for analyses. Such computations may involve RR-file merges, recodes, transformations, and indices specific to a given analysis. Such A-programs first begin by processing RR-files to create these A-files. Next, the A-files are analyzed to output tables and graphs. Analytic programs are sequentially numbered and dated for quick reference and retrieval.

Analytic files may be recreated by rerunning the saved program. Reapplying saved A-programs on corrected RR-files will also update outputs. This strategy permits investigator oversight while allowing analysts flexibility to work with the data without changing the original RR-files. Certain variables repeatedly computed in analytic files may be elevated to RR-status by formally modifying the RR-program (described above).

Summary

Clinical investigators may better manage studies by effective use of a DAS that divides research data into three file types. Summarized in Table 1, these types represent outputs from: (1) collecting, (2) disseminating, and (3) analyzing research data. We believe that the use of these file types will enhance investigator oversight, documentation, and data security without requiring changes in the choice of statistical software or adding significant investigator burden or administrative costs. As such, it offers an appealing methodology for investigators facing small budgets with limited access to viable data management centers.

References ■

1. Fleiss JL. Design and Analysis of Clinical Experiments. New York: John Wiley, 1999.
2. Knatterud GL, Rockhold FW, George SL, et al. Guidelines for quality assurance in multicenter trials: a position paper. *Control Clin Trials* 1998;19:477-93.
3. Marks RG. Validating electronic source data in clinical trials. *Control Clin Trials* 2004;25:437-46.