Innovative Open Source Approaches to Automating Clinical Data Management for a Large Multi-centre Cohort Study

Narshion Matai Ngao

Childhood Acute Illness & Nutrition Network, KEMRI Wellcome Trust Research Programme

25-11-2022

The Global Health Network Conference 2022, University of Cape Town, South Africa
Background

Data Management in Research Studies

• Do you agree with this?

Source: twitter.com random user post
Background

Data Management in Research Studies

• Do you agree with this?

Source: twitter.com random user post

• Accurate, Reliable and Statistically sound data for analysis and decision making
Background

CHAIN Network Data Issues Experience

- A Multi-centre, longitudinal (multi-event) cohort study

- We experienced these data related issues: -
  1. CRF design issues & database laxity – allowing errors to occur. (unforeseen at first)
  2. Inconsistent values entered (wrong dates, abnormal measurements)
  3. Missing data
  4. Programmer issues – formatting, conversion errors, data type mapping, renaming issues
  5. Time to chase all these in a multi-centre environment
Background

Unique Challenges of a Multi-centre Study

Complexity of set up – clinical, social, laboratory data

Infrastructural Environment – different capacity at sites

Data Cleaning turn-around - Query timings

Lack of standards/Standardization – Consistency of approach

Data Reporting – Single point of truth
Methods

What we need to see happen

Monitor real time data quality checks.

Run data cleaning queries & post them automatically.

Access to Standardized Study Reports

Improve productivity of data team
Methods

Solution: Data Management as an Embedded Service

Study Process

1. Labelling and grouping of questions on CRF
2. Database validations, required and range checks
3. 24hr data entry policy

Data Collection & Data Entry

1. Second clinician verification before entry
2. Real time data visualization dashboard
3. Automated queries and query tracking via issue management system

Extract, Transform, & Load

1. Data types, formatting and conversion errors
2. Source Document Verification
3. Data curation process – transformation and loading for analysis

Data Analysis & Reporting

1. Data visualization
2. Numeric data cleaning techniques
3. Categorical data cleaning techniques
4. Pre-analysis & targeted data cleaning cycles.

Data Cleaning

Detection

1. Missing data
2. Duplicated data
3. Outliers and inconsistencies
4. Strange patterns
5. Suspected wrong results
6. Exploratory data analysis
7. Data Visualization

Diagnosis

1. Confirm errors, missing data
2. Central tendency and dispersion techniques to verify outliers and inconsistencies
3. Data distributions
4. True extremes, true normal
5. No diagnosis, still suspicious data

Data Editing

1. Correct data
2. Delete
3. Leave unchanged
4. Document action

Automated Queries (Real-time Dashboards)

Interactive Reports/Curation (Data Cleaning Sprints using SCRUM)

Task Management (Query Resolution workflow)
Methods

How We Did It
Methods

Our Tools of Trade

- REDCap is a mature, secure web application for building and managing online surveys and databases.

- Kilifi Integrated Database Management System. Primary storage for laboratory meta data and test results.

- Rstudio and Shiny: Great combination of both packages & frameworks to produce a dashboard.
Methods

**Benefits of our approach**

1. A sub-set of risk-based approaches to data management.
2. Real-time check on both item and variable missingness.
3. Up-to-minute capturing of data inconsistency and anomaly detection.
4. Standardized yet interactive study progress reports (accreditation curves, lost to follow up and missed visits summaries).
5. Customized utility reports for planning future events such as enrolment targets.

![CIN-NeOBAC IDs without corresponding clinical data](image_url)
Methods

Benefits of our approach

1. A sub-set of risk-based approaches to data management.
2. Real-time check on both item and variable missingness.
3. Up-to-minute capturing of data inconsistency and anomaly detection.
4. Standardized yet interactive study progress reports (accrual curves, lost to follow up and missed visits summaries).
5. Customized utility reports for planning future events such as enrolment targets
Methods

Benefits of our approach

1. A sub-set of risk-based approaches to data management.
2. Real-time check on both item and variable missingness.
3. Up-to-minute capturing of data inconsistency and anomaly detection.
4. Standardized yet interactive study progress reports (accrual curves, lost to follow up and missed visits summaries).
5. Customized utility reports for planning future events such as enrolment targets
Methods

Benefits of our approach

1. A sub-set of risk-based approaches to data management.
2. Real-time check on both item and variable missingness.
3. Up-to-minute capturing of data inconsistency and anomaly detection.
4. Standardized yet interactive study progress reports (accrual curves, lost to follow up and missed visits summaries).
5. Customized utility reports for planning future events such as enrolment targets

<table>
<thead>
<tr>
<th>2730</th>
<th>2607</th>
<th>408</th>
<th>1248</th>
<th>978</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neonatal admissions</td>
<td>Blood cultures done</td>
<td>Inpatient deaths</td>
<td>D28 consented</td>
<td>Alive at D28 (Kilifi &amp; Mbagathi)</td>
</tr>
</tbody>
</table>
Methods

Benefits of our approach

1. A sub-set of risk-based approaches to data management.
2. Real-time check on both item and variable missingness.
3. Up-to-minute capturing of data inconsistency and anomaly detection.
4. Standardized yet interactive study progress reports (accrual curves, lost to follow up and missed visits summaries).
5. Customized utility reports for planning future events such as enrolment targets
# Results and Conclusions

<table>
<thead>
<tr>
<th></th>
<th>Seamless data extraction and processing with no human intervention.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Run automated cleaning scripts and monitor data quality checks in real time.</td>
</tr>
<tr>
<td></td>
<td>Up-to-minute access to standardized study reports by sites’ PI and study sta.</td>
</tr>
<tr>
<td></td>
<td>Real time study progress reports helped planning of recruitment and other activities.</td>
</tr>
<tr>
<td></td>
<td>Improved productivity of the data team and reduced over-reliance on the central team</td>
</tr>
</tbody>
</table>
Results and Conclusions

Recommendations

For a Multi-centre Study: -

1. Standardizing data management practice across all participating sites
   • Ensure a uniform data collection and management environment for all the sites.

2. Database for data collection should be hosted online (via internet) so that it is the same or if using offline data capture, then data should be synced to the same online server as soon as possible.

3. Real-time (or almost) data monitoring and query resolution dashboards be implemented and hosted online.
   • Cut back data managers time spent on repetitive tasks & standard reports.

4. Participating sites to take up consumption of the centralized dashboards whenever they’re availed.
Acknowledgements

Special thanks to the patients, staff and leadership at our clinical sites.