Paper DH03

Optimizing Data Validation

Andrew Newbigging, Medidata Solutions Worldwide, London, United Kingdom

ABSTRACT

Much effort goes into the specification, development, testing and verification of programmatic edit checks to ensure that the error rate in clinical trial data is sufficiently low as to have no statistically significant effect on the overall trial results. An analysis of several thousand clinical trials, containing over 1.1 billion data values and 1.1 million edit checks, shows that the majority of edit checks (60%) have no impact on data quality; none of these 678,000 edit checks have generated a single data query or discrepancy. What can be learnt from this analysis; can we reduce the overall number of edit checks without compromising data quality; can we identify the 'high-performing' edit checks and improve CRF design to avoid data entry errors; are there novel methods that might achieve similar standards of data quality with less effort?

INTRODUCTION

Electronic data capture systems provide a variety of implicit and explicit methods to avoid, identify and correct errors and inconsistencies in clinical trial data. The objective of such methods, in conjunction with manual processes such as data monitoring and data audits, is to reduce the data error rate to a level that has no bearing on the statistical analysis of the trial results.

Opinions vary on the threshold of acceptability for data error rates:

'… articles regarding this topic maintain that data as well controlled as clinical trial data should have errors only in the range of 10 to 50 per 10,000. This translates into .1% to .5%' [i]

Recent draft regulatory guidance acknowledges that some data is more critical than other data in its impact on study results and that therefore a risk-based approach can be used to determine an acceptable error rate and the consequent level of effort required:

'There is increasing recognition that some types of errors in a clinical trial are more important than others. For example, a low, but non-zero rate of errors in capturing certain baseline characteristics of enrolled subjects (e.g., age, concomitant treatment, or concomitant illness) will not, in general, have a significant effect on study results. In contrast, a small number of errors related to study endpoints (e.g., not following protocol-specified definitions) can profoundly affect study results, as could failure to report rare but important adverse events.' [ii]

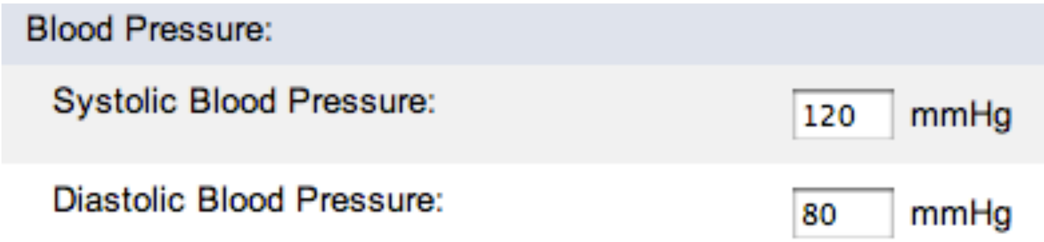
This paper does not attempt to address the question of what error rate is acceptable but does investigate the effort put into data validation checks and their effectiveness and impact on data quality.

ELECTRONIC DATA CAPTURE

Electronic data capture (EDC) systems are increasingly used to collect clinical trial data and have been shown to have low rates of data error[iii]. EDC systems typically provide features that help to reduce data errors and inconsistencies, ranging from simple constraints to complex programs:

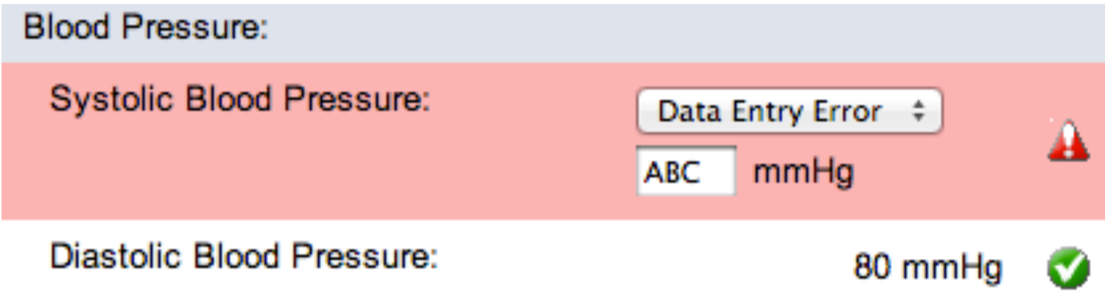
DATA ENTRY CONSTRAINTS

Unlike paper forms, electronic data collection forms may provide hard constraints on data entry fields to prevent out of range values. In the following example the data entry screen will not accept more than 3 characters in the blood pressure data entry fields, thus preventing typing mistakes leading to scientifically impossible values:



DATA TYPE CONFORMANCE

Each data entry field is usually associated with a simple data type (integer, float, string, date, time) which can be used to automatically apply data type conformance checks. In the example below the Systolic Blood Pressure field has been defined as an integer type which allow the electronic system to automatically check data entry for conformance against the type definition:



RESTRICTED VALUE SETS

Data entry fields may be associated with a set of permissible values and the electronic data capture system may restrict the user to selecting a value from this list, using data entry controls such as selection lists or radio buttons. In the following example the Outcome variable on an Adverse Event form is restricted to one of 6 values; no other data value can be entered by the user:

EDIT CHECKS

Electronic data capture systems usually have the ability to program edit checks that can provide further validation of data values beyond the constraints and conformance checks illustrated above. The construction and testing of edit checks varies from system to system, from graphical interfaces to free-format program code. In this paper we represent programmatic edit checks using pseudo-code, for example:

```
if SystolicBloodPressure > 180 then OpenQuery

if SystolicBloodPressure < DiastolicBloodPressure then OpenQuery
```

EDIT CHECK COMPLEXITY SCORE

To assess the effort required to create an edit check and to verify that it functions correctly we developed a complexity score algorithm to calculate a numerical score derived from the number of logical steps, number of logical branches ('or' clauses) and the number of data fields included in the check.

A simple uni-variate range check has a complexity score of 3, one point for the creation of the check and one point for each data value required to test that the check has been implemented correctly:

```
if SystolicBloodPressure > 180 then OpenQuery
```

Test data	Result
180	No query
181	Query opened

Adding an additional logical clause to the edit check raises the complexity score to 9, reflecting the increased size of the check specification and the increase in the number of values required to test that the check function correctly.

```
if SystolicBloodPressure < 90 or SystolicBloodPressure > 180 then OpenQuery
```

Test data	Result
89	Query opened
90	No query
180	No query
181	Query opened

SOPs for building studies may vary in their requirements for rigorous testing of edit checks, and in the case of simple checks such as the above examples a simple Quality Control inspection may be deemed sufficient to verify that the check has been implemented correctly. However many edit checks are significantly more complex and verification by inspection is not feasible. In our analysis we found that

30% of edit checks in our sample have a complexity score greater than 9, 5% have a complexity score greater than 100 and the most complex check scored 15,945 points on our rating scale (this check cross-references 41 data fields and comprises more than 2,000 lines of code with multiple logical branch points).

No attempt has been made yet to correlate the complexity score rating with actual effort by study builders. While the range in scores [3-15,945] may seem large it does not seem unreasonable given the corresponding variation in size and complexity of the edit checks; consider equating the complexity score with minutes of effort. At the extreme end of the range the most complex check would equate to approximately 30 days of effort to create and test. Measuring productivity by lines of code is notoriously challenging but 67 (2000 lines/30 days) lines of code per day is within the range that may be expected for typical software development.[iv]

METHODS

Our analysis uses data from the Medidata Rave® Electronic Data Capture system, combining data from multiple sponsor databases including global pharmaceutical companies conducting hundreds of studies, biotechnology companies running single studies and academic medical research centres. All phases of clinical research from Phase I through to Phase III and post-marketing studies are included in the sample. The sponsors and clinical sites included all major geographies (Americas, Europe, Asia Pacific). Most studies were conducted in English, with Japanese, Chinese and other languages being used for data collection on a small number of the total studies. Our analysis at this stage does not attempt to investigate differences between research phases, type of sponsor or therapeutic area.

Test and other non-production data was excluded from the analysis dataset, as were system-derived data fields (for example where the EDC system automatically calculates Body Mass Index from Height and Weight). It was not possible reliably to distinguish between data entered by human users and data entered by system processes such as batch loading of laboratory data.

Our final dataset for analysis included 1,137,496 edit checks, 1,160,836,888 datapoints and 29,455,496 data queries generated by edit checks.

RESULTS

CHANGES TO DATA VALUES

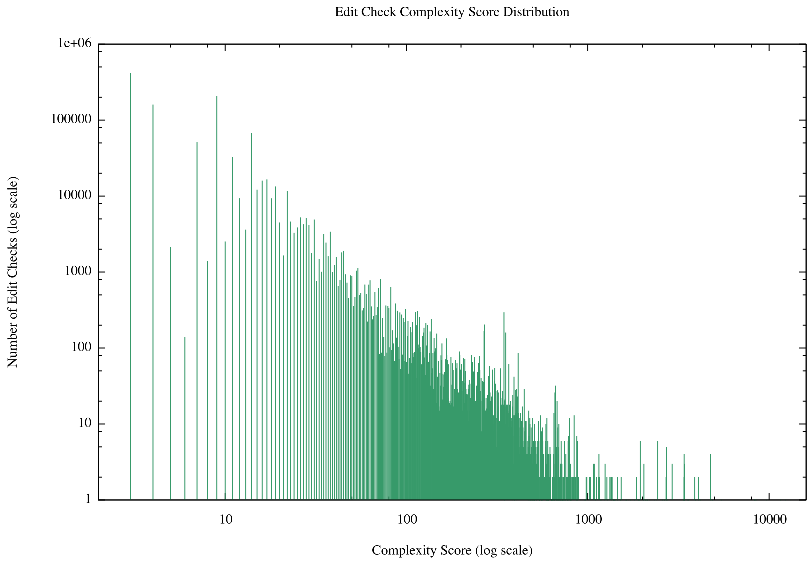
Of the 1.1 billion data values in our dataset only 60 million (5%) have been changed since initial data entry with the majority of data values (95%) being entered once and never changed. The Data Entry Constraints and Restricted Value Sets described earlier will have contributed to the high level of unchanged data since they prevent the entry of data that might be erroneous at the time of data entry.

DATA QUERIES

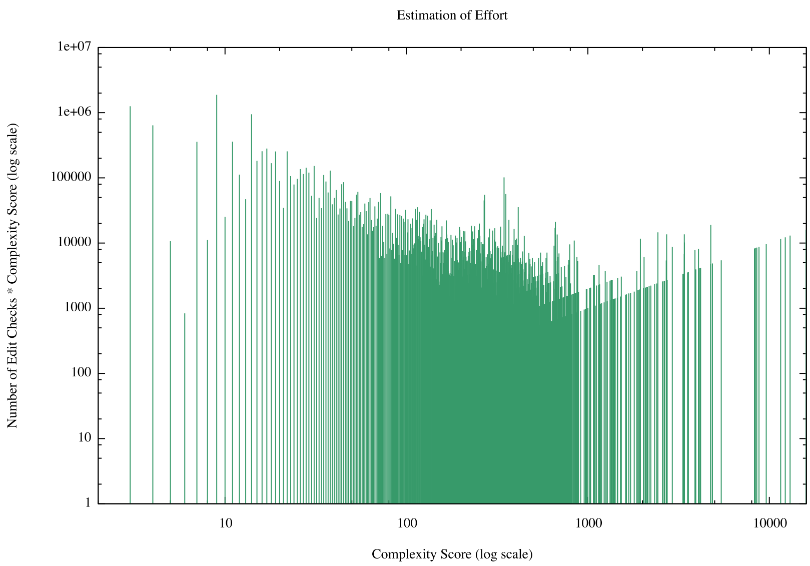
29 million data queries were created as a result of edit checks with a further 5 million manual queries being raised by users of the EDC system.

EDIT CHECK COMPLEXITY

The following chart shows the distribution of edit checks by complexity; the majority of checks (70%) having a complexity score less than 10 and 95% less than a score of 100. The chart is shown with log-log axes to show the tail of highly complex checks.



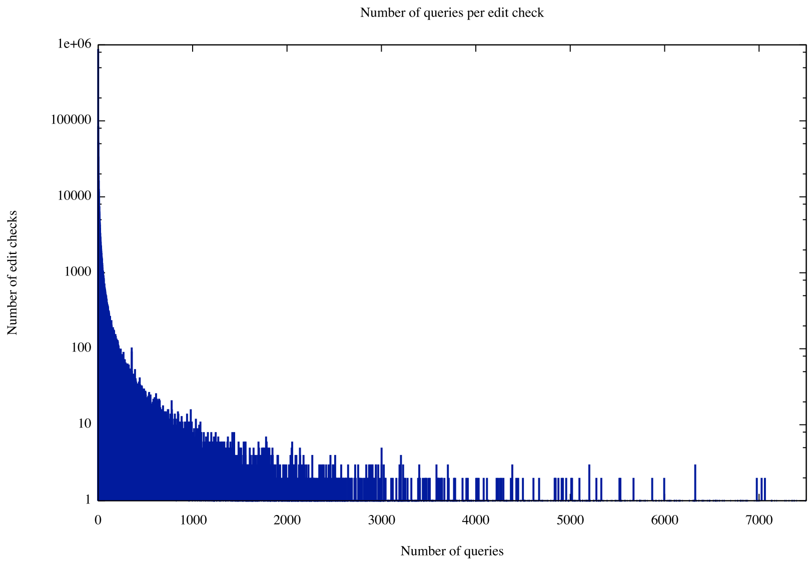
Using the complexity score as an estimate of effort to construct and test edit checks we multiply the number of checks for each complexity score by the complexity score to estimate the total effort. This graph shows that the more complex edit checks have a disproportionate impact on total effort; the edit checks with a complexity score greater than 100 (5% of the total number of edit checks) consume 25% of the overall effort.



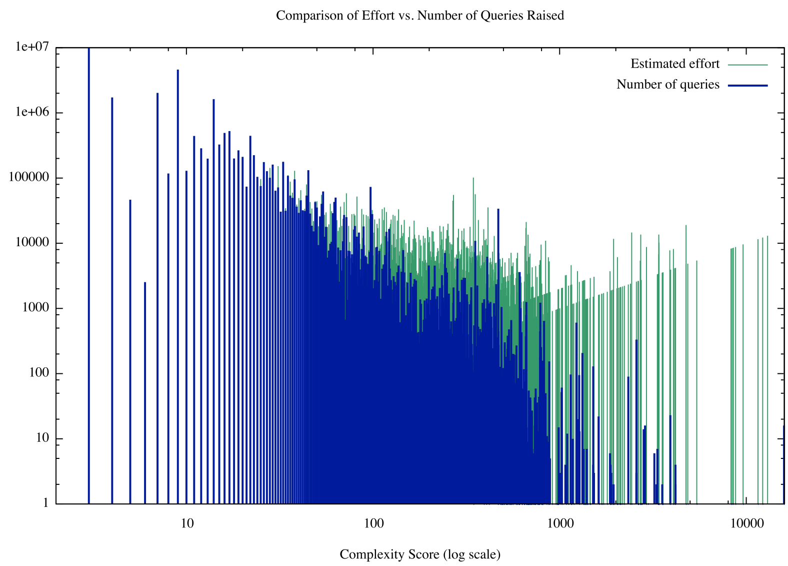
Note that the apparent trend lines on the right of the chart (complexity scores greater than 1,000) are due to the small number of edit checks at this end of the scale, and the large score, ie. each line on the right hand side of the chart represents a single, complex edit check.

EDIT CHECK IMPACT

The number of queries raised for each edit check shows that 675,966 (59.5%) of edit checks did not raise any data queries at all and so had no direct impact on data quality; no data was changed as a result of these edit checks.



We can compare the number of queries raised for each level of complexity score with the estimated effort calculated earlier. This chart shows that the number of queries raised declines as the complexity of the check increases, probably because the data criteria that need to be satisfied to trigger the check are less likely to occur:



DISCUSSION

ARE EDIT CHECKS NECESSARY?

95% of the data values in our analysis dataset were not changed from initial entry, ie. the values are acceptably clean without the application of edit checks. This is an encouraging metric but leaves a significant proportion of data which has been changed and a potential error rate that is unacceptably high; clearly edit checks have played a role in identifying which data needed to be changed. Electronic data collection without edit checks is unlikely to lead to an acceptable data error rate.

CAN WE SCALE BACK EDIT CHECKS?

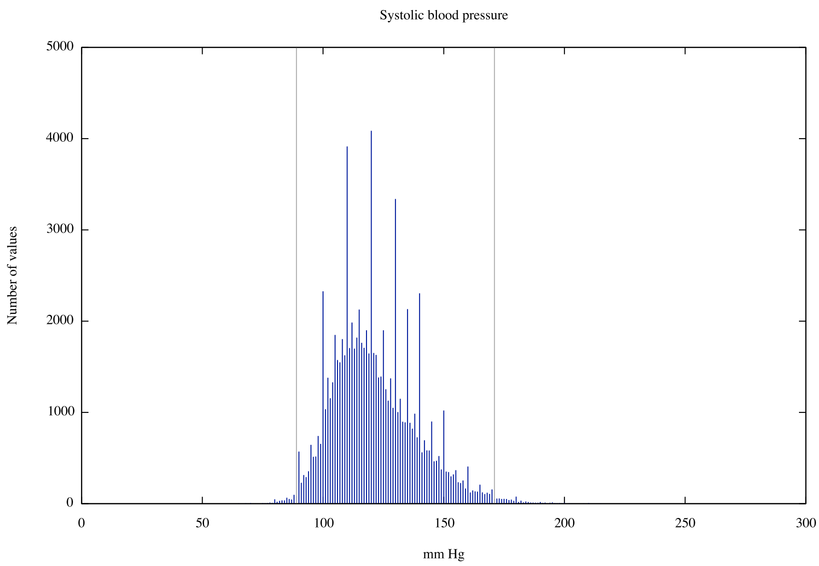
The majority of edit checks analysed (60%) have no impact on data quality because the conditions to trigger the edit check, and therefore to raise a data query, have not been met. A further 20% of the total number of edit checks raised a small number (1-6) of data queries.

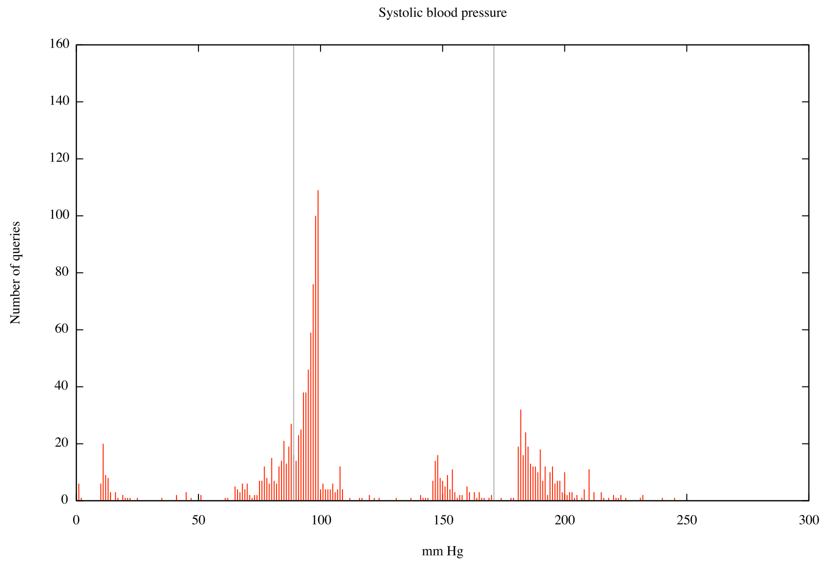
One method to identify unnecessary edit checks that are unlikely to improve data quality is to look at historical data. For example we found an edit check with complexity score of 87 that had been used in 231 clinical studies by one sponsor. This edit check had raised no queries in any of those studies and clearly could be dropped from future studies with negligible risk to data quality. This edit check was one of many similar examples that we found.

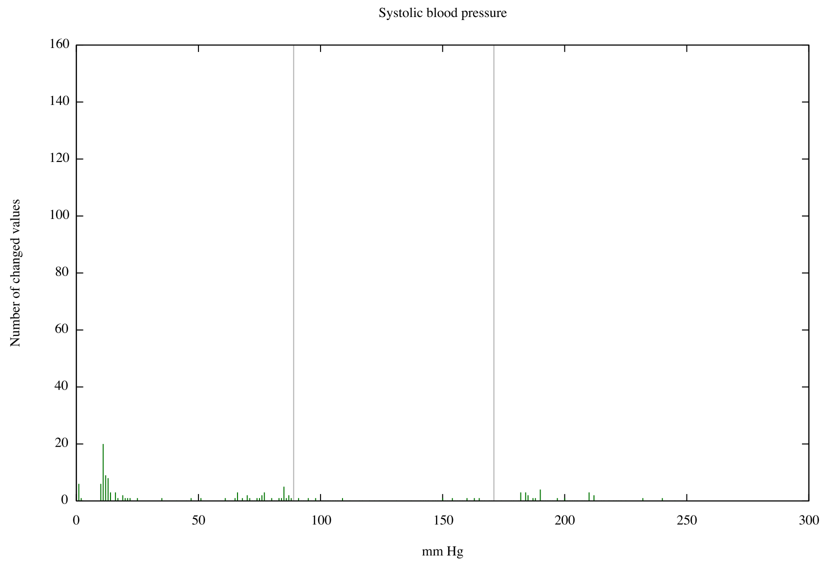
CAN WE USE DATA TRENDS TO OPTIMIZE EDIT CHECKS?

We examined data values and edit checks for Systolic Blood Pressure for 290 studies conducted by one sponsor. We found 84,641 data values with 1,306 queries which caused 130 data changes (0.15% of the total).

The distribution of data values, queries and data changes are shown below, with vertical lines indicating the 1st and 99th percentile values (note the peaks at values of 100, 110, 120, etc. due to the accuracy of typical Sphygmomanometer scales):



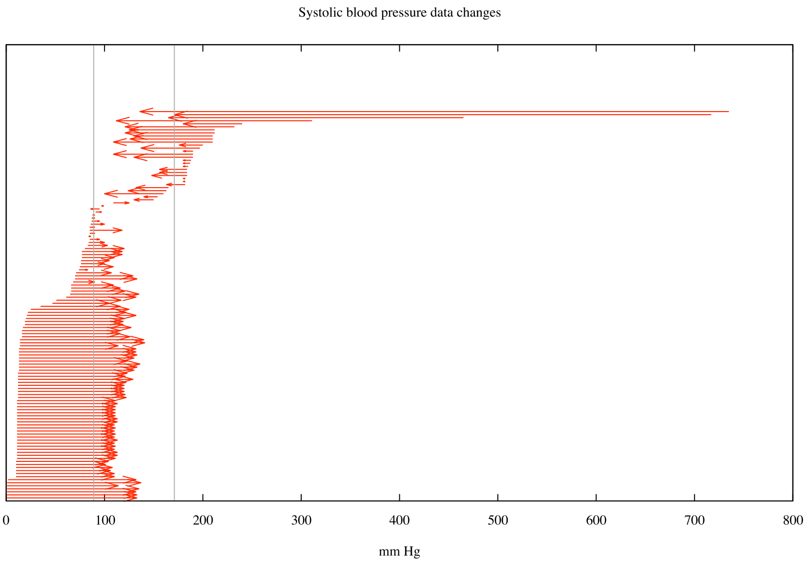




From the charts we can see that most (90%) queries did not lead to data changes but will have required effort from investigator site users to respond and data managers to close. By setting the edit check ranges at the 1st and 99th percentiles these false positives could have been avoided, with very little impact on data quality. There were 8 data changes within the 1st and 99th percentiles, ie. 0.01% of the total dataset. Such a small number of changes is very unlikely to affect the results of the trial; the actual data changes are shown below:

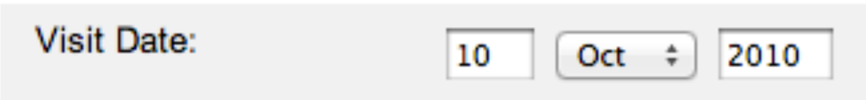
Original value	Value after query
91	97
95	86
98	99
109	125
150	130
154	140
160	100
163	124

Changes to data values outside the 1st and 99th percentiles were often large because the original values were clearly impossible:

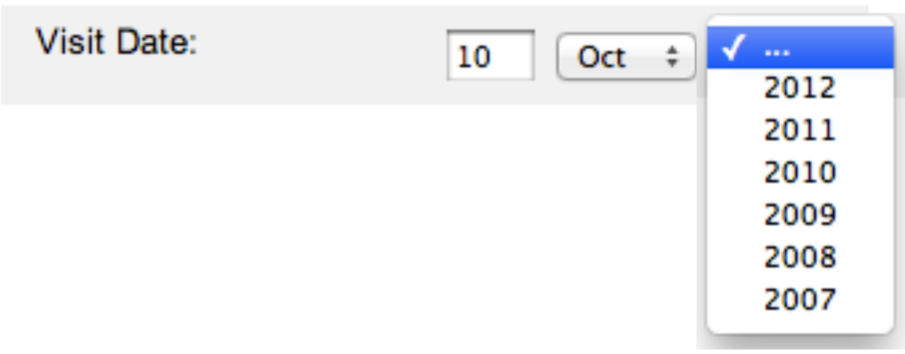


CAN DATA COLLECTION FORM DESIGN OR EDC SYSTEM DESIGN REDUCE ERRORS?

Earlier in this paper we looked at how data entry constraints and data typing can help avoid data errors by preventing invalid entries. The EDC system used in this analysis constrains date fields to be structured as day (2 character field), month (constrained value set) and year (4 character field):



A common, and simple, edit check is to query future dates, for visit dates, dates of birth, etc. In our dataset we found 297,720 checks of this type, which caused 266,398 queries to be raised.
An alternative form design which restricted the possible values for year would reduce the number of queries:



We estimate that this would reduce the number of future date queries by up to 20%. While the effort of constructing such checks is not large there would be a significant reduction in effort to answer and close these queries.

CONCLUSION

Edit checks are necessary to ensure data quality reaches acceptably high levels. However many edit checks have little or no impact on data quality and with careful review the same data quality could be achieved with a reduced number of edit checks. Some edit checks are overly restrictive and lead to many false positive queries; less restrictive checks would reduce the burden of query management without compromising data quality. Finally, changes to data collection forms could help to avoid data errors.

[ii] Draft Guidance for Industry
Oversight of Clinical Investigators – A Risk-Based Approach to Monitoring
http://www.fda.gov/downloads/Drugs/.../Guidances/UCM269919.pdf

[iii] Quantifying Data Quality for Clinical Trials Using Electronic Data Capture
Meredith L Nahm, Carl F Pieper, Maureen M Cunningham
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0003049

[iv] Code Complete, A Practical Handbook of Software Construction by Steve McConnell, Second Edition, Microsoft Press, 2004