The Automatic Clinical Trial: An Assessment of Clinical Data Entry Errors And Data Variations In A Clinical Data Repository

Anuj Shrivastava¹ & Dr Pramod Singh Khatri²

¹MSc Clinical Research, Department of Clinical Research, Amity University Gurugram.
²Head, Department of Clinical Research, Amity Medical School, Amity University Gurugram

ARTICLE INFO

Article History:
Received on 17th Sept 2018
Peer Reviewed on 29th Sept 2018
Revised on 11th October 2018
Published on 26th October 2018

Keywords:
EDC; Data management; Risk-Based Monitoring, Source Data Verification

ABSTRACT

Background: Source data verification (SDV) is a resource intensive method of quality assurance frequently used in clinical trials. There is no empirical evidence to suggest that SDV would impact on comparative treatment effect results from a clinical trial.

Methods: To accomplish the aim, To begin the experiment, the clinical trial management software OpenClinica was implemented as part of three anterior cruciate ligament (ACL) studies that were ongoing at Medanta Hospital. Following training of the online software system, each study had to be built and eCRFs for each study had to be created de novo.

Findings: The main errors observed were simple transcription errors from the paper source documents to the EDC database. This observation was to be expected, since every transaction has an inherent error rate. What and how to monitor must be assessed within the risk-based monitoring section of the comprehensive data monitoring plan.

Interpretation: With the advent of direct data entry, and the elimination of the requirement to transcribe from a paper source record to an EDC system, error rates should go down dramatically. In addition, protocol violations and data outside the normal range can be identified at the time of data entry and not days, weeks, and months after the fact. In this empirical comparison, SDV was expensive and identified random errors that made little impact on results and clinical conclusions of the trial. Central monitoring using an external data source was a more efficient approach for the primary outcome of OS. For the subjective outcome objective response, an independent blinded review committee and tracking system to monitor missing scan data could be more efficient than SDV.
INTRODUCTION
Collection of individual patient data on Case Report Forms (CRFs) in clinical research has traditionally been done by investigators in their offices summarizing medical charts on paper forms (pCRFs), a tedious method that could result in data errors and wrong conclusions. Electronic data capture has in recent years been increasingly used in both industry and academic research settings. The feasibility of electronic CRFs (eCRFs) has been documented by numerous studies analyzing data collected on websites, laptops or digital pens. Since the mid-1990s, eCRFs have increased data quality and completeness by using alarms, automatic completions and reminders, reducing losses and transport logistics, especially for multicenter trials. Moreover, use of eCRFs permits speedier database processing and shorter study periods, resulting in lower costs. Previous studies of eCRFs have primarily focused on the investigators’ point of view, while few have documented the perspectives of the other stakeholders. Despite their demonstrated usefulness, eCRFs have not become dominant.

Clinical data management (CDM) is a vital cross-functional vehicle in clinical trials to ensure high-quality data are captured by sites staff through paper case report form (CRF) or electronic case report form (eCRF) and available for early review. The integrity and quality of data being collected and transferred from study subjects to a clinical data management system (CDMS) must be monitored, maintained, and quantified to ensure a reliable and effective base for not only new drug application (NDA) submission and clinical science reports but also corporate clinical planning, decision-making, process improvement, and operational optimization. The gradually increasing use of electronic data-capturing (EDC) technology and eCRF to collect data in clinical trials has grown in recent years and has affected the activities of clinical research operations for industry sponsors, contract research organizations (CROs), and clinical sites.

SPECIFIC AIMS
The research project will determine whether an electronic system for data management should be implemented for all research studies at Medanta Hospital. The project will introduce an online clinical trial software system, OpenClinica, to manage all research data within three anterior cruciate ligament (ACL) studies. Within the software system, electronic case report forms (eCRFs) will be created.

MATERIALS AND METHODS
The internship project involved multiple steps in order to determine whether a clinical trial management software system was feasible for use by the clinic and whether it actually reduced errors and proved beneficial for research use. First, the clinical research coordinator was designated as data manager and was trained to use the software used to build electronic case report forms (eCRFs) for the...
five patients who had completed a Week 1 research visit for the ACL study. A list was made that included the subject IDs of those 105 patients, and each ID was assigned a consecutive number between 1 and 105. An online random number generator (Appendix A) was then used to randomize the subject files to be used for Week 1 source data verification. This online random number generator produced a table of 105 random numbers between 1 and 105 without using duplicate numbers. The 105 random integers listed were used to allow the corresponding ACL subject research files to be pulled at random. The ACL subject research files were pulled in groups of five, starting from the beginning of the generated list. Chart randomization and retrospective source data verification were then completed on subject research visits for Weeks 1, 4, 8 and 12, and the Return-to-Sport visit. The methods mentioned previously were used to randomize subject files for all visits. There were a total of 98 subjects who completed a Week 4 visit, 9 subjects for Week 8, 7 subjects for Week 12, and 57 subjects for the Return-to-Sport visit. Source data verification was completed on five subjects for each of the research visits, so that a total of 25 visits were retrospectively verified. Each subject research visit contained a different number of data points, depending on what data was collected at the time of that visit. In Weeks 1 and 4, 8 data points were collected at each visit. Week 8 contained 20 data points, Week 12 had 26, and the Return-to-Sport visit contained 34 data points. Since 5 visits were verified for each week, there were a total of 480 data points considered for retrospective source data verification (Table 1). To complete this source data verification, the paper chart (or source document) was compared to the Excel spreadsheet (or pCRF). Each data point collected on that visit was compared between the two documents, and if any mistakes or discrepancies were discovered, they were indicated on the “Source Data Verification Checklist” (Appendix B) by marking the type of error discovered. There were six types of errors which were considered during source data verification. If a data point had been recorded on the CRF but was not written on a source document included in the subject chart, the error was marked as “no source document.” This error included instances where a copy had been made of the original source and the copy placed in the subject’s research folder. An error was excused if there was a “note-to-file” contained in the research folder explaining that the original data could be found in the subject’s medical file. The next type of error, documented as “mismatched source/CRF,” indicated that there were different numbers recorded for the same data point on the source document and the corresponding case report form. In an instance where the data point had been recorded on the source document but was never
transcribed onto the case report form, the error was documented as “missing CRF data.” The total number of errors was counted for each visit and the total percentage of errors was calculated by dividing the number of errors by the total number of data points for that visit and multiplying by 100.

Prospective source data verification was completed using the same techniques as the retrospective source data verification. Beginning on July 1st, 2014, a data entry log was kept for all ACL patients who came in for a research visit (Appendix C). The data collected on that visit was entered into OpenClinica. There were two Week 1 visits, four Week 4 visits, four Week 8 visits, five Week 12 visits, and five Return-to-Sport visits for a total of twenty visits available for prospective source data verification. This resulted in a total of 428 data points which were considered for prospective source data verification (Table 2). When the source data verification occurred, the data on the subject’s paper chart (source document) was compared to the eCRF in OpenClinica.

RESULTS AND DISCUSSION

Retrospective source data verification revealed that among 480 total data points, 49 errors were discovered including data points with no source document available, a mismatch between the source documents and the pCRF, missing pCRF data, and data points that were crossed out on the source document and rewritten with no initial or date to document the change. Data points with no recorded source document were the most prevalent among retrospective charts at 32 errors, comprising 65.3% of the total errors. A total of 9 mismatched data points made up 18.4%, struck out and rewritten data points made up 6 errors or 12.2%, and 2 errors consisting of missing pCRF data made up 4.1% of the errors accounted for retrospectively (Figure 1).

No errors were discovered in which there were multiple sources for the same data point or the writing of the data point was illegible (Table 1). It was noted during retrospective verification that the research charts completed earlier in the year had fewer source documents contained in the subjects’ research files. As the study continued and a research coordinator was hired, the number of missing source documents decreased due to an understanding of the problem of missing source documents and an active effort to reduce this error. The responsibilities of this research coordinator included improvement of data capture by ensuring only one member of the study staff was collecting and recording data in a single database using consistent methods for each study.

Errors discovered in retrospective source data verification were also evaluated by the subject visit in which they occurred. Error rate was calculated by dividing the number of errors discovered by the total data points verified among all charts for that visit and multiplying by 100. It was discovered that the highest error rate (40%) occurred in the Week 1 visit. Five percent of errors occurred in Week 4, 6% in Week 8, 13.1% in Week 12, and 4.7% at the Return-to-Sport visit (Figure 2). Prospectively, only 6 errors were discovered upon evaluation of 428 data points that were entered in the eCRF format. It is interesting to note that no data points were found to be missing in the eCRFs (Table 2). This can be attributed to the fact that OpenClinica required all data points for each visit to be entered, or required an explanation of why the data point(s) were omitted, before the eCRF data could be saved in the system. These requirements leave no room for error in regards to missing data when transcribing from the source document to the eCRF. There were also no errors involving multiple source documents or illegible writing (Table 2). Three errors (or 50%) were attributed to no available source document. Mismatched source document and eCRF data accounted for 16.7% of the errors, and two data points were found to be struck out and rewritten, accounting for 33.3% of the errors in prospective source data verification (Figure 3).

The “no source document” and “strikeout/rewrite” errors are considered to be associated only to the source document and not the eCRF. These errors occur during data collection when they are handwritten on the source document, and therefore might not reflect on the error-reducing capacity of the eCRF compared to the pCRF. A mismatched data point simply occurs by human error during transcription of the data from the source document over to the eCRF.

The error rate was also evaluated by visit number within prospective source data verification. The error rate across visits revealed a downward trend in which errors appeared to decrease at each subsequent subject research visit. Week 1 had an error rate of 6.3%, the Week 4 error rate was 3.1%, Week 8 was 2.5%, and Week 12 was 1.5%. There were no errors discovered in the Return-to-Sport visits prospectively (Figure 4). When evaluating the overall error rates in the retrospective and prospective formats, a significant reduction in error
was obtained when the eCRF format was used. Retrospectively, there were 49 errors out of 480 possible data points, displaying an overall pCRF error rate of 10.2%. Prospectively, there were only 6 errors out of 428 possible data points, displaying an overall eCRF error rate of 1.4% (Figure 5). When the eCRF was used in place of the pCRF in the study, there was an 86% decrease in errors made during data collection and transcription, demonstrating the ability of a software system such as OpenClinica to increase reliability of data in a research study. A \( \chi^2 \) analysis with Yates’ correction for continuity was completed to compare the error rates discovered in the retrospective and prospective formats. Taking into account 1 degree of freedom and a \( \chi^2 \) value of 29.308, the analysis resulted in a p-value of <0.001, demonstrating a statistically significant decrease in total errors in data points following the use of an electronic data management system.

Figure 6 shows a direct comparison between the types of errors made in the pCRF and eCRF formats. There was a decrease in twenty-nine data points with no source documentation among errors made when the eCRF format was used. There was also a slight decrease in the number of data points (eight) whose source documentation did not match the CRF documentation among errors made when the eCRF was used. Another decrease in error was demonstrated with missing CRF data, in which there were two errors in pCRFs from missing pCRF data, while no eCRF data was found to be missing. There was also a decrease in the number of data points that were crossed out and rewritten once the eCRF format was implemented, regressing from six to two errors. Finally, error rates sorted by visits were directly compared (Figure 7). Consistent findings showed that the error rate decreased with use of an eCRF in each research visit. Data analyzed in weeks 1 and 12 of these ACL research studies in particular seem to be the most important in predicting how well a patient will do in his or her return-to-sport test. If the data at these crucial visits are incorrect or missing, the physical therapist might misinterpret the patient’s readiness for return-to-sport. This practicum project actually showed that use of an eCRF system has the potential to reduce errors, particularly at the Week 1 and Week 12 visits. In Week 1, errors decreased from 40% to 6.3% with the use of an eCRF \( (\chi^2 = 4.664, v = 1, \alpha < 0.05) \). Week 12 showed a decrease in error rate from 13.1% to 1.5% \( (\chi^2 = 11.13, v = 1, \alpha < 0.001) \). Aside from a measurement of error reduction when using eCRFs, this practicum report was designed to subjectively measure the feasibility of the electronic data management software, the possibility of multiple data capturing methods, and the accuracy of an audit trial and limited data access. While OpenClinica did demonstrate a reduction in the number of data errors and an increase in data quality, its feasibility and ease of use were lower than expected. The software required no installation as it is web-based, which facilitated installation and accessibility on any computer throughout the research site. Building each of the studies within the online system did not take much time to complete (around one working hour), and there was a step-by-step question and answer portion about each study that allowed the software to custom-format each study to meet the needs of the research team. Construction of the eCRFs within OpenClinica seemed to be the most challenging task. From start to finish, it took around 24 working hours to complete a CRF with 428 data points. Following completion of the eCRF and its implementation in the correct portion of the research study in OpenClinica, it took around 8 working hours to edit the eCRF and fix any problems that were not discovered at the time of construction. For this practicum, two eCRFs were created to fit the needs of the three research studies evaluated, totaling around 66 working hours on building the studies and eCRFs before data entry could even begin. The research site currently has five more actively enrolling studies, which would take an additional 165 working hours before data entry could begin.

It is important to note that some of the errors accounted for in source data verification are associated only with human error during data collection on the source document, while some are associated with human error during transcription of the data from the source document to the pCRF or eCRF. The “no source document”, “multiple sources”, “illegible writing”, and “strikeout/rewrite” errors can be attributed solely to human error during data collection and the error rate should be consistent regardless of the type of CRF used. The “mismatch and missing CRF data” errors occur during data transcription onto the CRF, and should be reduced by the use of an eCRF, which requires all data points to be entered and has regulations on a range of values that can be entered for each data point. So while the use of OpenClinica still included errors in data collection and transcription, the
method of data collection appeared to be the prime source of errors. A solution to this problem would be to use an electronic tablet with OpenClinica capabilities for data collection. This approach would eliminate the use of a paper source document entirely, and the source document and CRF are essentially the same. In this format, the data entry user would enter the ACL measurements in the CRF during data collection in the physical therapy clinic. The eCRF would require all relevant data points to the patient visit to be entered at the time of data entry, so no measurements could be inadvertently missed. No transcription of the data points would occur because all patient data will have already been entered into OpenClinica. This aspect would considerably reduce human error including the lack of a source document, mismatched data points, missing CRF data, multiple sources, strike-outs or illegible writing.

Implementation of electronic tablets is not yet feasible at the site for financial reasons, but could be considered for application in multiple studies in the future. OpenClinica did result in a much better system for decreased user access to data and creation of a time-stamped audit trail in comparison to the previously used methods. Currently, the data for all studies at the research site is contained on a network drive that can be accessed by multiple users in the physical therapy department. Although there are regulations governing which folders may be accessed by specific users, these folders are not password protected, raising the possibility of security breaches. Anyone from the research team is allowed to view the data from the research studies to which they are assigned, and they can open the Microsoft Excel database, make changes to the data, and overwrite the old database without any audit trail of the changes or the research team member who made them. In OpenClinica, all research databases are password protected, and the data manager has the capability to assign roles in the system to each member of the research team depending on their level of activity in the study. Some users can only view the data, while others can add new subjects or data or make changes to existing data. When these changes are made, OpenClinica keeps a time-stamped audit trail called a “Study Audit Log” for each research study and each subject folder documenting all changes (Appendix D). This study audit log tracks when a user creates a subject, when data is added, edited, or deleted, and any other actions made to the subject’s chart. If data is edited or deleted, the audit log shows what previous data was entered in case it needs to be changed back to the original number. This system holds research staff accountable for all changes made to the database, an element that is non-existent in the current procedures at the research site.

**Table 1: Results of Retrospective Source Data Verification**

<table>
<thead>
<tr>
<th>RETROSPECTIVE</th>
<th>Week 1</th>
<th>Week 4</th>
<th>Week 8</th>
<th>Week 12</th>
<th>RTS</th>
<th>Total Errors</th>
<th>% Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Source Doc.</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>4</td>
<td>32</td>
<td>65.3</td>
</tr>
<tr>
<td>Mismatch</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>18.4</td>
<td></td>
</tr>
<tr>
<td>Missing CRF Data</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>Multiple Sources</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Illegible Writing</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Strikeout Rewrite</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>12.2</td>
</tr>
<tr>
<td>Total Errors</td>
<td>16</td>
<td>2</td>
<td>6</td>
<td>17</td>
<td>8</td>
<td>49</td>
<td>100</td>
</tr>
<tr>
<td>Data Points</td>
<td>40</td>
<td>40</td>
<td>100</td>
<td>130</td>
<td>170</td>
<td>480</td>
<td></td>
</tr>
<tr>
<td>% Errors</td>
<td>40</td>
<td>5</td>
<td>6</td>
<td>13</td>
<td>4.7</td>
<td>10.2</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Results of Prospective Source Data Verification

<table>
<thead>
<tr>
<th>PROSPECTIVE</th>
<th>Week 1</th>
<th>Week 4</th>
<th>Week 8</th>
<th>Week 12</th>
<th>RTS</th>
<th>Total Errors</th>
<th>% Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Source Doc.</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>Mismatch</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>16.7</td>
</tr>
<tr>
<td>Missing CRF Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multiple Sources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Illegible Writing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Strikeout/Rewrite</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>33.3</td>
</tr>
<tr>
<td>Totals</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>Data Points</td>
<td>16</td>
<td>32</td>
<td>80</td>
<td>130</td>
<td>170</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>% Errors</td>
<td>6.3</td>
<td>3.1</td>
<td>2.5</td>
<td>1.5</td>
<td>0</td>
<td>1.4</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Types of errors discovered in pCRF source data verification

![Pie Chart: pCRF Errors by Type]

- No Source Doc: 65.3%
- Mismatch: 18.4%
- Missing CRF Data: 12.2%
- Strikeout/Rewrite: 4.1%

Figure 2: Error rate by week discovered in pCRF source data verification

![Bar Chart: pCRF Errors by Visit]

Error Rate (%) vs Subject Visit
Figure 3: Types of errors discovered in eCRF source data verification

![Pie chart showing eCRF Errors by Type]

- No Source Doc.: 50.0%
- Mismatch: 33.3%
- Strikeout/Rewrite: 16.7%

Figure 4: Error rate by week discovered in eCRF source data verification Overall Error Rate Comparison

![Bar chart showing eCRF Errors by Visit]

- Error Rate (%) by Subject Visit:
  - Week 1: 8%
  - Week 4: 6%
  - Week 8: 4%
  - Week 12: 2%
  - RTS: 1%

Figure 5: Comparison of error rate between pCRF and eCRF formats ($\chi^2 = 29.308$, $\nu = 1$, $\alpha < 0.001$)

![Bar chart showing Overall Error Rate Comparison]

- Overall Error Rate (%):
  - pCRF: 10%
  - eCRF: 2%
DISCUSSION
The main errors observed in the present study were simple transcription errors from the paper source documents to the EDC database. This observation was to be expected, since every transaction has an inherent error rate. Clearly, early training on how to complete the date fields in the micturition diary log would have made an impact on the frequency of data changes. Once these types of errors and their magnitude are identified by the clinical research associates and by data management, the clinical sites could easily have been retrained. Changes to the text fields for medications and adverse events could easily be monitored remotely and rules for data entry established. For example, for medications, based on the observations from a simple database summary table, the sites could be instructed just to enter the drug name and not to enter units. For adverse events, the sites could be instructed not to enter any abbreviations or just to indicate the adverse event as “injection site reaction.” These types of data errors must be assessed early in the clinical trial as they could make the results from the form invalid. The choice between paper and electronic CRF is a significant step in the design and execution of clinical studies; it should be discussed with the involved stakeholders and based on efficiency. With the advent of direct data entry and the integration of EDC with EMR, and the elimination of the requirement to transcribe from a paper source record to an EDC system, error rates should decrease.
dramatically with a corresponding improvement of data quality.
In addition, protocol violations and data outside the normal range can be identified at the time of data entry and not days, weeks, and months after the fact. When the clinical research sites are able to bypass the use of traditional paper source documents as original data, and the pharmaceutical industry performs risk-based monitoring together with data management and statistical tools, there is the real
 possibility to:
1. Increase the quality of clinical trial data.
2. Reduce the time to database lock.
3. Stop development of ineffective or unsafe drugs early in the cycle.
4. Reduce unnecessary work.
5. Reduce the costs of clinical trials.
6. Accelerate the time to market

CONCLUSIONS
Data quality is an essential indicator of reliability within research studies. In clinical research studies, there is always the chance for human error to occur within data acquisition and transcription, allowing the possibility of impairment of data quality. In this practicum, use of an electronic case report form within clinical trial management software has been shown to decrease the likelihood of human error in the essential data management aspects of a clinical research study. Future research studies could identify the feasibility of implementation of multiple types of software that still improve data quality in physical therapy research.

REFERENCES
Your next submission with British BioMedicine Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats (Pdf, E-pub, Full Text)
- Unceasing customer service
- Immediate, unrestricted online access
- Global archiving of articles

Track the below URL for one-step submission

http://www.britishbiomedicine.com/manuscript-submission.aspx