**PhUSE 2016**

**Paper SP03**

# Upstream with your statistician and a data visualisation paddle

Nick Cowans, Veramed, Twickenham, United Kingdom

## ABSTRACT

While a clinical trial is ongoing, data flows in a continuous stream from the patient to the statistician with multiple steps in between. At one end of this, clinicians enter small trickles of data about individual patients. At the other, statisticians summarise and analyse large amounts of data into single numbers and points on tables and figures. Data management traditionally sit close to the source of this flow, and so might not get the best view on how their data checks will affect the final outputs. This can lead to inefficiencies and frustrations as data essential to statisticians remains dirty. In order to improve the data cleaning process, statisticians can get involved in designing novel tools that visualise issues in the data important to them. These tools can be used by clinical and data management colleagues to identify the data queries most critical to the endpoints in the study.

## INTRODUCTION

In an ordinary clinical trial, subjects are randomised to different treatment groups and various data about these subjects are collected. Inferences about the effects of the treatments of interest on endpoints of interest are then made. This process starts with a research question, from which a hypothesis is generated. The trial is then designed and run to collect data in an attempt to provide evidence to test this hypothesis. The results of clinical trials are finally published in reports and peer reviewed articles, for various regulatory, payer and other bodies to interpret to decide policy.

To allow those who must interpret the results of clinical trials to have confidence in the findings, the integrity of the data is crucial. Although those reading the final reports on a clinical trial may focus on single numbers in a table or in text, such as percent reductions in risk, a change from baseline, or a p-value, it is important to keep in mind that these numbers are summaries or analyses from hundreds or thousands of data points. If the data is not collected correctly, or there are mistakes in the recording of the data, then this could have an impact on the results and the interpretations made from them.

One part of the solution to this is clinical data management, which has become a fundamental part in the clinical trial process. The data management department facilitates the collection of data in a trial. It also holds the keys to process of checking data for inaccuracies and making attempts to correct errors via the issuing of queries back to the sites. However, it is the statisticians in the clinical trial who will plan and carry out the analyses of the data. It is the statisticians who will know most about which data issues are of most importance to the analyses of the data. Statisticians should contribute to data cleaning to ensure the data are of sufficient quality for analysis. While it is common for statisticians to contribute to the initial design of data checks, to catch unforeseen issues statisticians need to collaborate with data management on data cleaning during the trial.

In this paper, the author will draw upon his experience of working as a statistician on a large, phase III respiratory trial looking at the effects of investigational product (IP) on various clinical endpoints in chronic obstructive pulmonary disease (COPD) patients. This study was an event driven trial that had staggered subject recruitment over 4 years and a further year of follow up after the last subject was recruited. This resulted in a vast amount of data collection from over 16,000 subjects in more than 1300 sites in 43 countries. During this five year period, all involved in the trial were blinded to the randomised treatment groups the patients were taking. Monthly extracts of the live data was sent blinded to the statistics and programming team, who used it to program tables and figures using dummy randomisations. This allowed the statistics department to recognise data issues that may not have been picked up by data management. In addition, various visual tools were designed and programmed in SAS and other programming languages to demonstrate to both the clinical and data management departments potentially unclean data, and highlight how not attempting to clean it might affect the statistical analysis of our endpoints.

This paper aims to highlight the benefits of statisticians continued engagement in the data cleaning process during the running of a clinical trial and in particular the use of graphical tools to facilitate clinical review of suspect data.

## FLOW OF DATA IN A CLINICAL TRIAL

Data flows in a clinical trial from subject to final report via several steps. It will begin as thousands of individual snippets of data: measurements, dates and events collected about individual subjects at multiple visits at different times around the world. It will pass through several transcription, data entry, mapping and data transformation steps before it is summarised and analysed on source tables or figures which are then referenced further in a report.

In the CDISC study in this paper, the flow of data from subject to report or publication is presented in Figure 1. Each of the over 16,000 subjects from over 1300 sites had a screening, a baseline and then up to 16 scheduled visits, and

any number of unscheduled visits where data was documented. The data from these visits, including measurements of FEV1, questionnaires, along with information about the subject disposition, log forms of events and concomitant medication was first entered into a physical medical record for each subject. This data was then entered by the sites into the subject's trial electronic case report form (eCRF), a live database operated by a third party CRO. Due to the fact that the study was in this ongoing phase for over 5 years, snapshots from the live database were taken every month and mapped to SDTM datasets by the same third party CRO carrying out the data management. These monthly SDTM packages were delivered to the statistics and programming department, working on the client's system. Dummy treatment groups were assigned to the subjects and ADaM datasets built. From these dummy TFLs were programmed, which after unblinding would become the final TFLs.

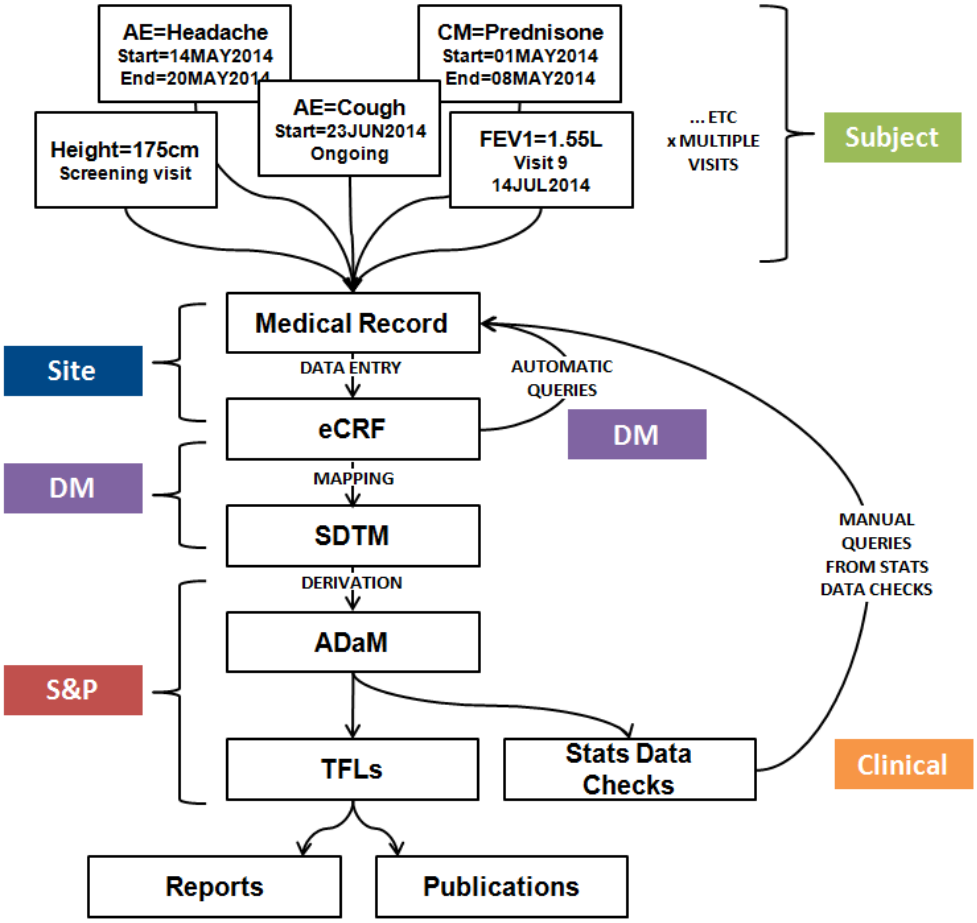


*Figure 1: Flow of data in the current clinical trial.*

## ENSURING DATA QUALITY

Data management collect and manage data in a clinical trial. Part of this process will usually involve programming data checks so that various data rules are not violated. Examples include ensuring that a height entered for a subject is nonnegative and within a sensible range, checking that start dates are not after end dates or that a subject's age does not fall outside of the inclusion criteria for the trial. Where true electronic data capture (eDC) occurs, and the information is automatically recorded into the eCRF as source, these data checks can be built into the system so that the invalid data can not physically enter into the database. However, often it is more practical to run scripts of data checks on the data after it has been entered, and then generate automatic queries to the site after issues have been identified.

In addition to general data checks that are common to the majority of clinical trials, it is helpful for the statisticians in the trial to get involved with the design of study specific data check rules before the study has begun collecting data. In the current study, several of our endpoints required the measurement of FEV1, which is a measure, in mL, of lung function. Before the study began, the statistics team worked with data management to create some rules on study specific issues including FEV1 - see below.

Despite this, and especially relevant in a large, late phase study with long study exposure, it is inevitable that issues with data will arise that were not foreseen before the study began (and contracts were signed). While a capable and experienced data management department will always identify and attempt to fix any new issues that arise, as

shown in Figure 1, they sit at the other end of the flow of data to the analyses and reports. It is the statistician who sits closest to the TFLs and analysis, and who can see exactly how data issues might affect the individual numbers on a table or figure.

## VISUAL TOOLS

Two examples of visual patient profile tools, which were created by the statistics and programming team in this study, will now be described. These examples examine two endpoints common to respiratory trials: FEV1 and exacerbations. These tools were designed to demonstrate to clinical and data management colleagues the effect that potential data issues would have on the final statistical analyses of our endpoints if no attempt was made to clear them up.

### FEV1

Various spirometry measurements can be collected in a clinical trial in order to assess lung function. Forced Expiratory Volume in 1 second (FEV1) is measured by a subject performing a forced expiratory manoeuvre: following a full inhalation, the patient exhales fully as quickly as possible into a spirometer – a medical instrument designed to measure volumes of air exhaled. Taken from at least three attempts, FEV1 is the maximal volume of air exhaled in the first second of a forced expiratory manoeuvre, and is a proxy for lung obstruction; those with greater pulmonary obstruction will be able to exhale less air in the first second than those with less obstruction.

With age and disease progression, lung function declines for all patients. However, several long acting therapeutic interventions have been demonstrated to slow the rate of this unavoidable decline. It is therefore common in a respiratory trial to collect FEV1 at baseline and then at various scheduled visits spaced out throughout the trial, for example every 3 months, to allow the rate of decline in FEV1 to be evaluated.

In the current study, the effect of drug interventions on the rate of decline in FEV1 was a secondary endpoint. The primary model used to assess this was a random coefficients model where post-baseline FEV1 was the response variable, the slope estimate of interest was that of the treatment by time interaction and subject effects were treated as random. In a random coefficients model such as this, subjects with more data are given more weight. A sensitivity analyses was therefore included where an individual slope was estimated for each subject, giving each subject the same weight.

Since FEV1 was used in a secondary endpoint, an appreciable amount of time was spent by both data management and statisticians in ensuring that the data collected was clean. Before the study began, rules were set up that, when broken, would send out automatic queries. For example, when an FEV1 measurement was entered into the eCRF that was outside a biologically plausible range for any subject, a query was automatically sent to the site. Moreover, rules were also created to issue queries when an FEV1 measurement was entered that was questionable given that subjects baseline FEV1 measurement, or given the recording at that subjects previous visit. Deciding on the parameters of this longitudinal data query rule also required input from the clinical team.

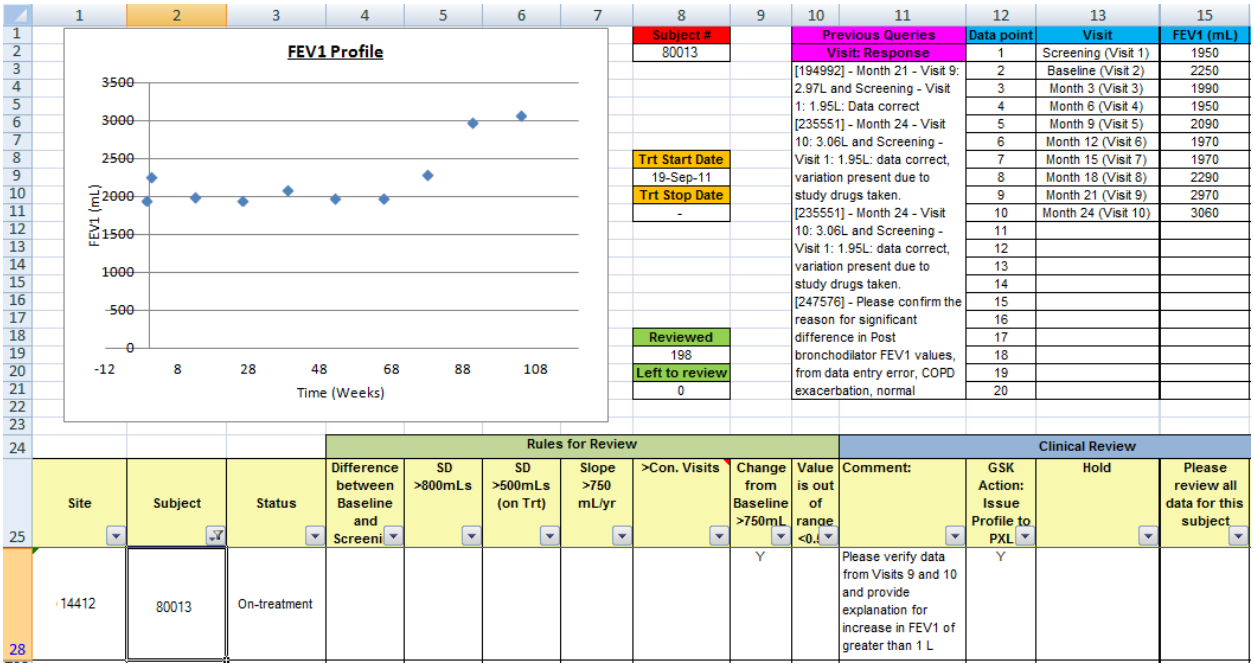

*Figure 2: FEV1 Profile tool in Excel for clinical colleagues to review.*

Despite these rules being in place, questionable FEV1 data was seen coming through in the blinded monthly extracts. When it was investigated, it became apparent that although the queries were being sent to verify the potentially incorrect result, there was a high frequency of rapid site responses stating that the *"data is correct"*. The receipt of any response automatically resulted in the query closing, which led to the data coming through unchanged. Of course, there are many instances where data that may appear questionable are entirely correct. Despite this, it was felt that some of issues would benefit from re-querying. Given the size of the trial the effort of reissuing all queries would be huge and largely fruitless. The statistics and programming team decided to invest some time in designing an FEV1 profile review tool that could maximise efficiency of this task.

A program was written in SAS to identify subjects with questionable FEV1 data, not just those that had been queried before, but also those that would have significant consequences on the statistical analyses if not cleared up. With a sensitivity analyses in mind in particular, those with extreme slopes, such as subjects with just two post baseline FEV1 visits very close together, could have a significant impact on the results. All the FEV1 data about the subjects of interest was extracted from our ADaM datasets, along with all the query information to date, provided by data management. This was then imported in a hidden lookup sheet in Microsoft Excel. In a different, visible, sheet in Excel, a spreadsheet was designed which allowed a user to select an individual subject using the AutoFilter feature. Once the filter was applied, the spreadsheet used the VLOOKUP function to populate the visible cells with data about that subject, and plot the FEV1 measurements on a graph. An example of this tool in use is shown in Figure 2.

This method may sound coarse, and one could argue that there are alternative, more appropriate tools available in which to do this job (Spotfire, R, SAS). However, Excel was chosen because it was a piece of software with which everyone was already familiar, which meant that clinical colleagues were able to use with no additional access or training barriers. Periodically, throughout the blinded phase of the trial, this process was run and produced a spreadsheet for a member of the clinical team to review. The reviewer was then able to enter bespoke, more detailed, comments in this spreadsheet, which were fed back into SAS in order to produce a PDF showing the FEV1 profile of the subject and the query from the clinical team, an example of which is shown in Figure 3. This PDF was sent to the individual sites via data management, and the investigators were asked to respond to the query in a box on the PDF and sign and return to us.

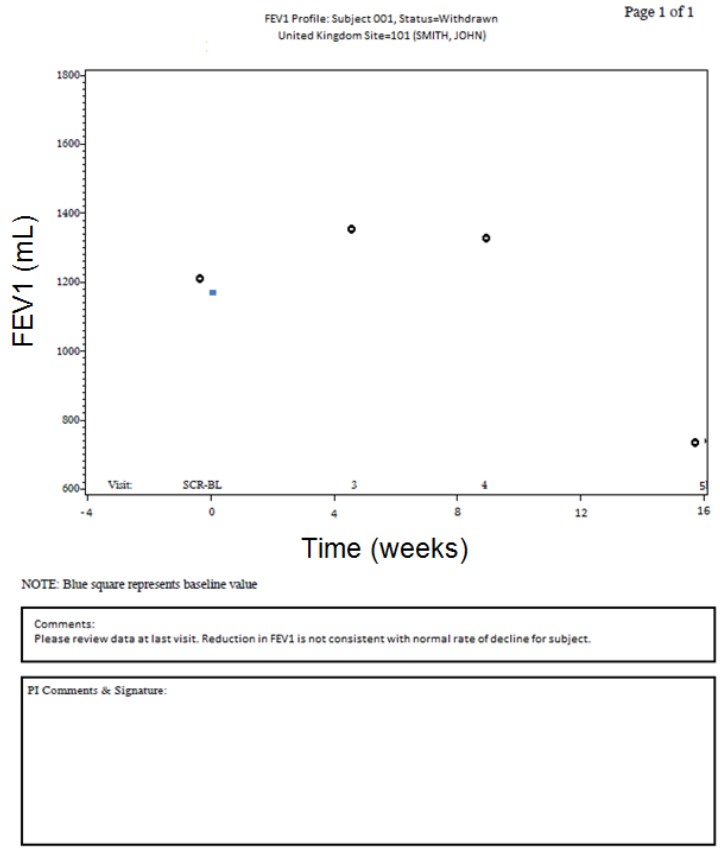


*Figure 3: Example FEV1 Profile PDF sent to sites.*

Having this additional level of oversight, where specific issues were carefully raised on an individual basis with the site about a particular subject resulted in improved responses compared to the automated queries. Although it was

still found that many of the data points remained unchanged, there was now more detailed information as to why the data points looked unusual, which resulted in more confidence in the FEV1 data.

**EXACERBATIONS**

Patients with COPD often suffer from daily symptoms such as breathlessness, coughing and wheezing. In addition, such patients are also prone to sudden worsening of symptoms that last for several days and may require treatment with antibiotics or systemic corticosteroids, and in severe cases, result in hospitalisation. These sudden deteriorations are known of as exacerbations, and are usually brought on by an infection or environmental change.

A desirable quality of an interventional therapy is to reduce the probability of experiencing exacerbations. In the current study, a tertiary objective was to study the effect of IP on COPD exacerbations. This was carried out by analysing both the annual rate of exacerbations as well as time to the first exacerbation.

For this trial, a COPD exacerbation was recorded as an adverse event on the adverse event page of the CRF. Investigators were able to enter text to describe the diagnosis of each adverse event, which was coded to a MedDRA preferred term. For every adverse event entered, the eCRF asked investigators the question "*Did this AE meet the protocol definition of moderate / severe COPD exacerbation?*". For the analyses, it was the answer to this Yes/No question that was used to identify COPD exacerbations in this study. In a study with nearly 50,000 adverse events recorded, it was important for these data to be clean. The statistical team worked with the clinical team to ensure that unacceptable terms for COPD exacerbations, where the investigator had clearly answered 'Yes' to this question in error, were queried and cleaned up.

A generalised linear model assuming the negative binomial distribution was used to analyse the rate of COPD exacerbations. For rates, it is imperative that the same events are not counted more than once. Due to the way that exacerbations were recorded in this study, it was found that on several occasions the investigators had entered two adverse events with the same start and end dates for the same subject. One would have a clinical term, such as "Bronchitis" and the other would have the term "Exacerbation of COPD". Medically, this is just one event: an exacerbation of COPD caused by a bronchitis infection. However, when the investigator had assigned both adverse events as meeting the protocol definition of a COPD exacerbation, this would translate to two events in the dataset. The investigator, who sits even further away from the analysis of data than data management, may not have realised how COPD exacerbation events were being collected in this study, and had inadvertently double counted this event.

In other situations, two exacerbations had occurred, but the second had started less than 7 days after the first had ended. Without querying, these would have been recorded as multiple events, despite the general clinical opinion that the second event is more likely a relapse of the first, and that this is only one event. Although a statement for collapsing such exacerbations was included in one of the statistical analysis plans, the preference was to have these tidied up in the source database where possible, given that the data would be available for share at the end of the study.

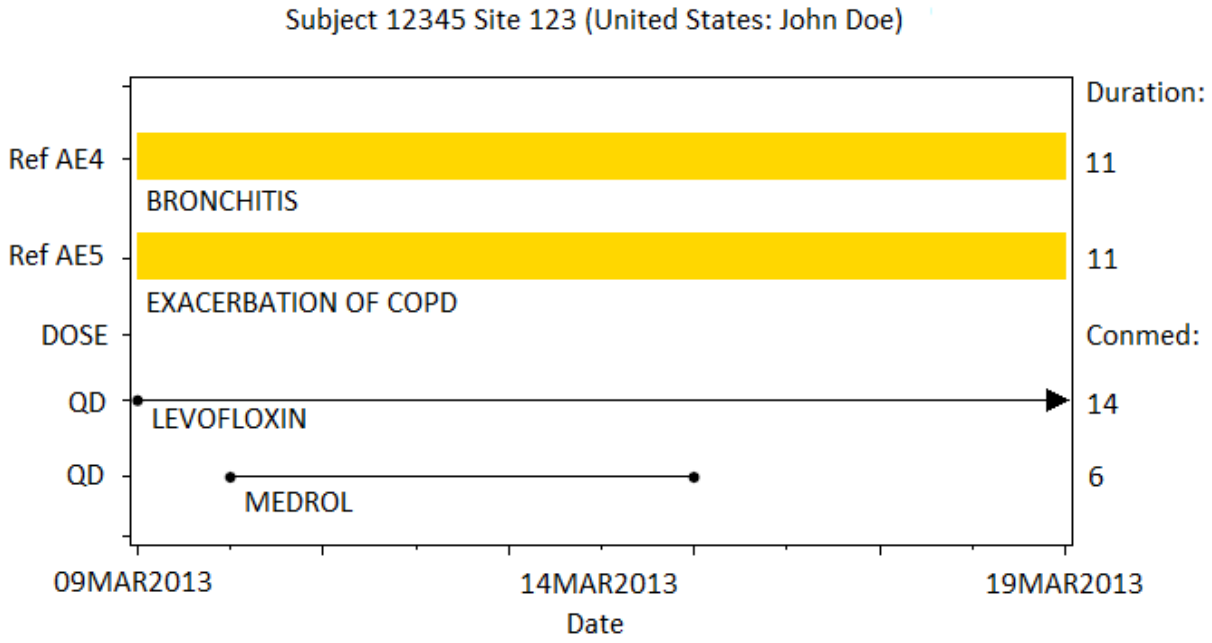


*Figure 4: Example of overlapping exacerbations patient profile reviewed by clinical colleagues.*

In order identify cases such as this, and have them reviewed before issuing queries, an exacerbations profile tool was designed and produced. Overlapping adverse events that were flagged as meeting the protocol definition of being COPD exacerbations were identified in our ADaM datasets. In SAS, a program was written that plotted each overlapping exacerbation, shown in Figure 4. Along the bottom axis was date, and for each overlapping exacerbations, identified by their adverse event reference ID on the left axis, a rectangle was plotted along with the verbatim term entered by the investigator. The duration of the exacerbation was annotated onto the right hand axis. On request from the team reviewing these profiles, underneath the exacerbations, any concomitant medications that were taken during the period of interest were also plotted, along with duration and dosage. These plots were sent to the clinical team to review, who made decisions on which to query, and again were able to create tailor written queries.

As was seen with the FEV1 profile tool, these individually written queries, where the effect of the unclean data had been visualised, resulted in better responses from sites than the automated queries.

## CONCLUSION

Cleaning of clinical trial data is traditionally carried out by data management, who sit close to the collection of the data. Statisticians sit at the other end of the process, but need to get involved in data cleaning to identify data issues of most importance to the analyses of the endpoints. Continued collaboration with data management beyond initial setting up of data checks is required to address subtle issues that emerge during data collection. Data can be difficult to review when displayed in typical dataset standards such as listings, so unique visual tools can be created to help our clinical and data management colleagues review more easily. By concentrating on the data issues of most importance to the analyses, statistical involvement in data cleaning can increase efficiency in the data cleaning process and improve the quality of the final analyses.

## ACKNOWLEDGMENTS

Abigail Fuller (Veramed Limited) for her input into this paper and design and production of the FEV1 profile tool.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nick Cowans
Veramed Ltd
5th Floor Regal House, 70 London Road
Twickenham TW1 3QS
United Kingdom
Work Phone: +44 (0) 203 696 7240
Email: nick.cowans@veramed.co.uk
Web: http://veramed.co.uk