

Written and researched by:
Dr. Kirk Mousley
and Karl Mousley

In the next issue of
EDC Today:

Data Marts

About EDC Management:

EDC Management is the leader in Clinical and Data Management and Electronic Data Capture (EDC) consulting services for the biopharmaceutical industry. EDC Management publishes well-researched and timely information about Electronic Data Capture technologies and processes through *EDC Today*® and *EDC In Depth*. We do not sell or endorse any specific EDC software application or vendor. Improve process today; position for tomorrow.

EDC Management

P.O. Box 384
Conshohocken, PA 19428
484-530-0300 (voice)
610-567-0357 (fax)
info@edcmanagement.com
www.edcmanagement.com

Edit Checks

EDC Today is an independent publication on current information and issues in Electronic Clinical Systems (ECS) strategies and technologies for the Biotechnology and Pharmaceutical (Biopharma) industry. Each month we examine topics related to ECS theory, technology, practice, or implementation.

Recently EDC Management was asked, “Just how many edit checks do I need to be sure my data are clean?” and “Why can it take so long to get edit checks written?” We were also asked to comment on how the process of specifying, writing, and testing edit checks might be streamlined and made more efficient. This was no simple question!

In this issue, we discuss edit checks and focus on the nature of clinical data errors, how some of these errors might be avoided, the nature of edit checks, and how a mix of considerations might help you know how many edit checks you need to be sure your data is clean enough and how to maximize the return on your investment in edit check software.

Introduction

Edit checks are usually small programs that are written to confirm that the clinical data residing in a database are reasonable. These edit checks represent a significant effort in the set up of a data management system used for collection and cleaning of clinical trial data. An understanding of the nature of data errors in the clinical trial data can assist in the specification and programming of these checks.

There are a couple fundamental purposes for edit checks. One is to ensure that clinical trials data captured during a study conforms to regulatory standards for “quality”. Another is to ensure the data is sufficiently “clean” for meaningful analysis.

In this issue of *EDC Today*, we discuss the sources of data errors, the cost associated with these errors, and how to avoid errors. We then describe how to determine how many edit checks are needed, the different types of edit checks, and how to balance the effort of programming the edit checks with alternative ways of searching for data errors such as listings and charts.

(continued on page 2)



EDC Management believes that Electronic Data Capture (EDC) and Clinical Data Management Systems (CDMS) software should provide native support for most of the straightforward edit checks that a data management department needs to perform. EDC Management believes that a large percentage of edit checks should be implemented as part of specifying the metadata for the protocol in the data management system.

Sources of Errors

In any discussion on edit checks, it is important to be cognizant of the sources of error they are designed to uncover. In broad terms, errors can be classified as being an inappropriate response to a clinical trial question or as a transcription error.

The first category of errors includes those due to misunderstood or ambiguous Clinical Protocol requirement/specifications. These errors might inadvertently cause Protocol Violations (PV). Misunderstood or ambiguous Case Report Form (CRF) instructions, prompts, labels, scheduling, and other elements can also lead to inappropriate response errors. Furthermore, these errors may be “systematic” in nature (i.e., the error is repeated for each/most occurrences of the misunderstood element).

The second category of errors includes transcription errors. These errors can occur during Source Document Verification (SDV), the process of comparing clinical data at the source to that transcribed onto a CRF (or eCRF). Data entry, the process of entering clinical trials data into a data repository, can also introduce transcription errors. Unfortunately, from beginning to end, the “data capture” process for a clinical trial can be a complex one, with the potential for introducing errors at every step.

Errors are Costly

The cost of discrepancy processing can be several hundred dollars per discrepancy. With traditional paper CRF processes, it is not uncommon to encounter a discrepancy on every page of a completed CRF. Even with an Electronic Data Capture (EDC) system, discrepancies will occur at a significant rate – however, these errors should be due less to “inappropriate response” and due more to clinical data complexity (that is, errors uncovered by cross module and cross page checks).

In order to determine the “cost” of an error it is necessary to categorize what it took to get “resolution”. One way to do this is by classifying resolutions as shown in Table 1. If nothing else, CDAs should be alert to errors that are both systematic in nature and expensive to resolve.

Table 1. Potential Resolution Classifications

No action taken beyond review of the data (e.g., the discrepancy itself is an error)
An entry error correctable in-house
An error corrected by editing ground rule (e.g., a missing item that can be safely assumed to be a certain value)
A correction made in-house but requiring notifying the investigator site
An error that requires clarification of the part of the investigator (e.g., a Data Clarification Form)
A “Protocol Violation” (i.e., a situation that involves removing a subject from the trial)

(continued on page 3)



In an article published by the Drug Information Association, Kenneth Buchholz and Diane Ascoli of Amgen Inc. state that they encountered 65,000 discrepancies while conducting a 2760 patient clinical trial using a 57 page CRF.¹ This works out to a discrepancy per every 2.4 CRF pages. They go on to state that they required nearly 22,000 hours of in-house staff time to resolve these errors. Obviously discrepancies are an expensive obstacle. Programming edit checks may also appear to be an expensive endeavor, about 300 staff hours are needed to program the edit checks for a typical phase III. However, this cost could be viewed as a bargain compared to the cost of resolving the errors they uncover.

Avoiding Errors

It should be cost effective to expend some effort (and money) in attempting to avoid as many errors as possible. Reducing the number of times an item is transcribed is almost the only way to reduce transcription errors (in some cases errors can be reduced by better formatting of a form). EDC offers what seems to be an easy way to reduce transcription errors by nearly half since it does away with one of the transcription phases (i.e., the transcription from source document to CRF) found in the traditional process. Offsetting this improvement to some degree, however, is the fact that EDC exchanges trained data entry professionals for untrained part time investigator staff.

Reducing the “inappropriate response” errors (those errors caused by misunderstanding the Protocol or the data entry form) can often be accomplished with better writing (i.e., clarified prose in Protocols), better forms (i.e., more understandable instructions, better field labels, improved field checks, error messages that explain the nature of entry errors in a clear way), and improved training in a number of areas (for example, professional writing, communicating, form designing, expressing clinical trial goals and procedures).

Note: a Clinical Research Associate (CRA) or medical monitor can do “reasonability checks” while performing Source Document Verification (SDV) but shouldn’t be depended upon to do a complete job of it since SDV is difficult enough to do on its own. Furthermore, we feel a Quality Assurance (QA) reviewer should not need to perform reasonability checks at all.

Stopping Errors

In order to prevent errors during an ongoing clinical trial, it is also important to identify and stop systematic errors as soon as possible. In order to do this, errors will need to be categorized by source (e.g., Clinical Protocol ambiguity, CRF page design or layout, Investigator Site Personnel issues, Data Entry staff, and more sensitively, specific Sites and Data Entry staffers) and closely monitored. Preventing errors in future clinical trials involves holding post-mortem meetings that identify sources of error and outline steps that will be taken to avoid them (e.g., usually this will focus on CRF design but should also include review of instructions and training methodologies). These post-mortem meetings are rarely held due to perceived professional staff time constraints but if one hears Clinical Data Associates (CDAs) talking about CRF pages that are “Data Clarification Form (DCF) generators”, a red warning light (and accompanying siren) should go off amongst their CDA management! Table 2 summarizes procedures that will help prevent errors.

(continued on page 4)



Table 2. Preventing Errors

Write and edit Protocols to reduce ambiguity!
Design and redesign CRF pages to reduce discrepancies!
Enhance training of site personnel to reduce discrepancies!
Design and redesign edit checks to reduce false discrepancies!
Develop and use software that categorizes sources of error!

How Many Edit Checks Are Needed?

In deciding how many edit checks are “required” it is necessary to consider how many of the items collected for the clinical trial are pivotal to successful analysis or early detection of subjects that are “protocol violators”. On this basis, many Biopharma Clinical Data Management (CDM) groups lack sufficient edit-checks or have way too many of them!

Interestingly, items that are critical (i.e., “must have”) to analysis are often not “required” by the data entry system and no edit check exists to uncover them if the item is missing. Sometimes, instances that are obviously violations of the protocol are not specified at all (e.g., a standard weight item might be used where the range check associated with it is not restrictive enough for a specific clinical trial).

On the opposite extreme, some CDM groups specify edit checks that are either redundant, meaningless, or cover cases that cannot happen. Edit checks are an investment; it is necessary to use standards to maximize their return on your investment.

You can determine when your data are “clean” enough by:

- Knowing the criticality of an item or data point to analysis helps determine the edit-checking effort level.
- Reading and understanding the clinical protocol. Clinical groups would benefit by having a meeting to discuss the new Protocol, its research intentions, and how Clinical Research intends to execute the trial. This meeting should include both CDM and CDM-oriented Information Technology staff members.
- Defining your edit check philosophy. Will you attempt to flag any data that looks suspicious for review or only data that is almost certainly in error?

As mentioned previously, the use of an EDC application can reduce errors that occur during the course of a clinical trial. Avoiding transcription errors is one definite benefit of EDC. Another benefit lies in making use of edit checks as data are entered. These edit checks are commonly known as field-level checks or front-end checks. Carefully designed front-end checks will both prevent inappropriate responses and assist the data entry person by explaining, in greater detail than that provided on the eCRF, not only what data is expected but also what the proper data format should be. These checks, by their nature, tend to be limited to a field (i.e., item or question) or eCRF page. Checks of a more involved nature, that is, those that span fields and/or eCRF pages, are usually programmed to execute against the more complete, as of the time of execution, CDMS database. These checks tend to be more difficult to specify and to write and often have an underlying “clinical assessment” that needs to be codified (e.g., a laboratory test result that changes more than 3 percent from the previously collected value).

(continued on page 5)



In summary, there should be an edit check to ensure reasonability for every item that is critical to the analysis of the data collected for a trial; an edit check for every newly defined item in the clinical management system, and an edit check for every item known to be troublesome (but hasn't been corrected for whatever reason).

The Old Economic Maxim

There is an old economic maxim called the 80-20 rule, which is rooted in what is known as the Pareto Principle, that states eighty percent of output are the result of twenty percent of the effort.² Such is the case with edit checks in two ways. One, 80% (or more) of all the edit checks specified and coded for a clinical trial are pretty basic and simple to write and develop, leaving 20% (or less) of all the edit checks specified and coded that are difficult to specify, write, and test. Furthermore, it might be noted that 20% of the edit checks uncover 80% of the discrepancies encountered during a clinical trial. What the principle doesn't tell us, is which 20% of the edit checks will be hard to develop and which 20% of the edit checks will uncover most of the errors in the data.

The point to be made here is that an EDC or CDMS application should natively (and actively) support the writing of an edit check by meta-data definition for most, if not all instances of "easy to write edit checks". An EDC and/or CDMS vendor should note that there are few escalating levels of edit checking, and fundamentally, very few different types of edit checks, as shown in Table 3. Sponsors should continue to press EDC and CDMS vendors to improve the "native" support for edit checking and insist on being able to perform many if not most edit checks by specifying meta-data and not by scripting. At minimum a vendor that understands the Bioparma industry will provide well documented, generic but truly useful, supporting procedures that can be called from within their edit check scripting environment.

EDC Management suggests that there are only seven (7) types of edit checks while admitting the last type is a "catch-all" in nature.

1. Item is missing (or not missing)
2. Item is not in range
3. Item is not in "codelist"
4. Item is not in dictionary
5. Item value does not properly relate to another item value (e.g., checks involving LT, GT, EQ, "before", "after")
6. Other, specify (e.g., Sex is "Male" so no pregnancy test results should be entered, if "Other" is checked, please specify)
7. One or more combinations of the above (e.g., BP systolic and diastolic values checked against range and each other)

(continued on page 6)



Table 3. Levels of Edit Checks

Edit Check Level	Examples
Within the item edit check	A required item, a range check, and stretching things a little, codelist or dictionary lookup.
Within the record single item check	If this item is a specific value, some other item should be blank.
Within the record multiple item check	If this item is a specific value, and that item is some other specific value, then some other item on this record should not be missing.
Within the table cross record single item check	If this item is present, this same item in the next visit should be less than 3% larger.
Within the table cross record multiple item check	If this item is present, some other item in this table should be collected (not missing) in the next visit.
Cross table single item check	This item is “checked”, so a record should, or should not, exist in another table (e.g., A “No Adverse Events (AEs)” checkbox is selected, so there should be no AE records).
Cross table multiple item check	This date should be after that date in another table.

The Hard Checks

Another more advanced example of type 7 edit checks involve data calculated by complicated formulas such as body mass, Quality of Life (QOL) scores, tumor assessment, and clinical assessments such as endpoints. Often, results of the calculation(s) are not databased for valid reasons like “normalization” of data, but the values not being databased may lead to a fairly difficult to write and test edit check that in essence has to derive the data in addition to performing the actual check.

Another class of edit checks that tend to be difficult to program are those where there are no clear-cut “link(s)” between the data. This is often the case where dates need to follow a certain sequence (e.g., patient diary) or records of an Ongoing or Continuing nature (e.g., AEs with no end date because the event is continuing and will need to be linked to a subsequent visit or CRF page).

Example:

An AE record has a “Treated with RX? (yes/no)” question. If the answer is yes, the edit check needs to determine if there is a corresponding record in the MED table. Unfortunately there is usually no link between the AE and MED records so it is difficult to determine which (if any) of the MED records might be “the” Rx that was given to treat this particular AE.

(continued on page 7)



No One Can See The Troubles I See

Clinical review of listings may be more cost effective than developing difficult edit checks. It is easy to underestimate the power of a simple listing especially since massive amounts of computer processing power has become so prevalent. However, carefully designed listings can be just as or more effective than programming edit checks - e.g., a cross table listing of showing concomitant medications and AE verbatim terms in the cases where the AE is resolved by “treatment”. The human being is a marvelous pattern checking/misalignment discoverer.

Example: Diary Data Listing

<u>SEQ</u>	<u>DATE</u>	
1	10/01	As you can readily see, record 3 has the day and month transposed. These errors can be caught very easily by the eye when reviewing a simple listing.
2	10/02	
3	03/10	
4	10/04	

More modern visualization tools are available that will present data in a number of multidimensional charts that can also be used to visually detect “unreasonable” data. One such tool is IRIS Explorer produced by the Numerical Algorithms Group.

Conclusions

Data discrepancies are expensive and consume lots of clinical trial staff time. Cost effective, data cleaning is the result of taking a number of considered approaches.

Minimizing or avoiding “inappropriate response” errors by improved writing, communicating, and form design is one approach. Identifying and stopping systematic errors is another. Reducing transcription errors by using EDC is a third way to avoid errors.

Improved training of all clinical workers involved in a trial will help reduce errors. Focusing on what data is pivotal and essential to analysis can maximize the return on the edit check investment. Holding post-mortem meetings to identify the perceived (and measured) causes of errors found during a trial and using this information to re-write, re-design, and improve protocols, CRFs, and edit checks can be of significant value.

Understanding the nature of edit checks, both front end and back end, and their limitations can lead to more effective data cleaning. EDC and CDMS vendors should be encouraged to refine and improve the edit checking tool(s) they offer. Getting a clean database is a dirty job but someone has to do it!

Readers are encouraged to share recommendations that they may have for how vendors can improve their products. We would also like to encourage feedback and suggestions to this issue, and welcome suggestions of topics for future issues.

References

¹ Kenneth Buchholz, PhD and Diane Ascoli, “The Value of Computer-Assisted Data Review In The Clinical Development Process”, <http://www.diahome.org/content/abstract/1998/dij917.pdf>, Drug Information Journal, Vol. 31, pp 636 – 638, 1997.

² <http://www.aafp.org/fpm/20000900/76the8.html>



Who's behind the research?

Our lead researcher, Kirk Mousley, PhD received BS and MS degrees in Electrical Engineering from MIT and a PhD in Computer Science from Lehigh University. He has been the President of Mousley Consulting, Inc. since its founding in 1993 and has directed the company's efforts in the areas of clinical database design, data editing/cleaning, document management, and submissions.

Karl Mousley received his BS in Mechanical Engineering from Rose-Hulman Institute of Technology and a MS in Computer Science from Villanova University. He has been a senior member of the technical staff at Mousley Consulting, Inc. since 1993. Among his significant accomplishments are the investigation, evaluation, and implementation of new computer technologies for clinical data management systems and developing strategic plans for integrating these technologies into current systems. He has extensive experience preparing Standard Operating Procedures (SOPs).



EDC Today and EDC In Depth

EDC Management publishes well-researched, timely information about EDC technologies and processes.

EDC Today is a free electronic technical bulletin.

Each month we examine topic areas related to Electronic Clinical Systems (ECS) theory, technology, practice, or implementation.

Each *EDC In Depth* research report comes with an executive summary and may be purchased individually for \$395 or as a group of related reports for \$975. Available via downloadable electronic version or paper version sent via mail.

To subscribe to ***EDC Today*** or purchase a specific ***EDC In Depth*** research report:

Order online at
www.edcmanagement.com

Email us at
info@edcmanagement.com

Call us at
1-484-530-0300