
Optimizing Data Validation

Andrew Newbigging



Optimizing Clinical Trials:
Concept to Conclusion™

Acceptable data quality in clinical trials

... 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%

Draft FDA guidance

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.



How do Electronic Data Capture (EDC) systems help achieve acceptable data quality?


Data constraints

Blood Pressure:		
Systolic Blood Pressure:	<input type="text" value="120"/>	mmHg
Diastolic Blood Pressure:	<input type="text" value="80"/>	mmHg

Data type conformance

Blood Pressure:

Systolic Blood Pressure:  
 mmHg

Diastolic Blood Pressure: 80 mmHg 

Restricted value sets

Outcome:

- ...
- Recovered / Resolved
- Recovering / Resolving
- ✓ Not Recovered / Not Resolved
- Recovered / Resolved with Sequelae
- Fatal
- Unknown

Edit checks

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

```
if SystolicBloodPressure < DiastolicBloodPressure  
  then OpenQuery
```


Edit check complexity score

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

Complexity score = 3

Test data	Result
180	No query
181	Query

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

Complexity score = 9

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

Highest complexity score = 15,495

Dataset for analysis

Production data from 300+ sponsors' trials:

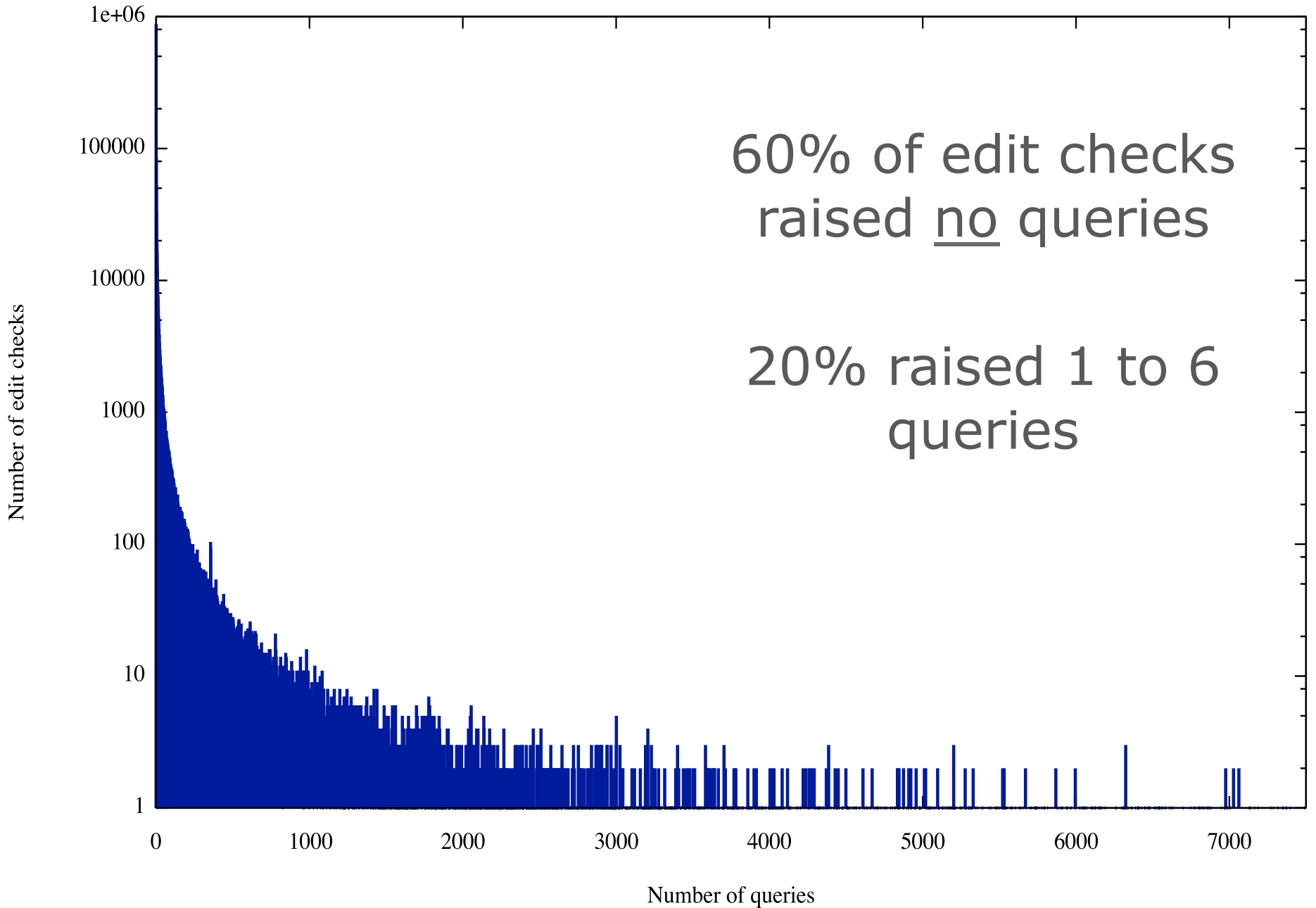
- Global pharma, CRO, biotech, academic
- Phase I, II, III, IV, post-marketing
- Americas, Europe, Asia Pacific
- English, Japanese and Chinese data entry

Dataset for analysis

Number of datapoints	1,160,836,888
Number of edit checks	1,137,496
Number of queries	29,255,296

95% of data values have not been changed
from initial entry

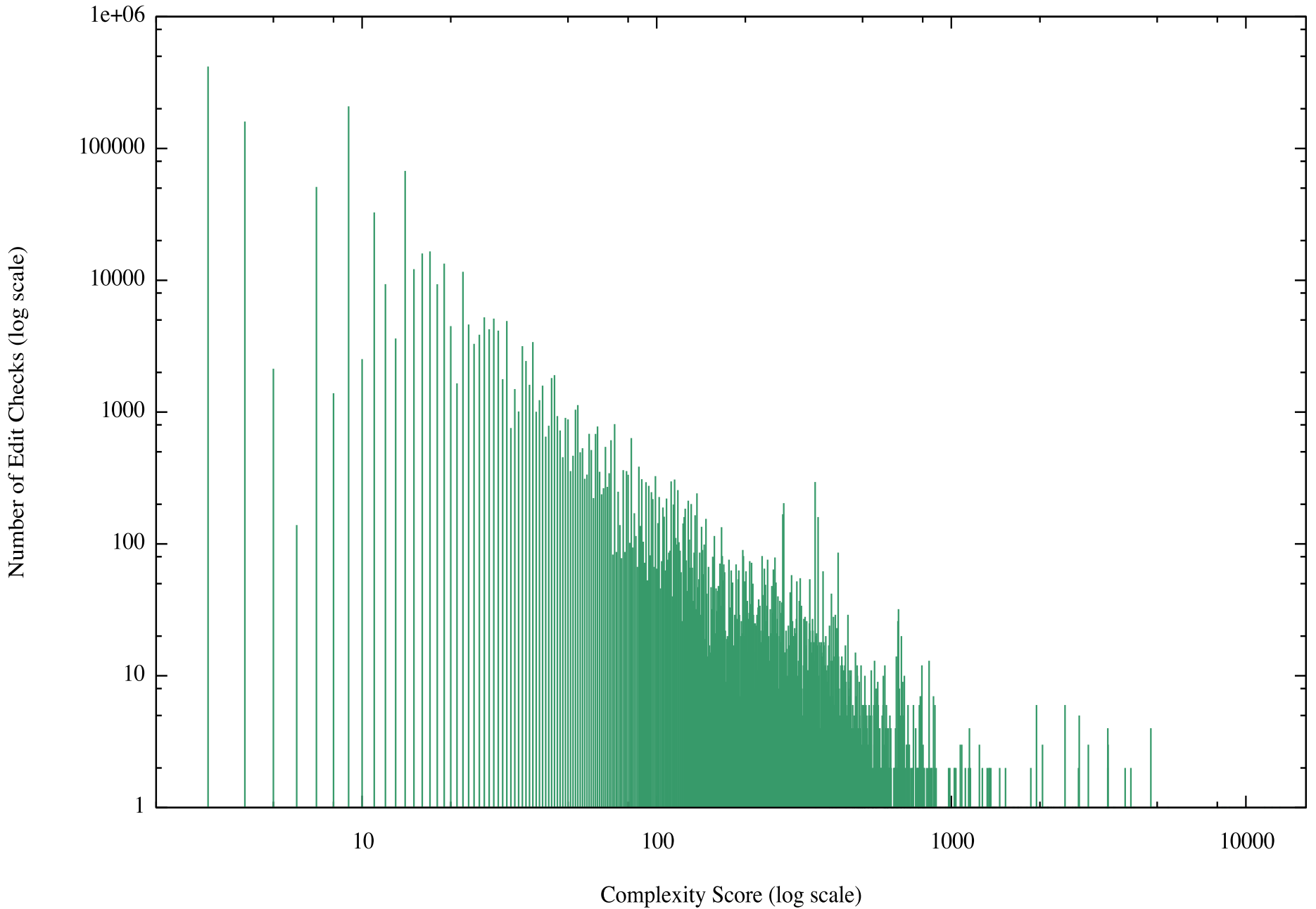
Number of queries per edit check



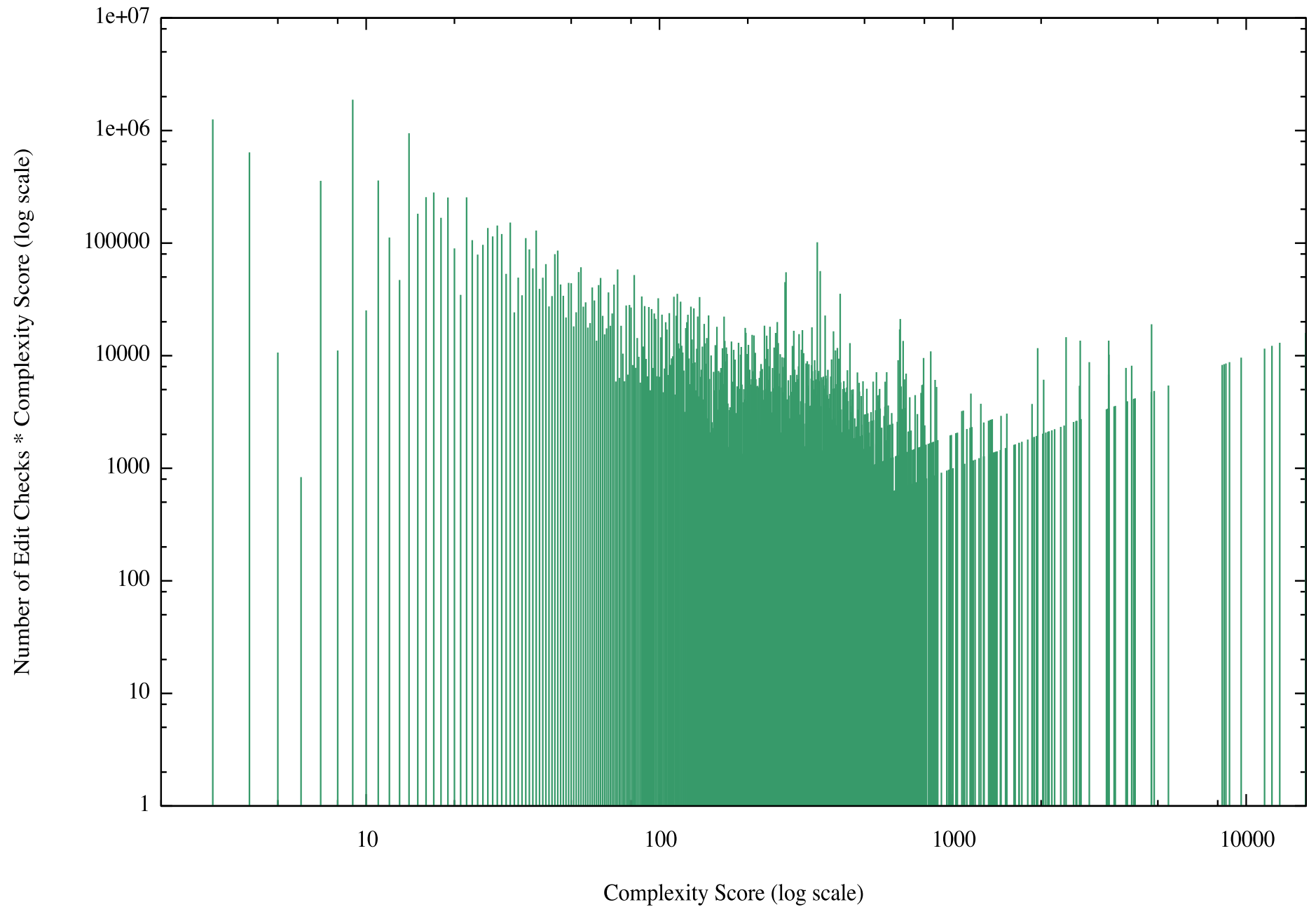
One sponsor copied the same edit check
(complexity score = 87) across 231 studies

No queries have ever been raised by the edit
check

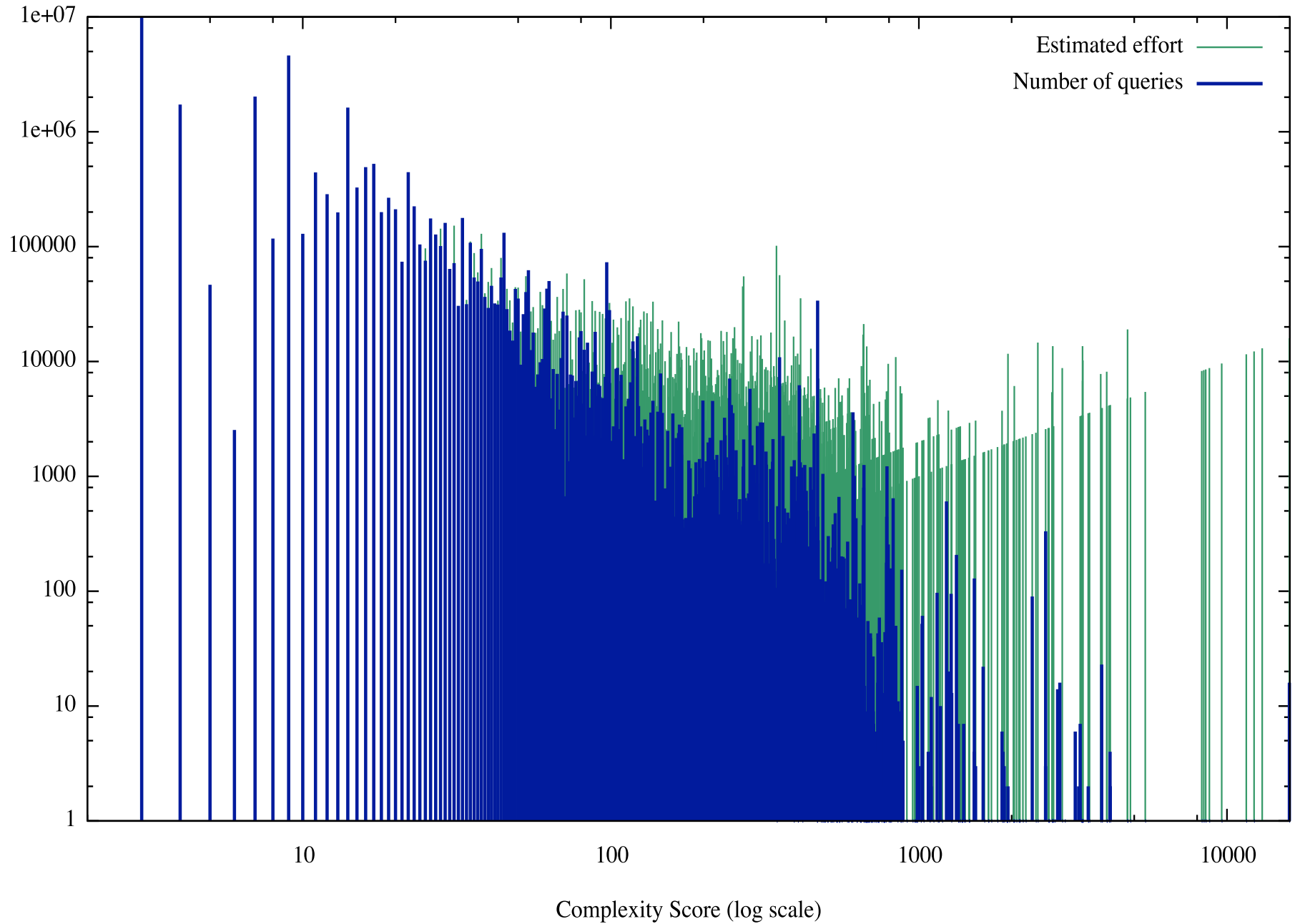
Edit Check Complexity Score Distribution



Estimation of Effort



Comparison of Effort vs. Number of Queries Raised

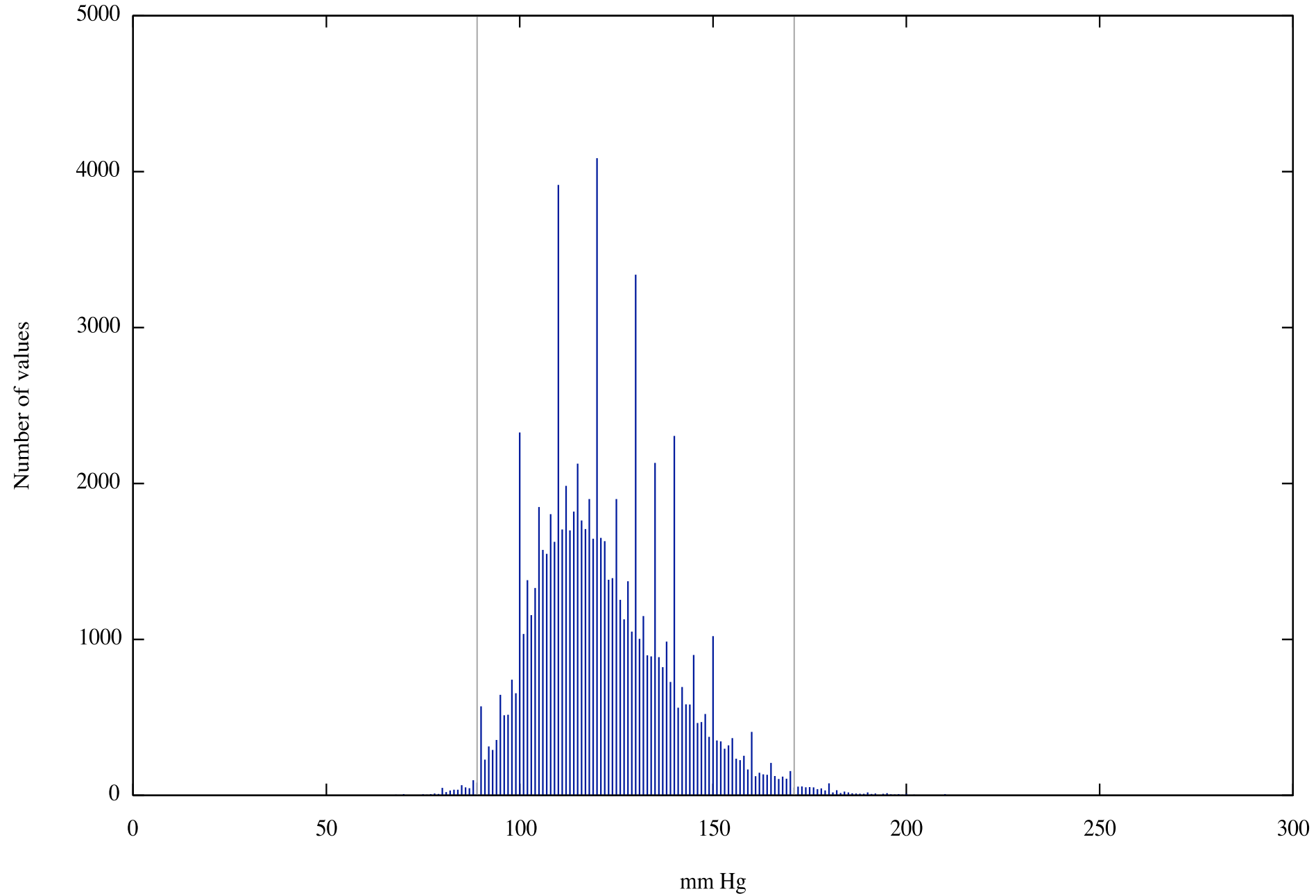


One sponsor, multiple studies

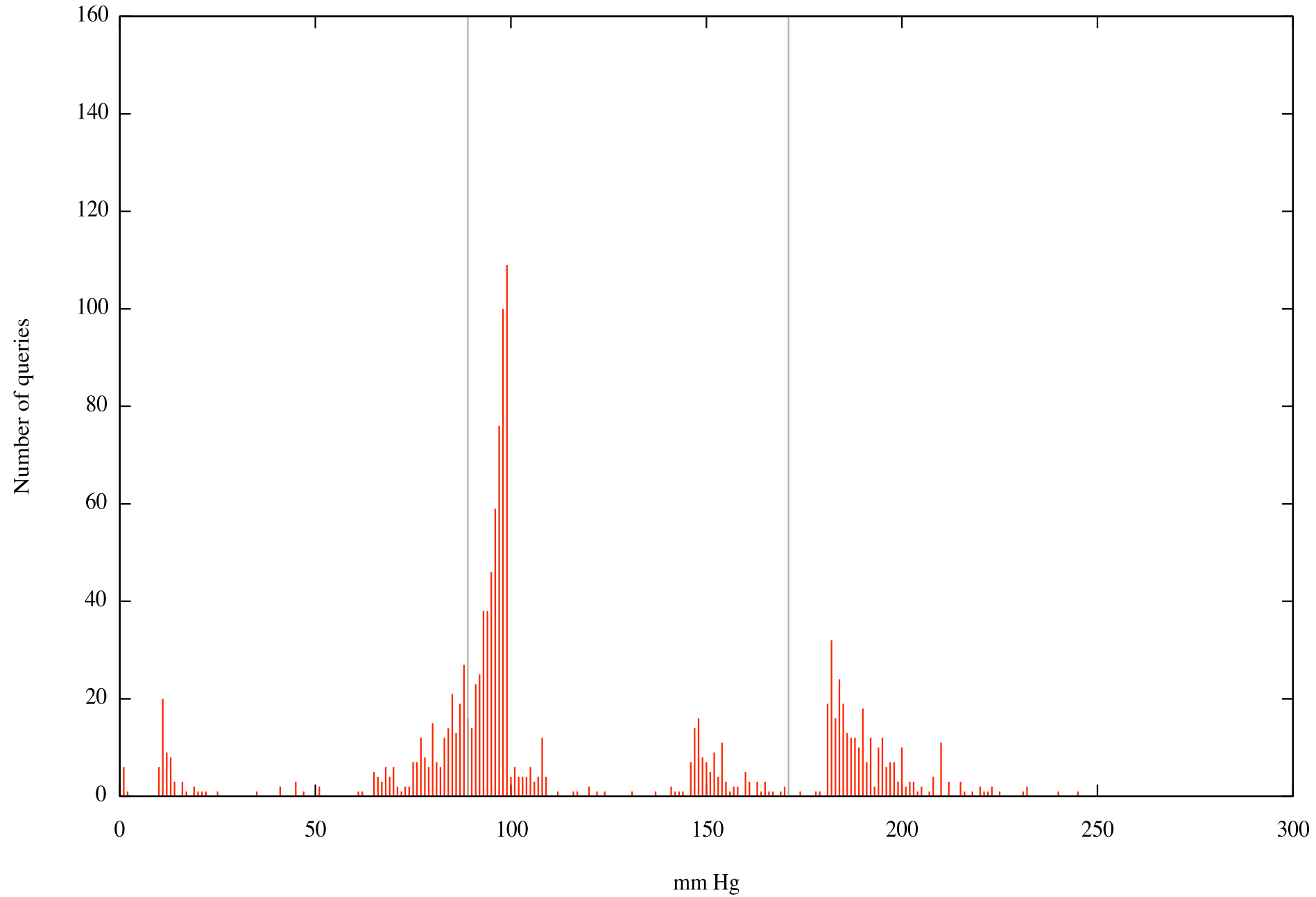
Systolic blood pressure

Number of datapoints	86,641
Number of queries	1,306
Number of data values changed	130

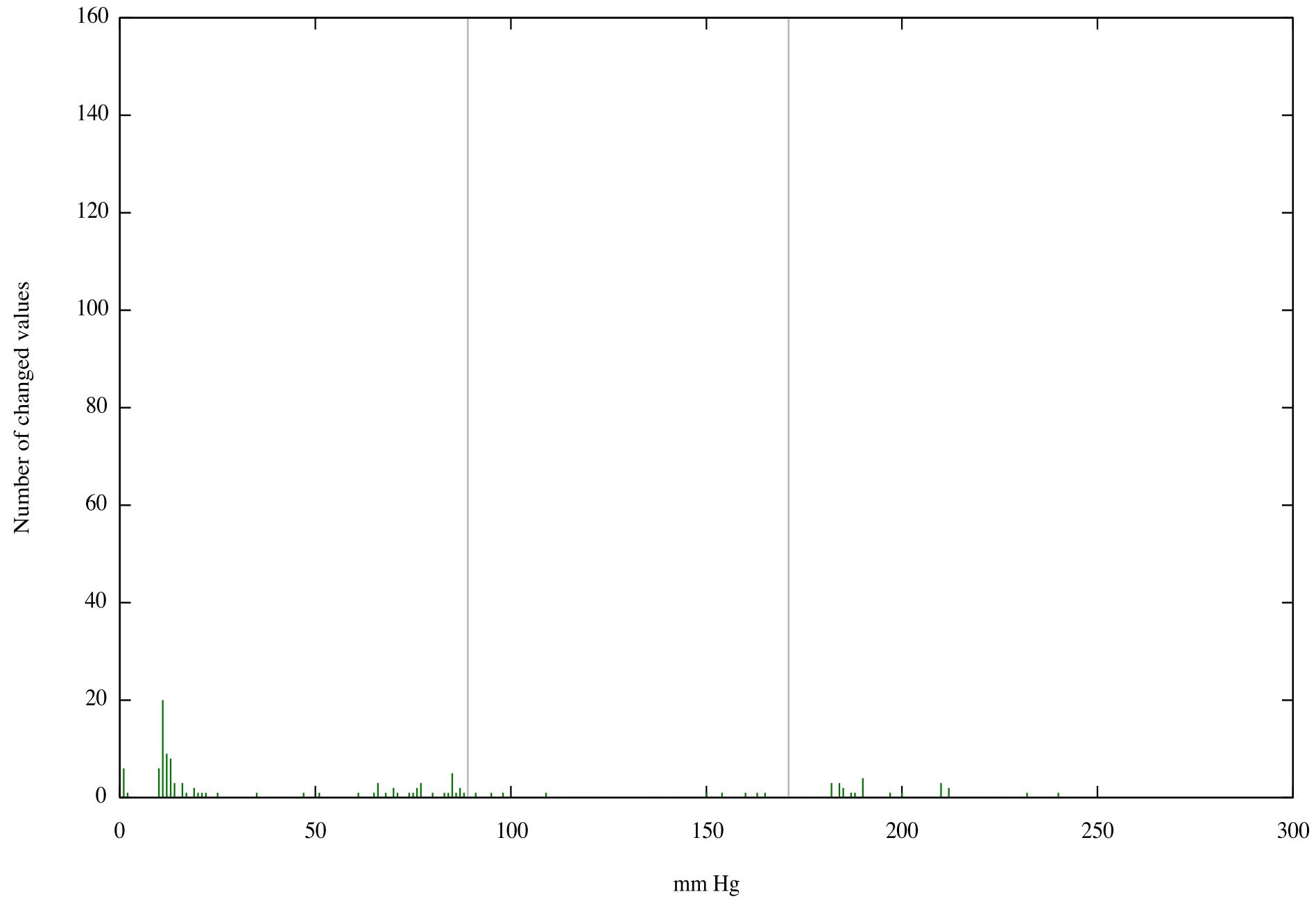
Systolic blood pressure



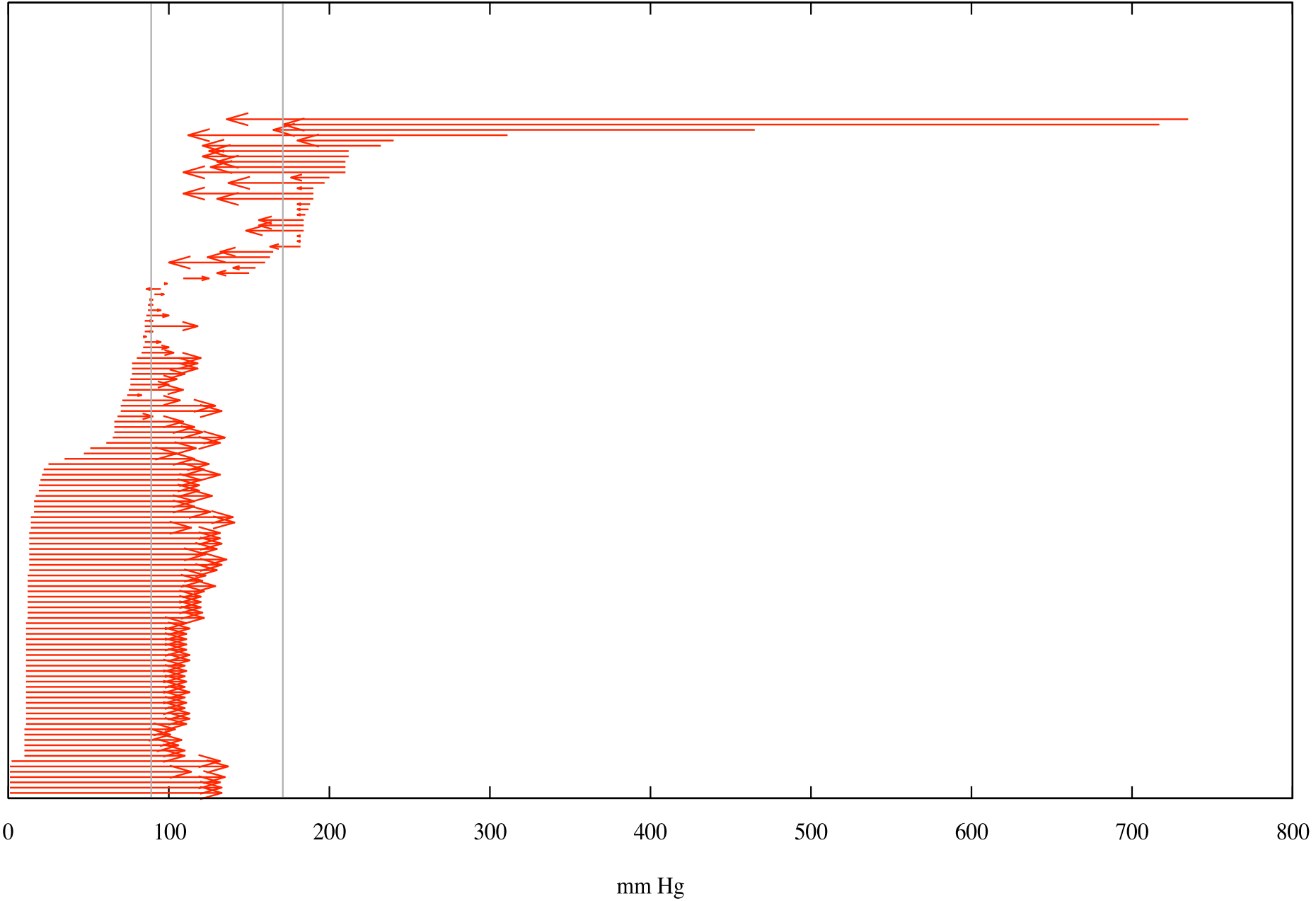
Systolic blood pressure



Systolic blood pressure



Systolic blood pressure data changes



Form design

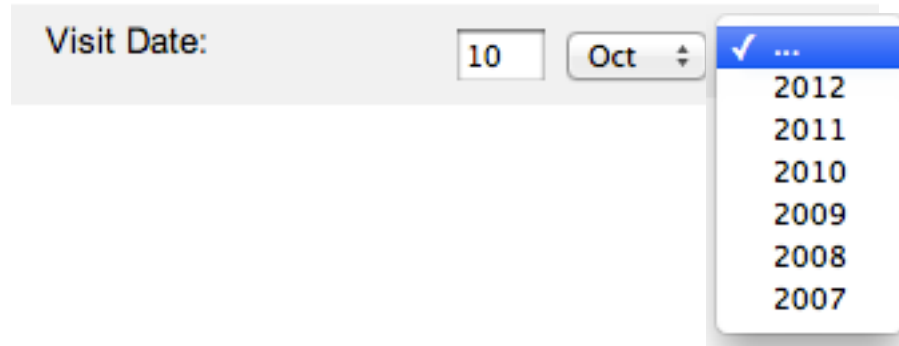
Visit Date:

`VisitDate <= Today`

297,720 edit checks

266,398 queries raised

Improving form design



Visit Date:

- 2012
- 2011
- 2010
- 2009
- 2008
- 2007

The image shows a form field labeled "Visit Date". It consists of three input boxes: a text box containing "10", a text box containing "Oct" with a small dropdown arrow, and a text box containing "...". A dropdown menu is open from the third box, listing the years 2012, 2011, 2010, 2009, 2008, and 2007. The top item in the menu, "...", is highlighted with a blue background and a checkmark.

Changing form design would reduce queries by up to 20% (53,000 queries)

Conclusion

95% initial data entry accuracy is good but not good enough

Therefore edit checks are necessary

But many edit checks have little or no impact on data quality

Data analysis can help target edit checks more effectively and reduce false positive queries

Improved form design can reduce data queries