Using Open Source Approaches to Automate Research Data Management

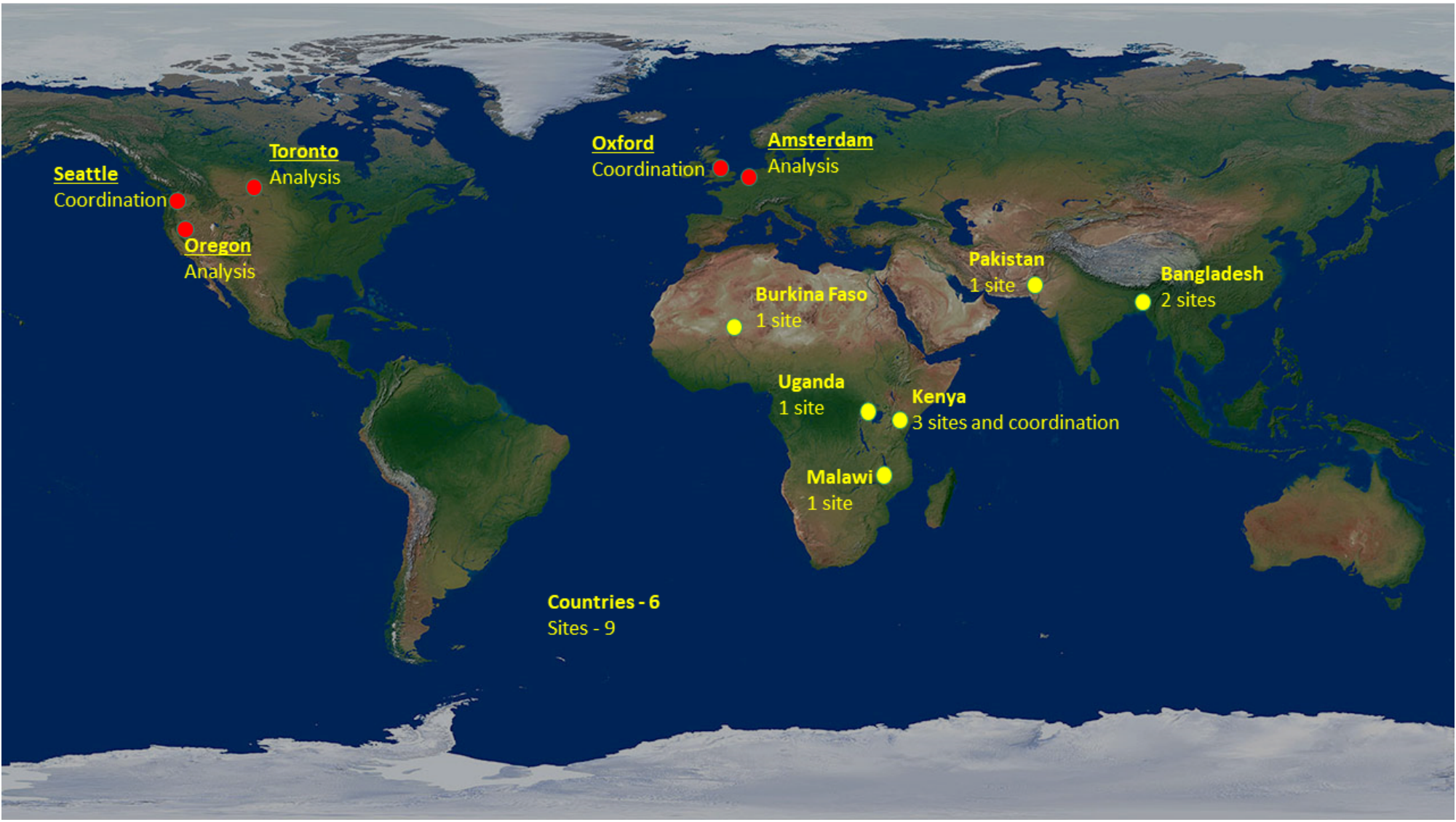
Christopher Maronga

January 7, 2021

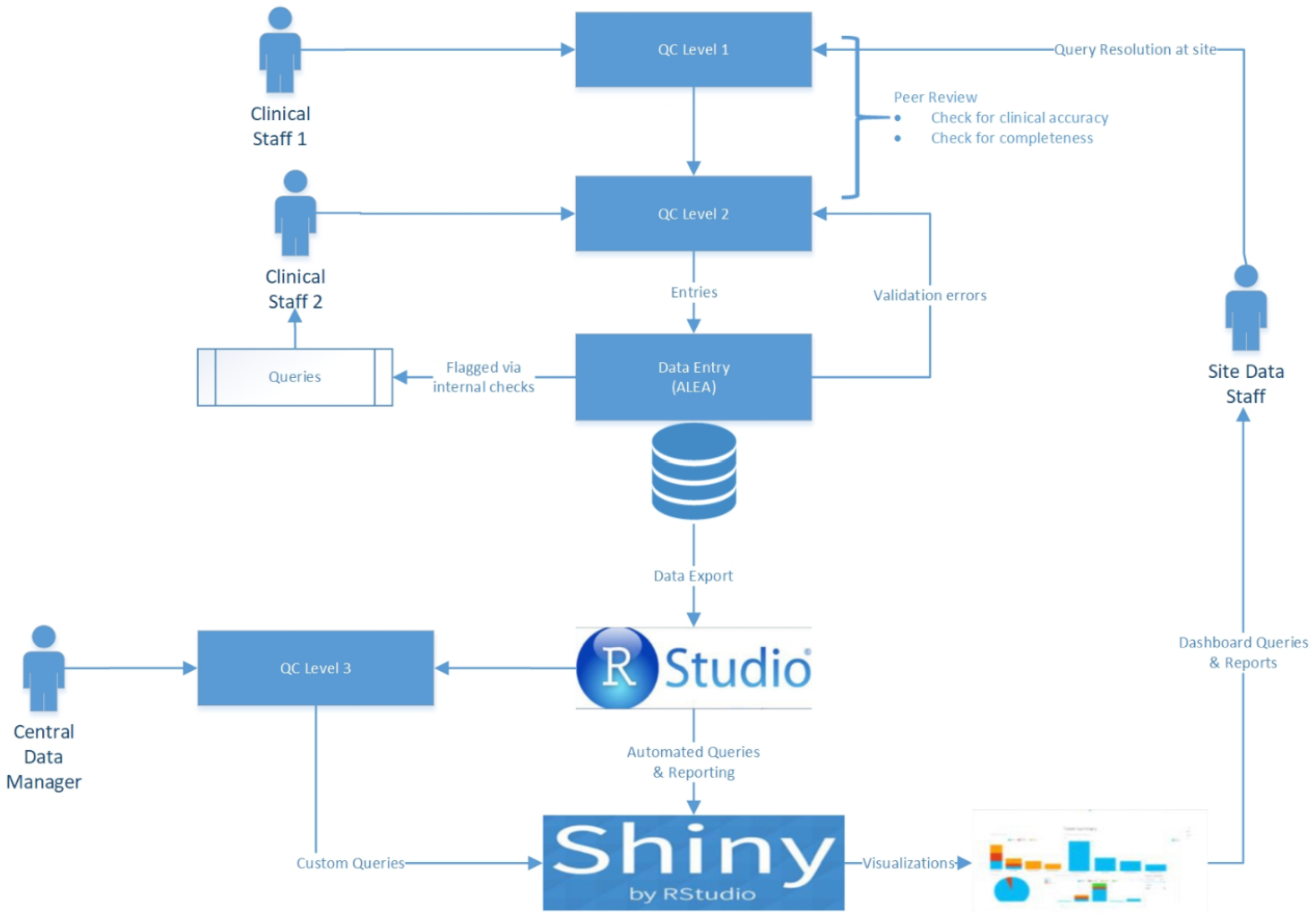
Contents

1. Introduction
2. Proposed QC model
3. Implementation
4. Fitting the pieces together
5. The Gains
6. The reality check
7. The Challenges
8. Future thoughts

- All health research involves collecting, managing and processing of study data.
- Data management can be costly and unaffordable in developing countries.
- Good data management is important and beneficial to the success of any study.
- Potential data management issues.
  - Data from different sources
  - Complex study designs (longitudinal)
  - Multiple individuals conducting data entry
- Open source technologies can be used to monitor real time data quality checks and up-to-minute study progress.


Seattle
Coordination
Toronto
Analysis
Oregon
Analysis
Oxford
Coordination
Amsterdam
Analysis
Pakistan
1 site
Bangladesh
2 sites
Burkina Faso
1 site
Uganda
1 site
Kenya
3 sites and coordination
Malawi
1 site
Countries - 6
Sites - 9

- Main cohort (4,714 records, 1,563 variables over 23 unique eCRFs)
- Over seven sub-studies (2,658 records, 1,985 variables over 46 unique eCRFs)
- Variables on demographics, clinical features, mental health, socio-economic phenotyping etc.
- Lab meta-data and patients' test results (long data), approx. 0.95 million records on more than 80 non-processed variables.
- 48 hour turn-around time for data base update (KIDMS for lab data and REDCap for clinical data)


Clinical Staff 1
QC Level 1
Query Resolution at site
Peer Review
Check for clinical accuracy
Check for completeness
Clinical Staff 2
QC Level 2
Entries
Validation errors
Site Data Staff
Queries
Flagged via internal checks
Data Entry (ALEA)
Data Export
Dashboard Queries & Reports
Central Data Manager
QC Level 3
RStudio
Automated Queries & Reporting
Custom Queries
Shiny by RStudio
Visualizations

- A sub-set of risk based approaches to data management.
- Real-time check on both item and variable missingness.
- Up-to-minute capturing of data inconsistency and anomaly detection.
- Standardized yet interactive study progress reports (accrual curves, lost to follow up and missed visits summaries).
- Customized utility reports for planning future events and tracking follow up trajectories by sites.

- 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.

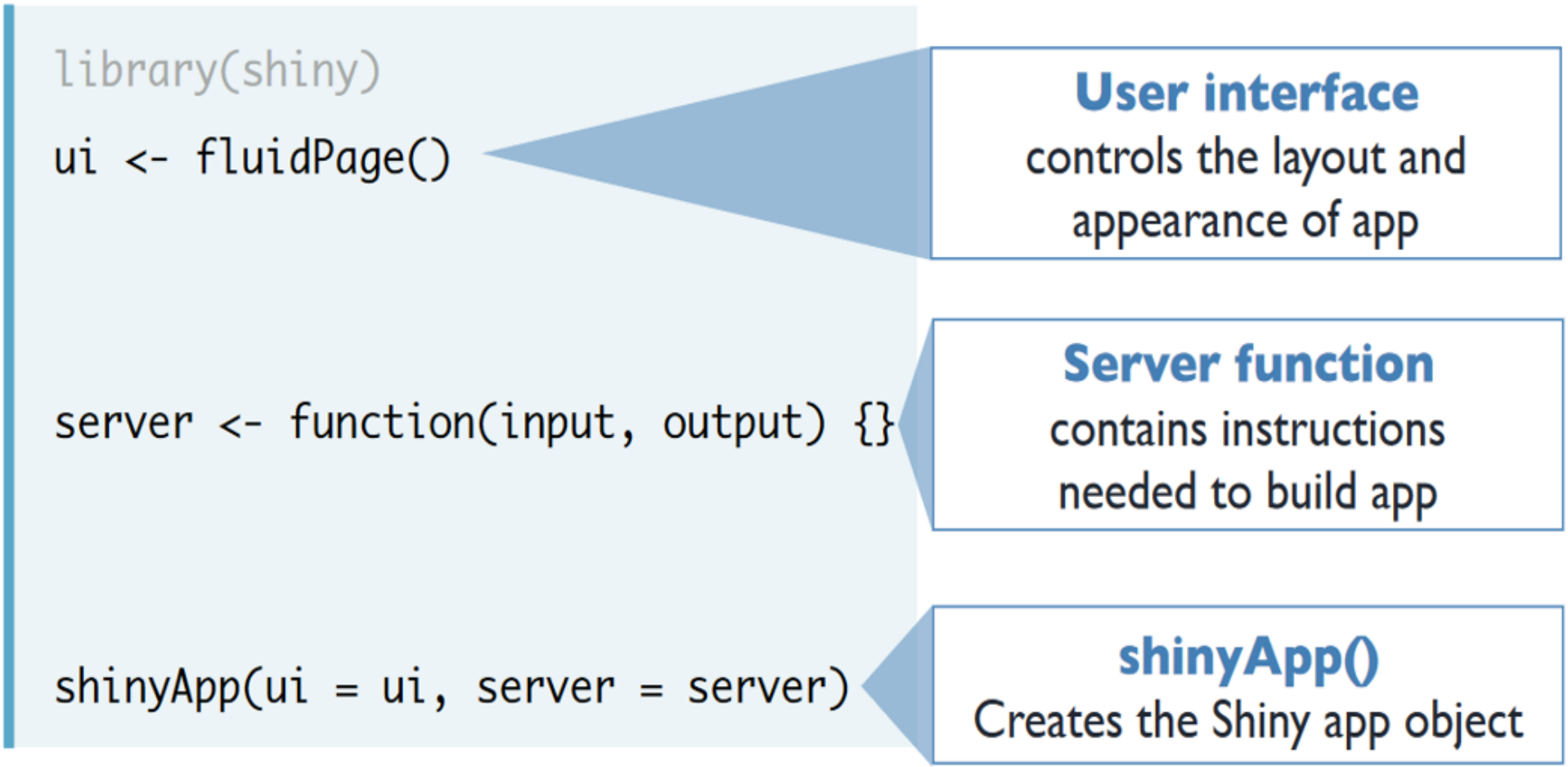
REDCap
Research Electronic Data Capture

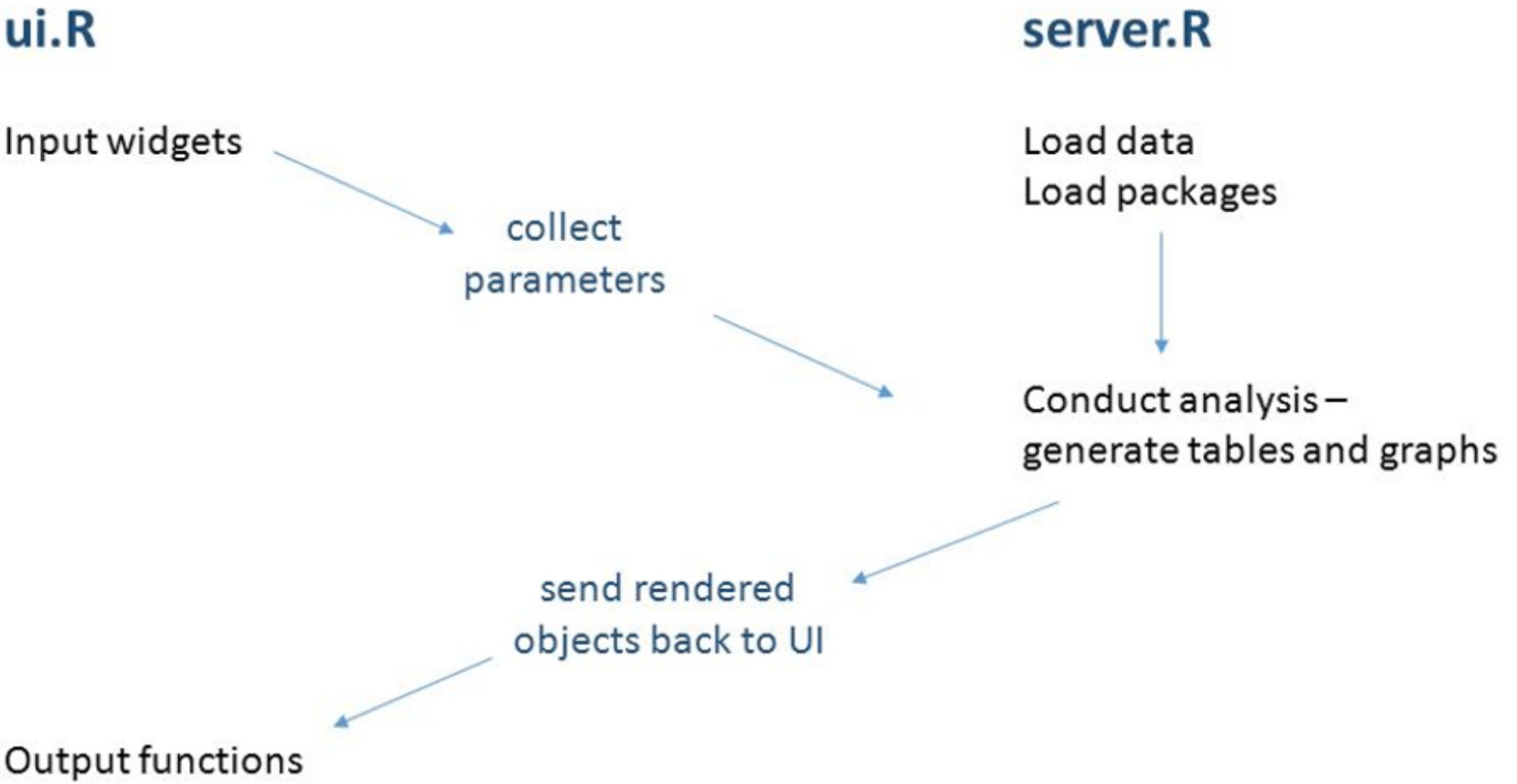
RDBMS (KIDMS)

RStudio

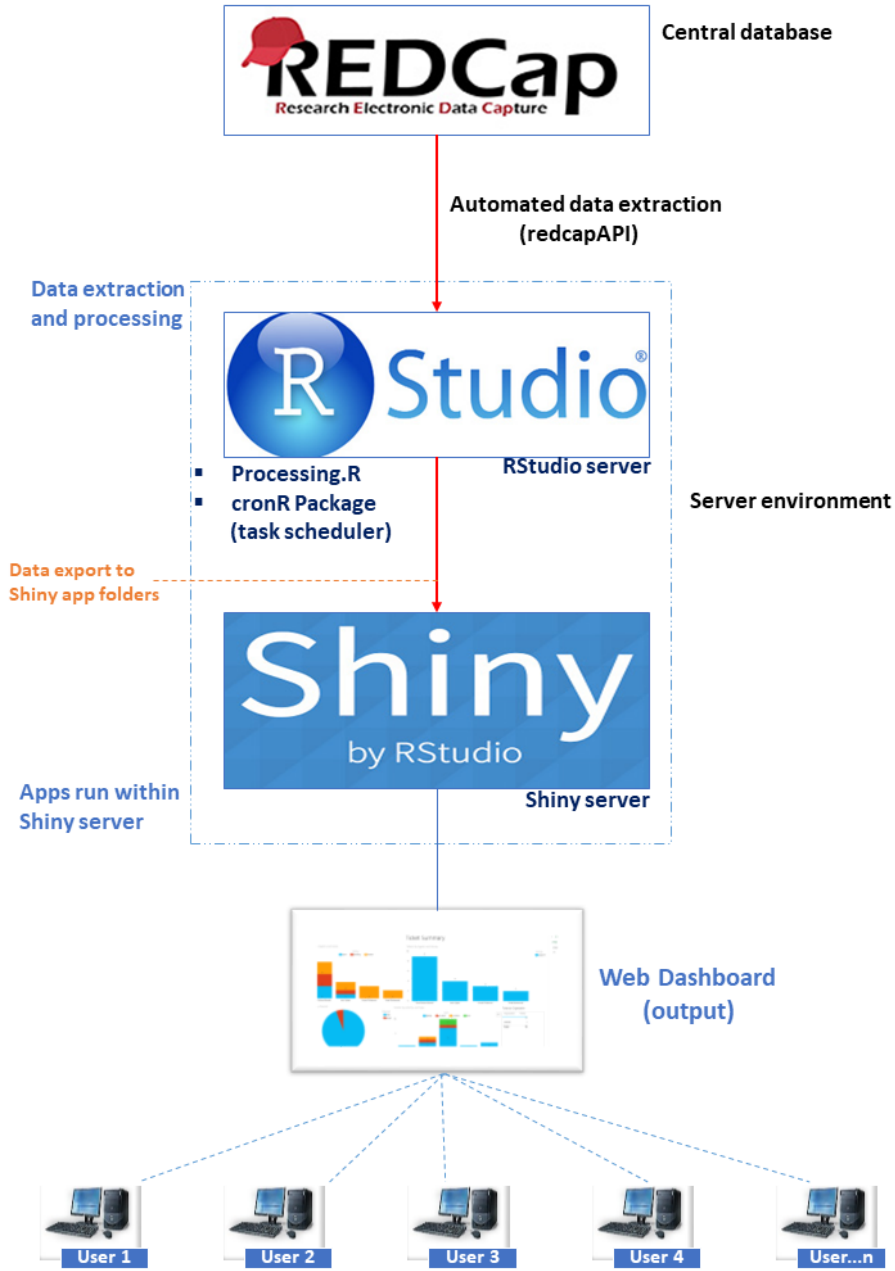
Shiny

The anatomy of a shiny application consist of user interface(ui) and the server section.




ui.R
server.R
Input widgets
collect parameters
Load data
Load packages
Conduct analysis – generate tables and graphs
send rendered objects back to UI
Output functions

- REDCap and KIDMS
  - Central data storage
  - Longitudinal design
  - KIDMS for lab data
- RStudio server
  - Task scheduling
  - Extraction
  - Processing
- Shiny server
  - Cleaning scripts
  - Reports authoring
- The Dashboard
  - Web-based
  - Reports and queries



CHAIN Network Reports Dashboard
Home
Banfora Samples
My Profile
Admin
Hi cmaronga, (Logout)
Clinical Reports (After 31-Jan-2019)
Launch
Clinical Reports (On/before 31-Jan-2019)
Launch
Laboratory Reports
Launch
Laboratory Storage Data
Launch
Clinical QC Reports
Laboratory QC Reports

Clinical Reports | Enrollment summaries | Follow up status | Survival curves | Study visits | Data completeness | Visualizations | About Dashboard

Recruitment summaries

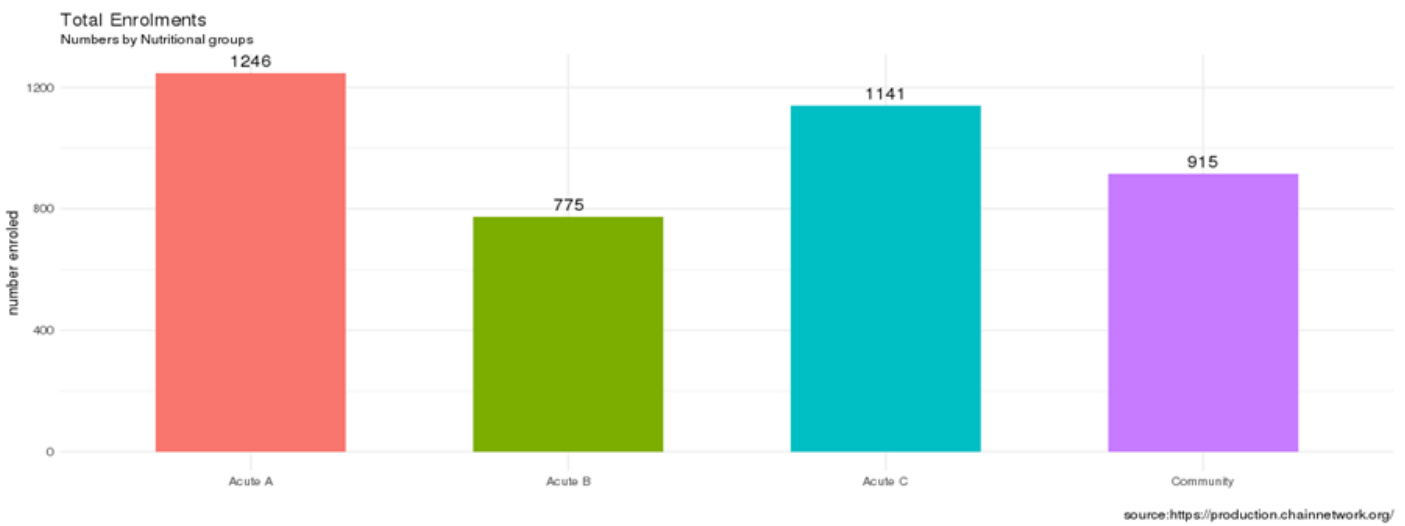
Select Date ranges

2016-11-01 up to 2019-02-11

Select site to view

All sites

Summary | ABC Groups | Enrolment targets | Age distribution | Monthly Enrollments



Missing CBC Results

Show 10 entries Search:

record_id	site	Date	event	test
No data available in table				

Participant discharge Queries | Verbal autopsy queries

Wrong discharge date and/or discharge date left hospital

Show 10 entries

record_id	site	disch_date	disch_date_left_hosp
No data available in table			

Showing 0 to 0 of 0 entries

Absconded patients with MISSING date of absconding(kindly populate Date observations done in the database)

Daily Reviews | Discharge & V.A | Follow up | Home Visits | Screening log | Missed visits | About Dashboard

Study Conclusion | Day 45 Follow Up | Day 90 Follow Up | Day 180 Follow Up

Day 180 follow up date must be before or same as study end

Show 10 entries Search:

record_id	site	Day 180 date	Study end date	comment
30001312	Blantyre	2019-05-17	2019-05-16	Study end happened before day 180
30001328	Blantyre	2019-06-19	2016-06-19	Study end happened before day 180

Showing 1 to 2 of 2 entries Previous 1 Next

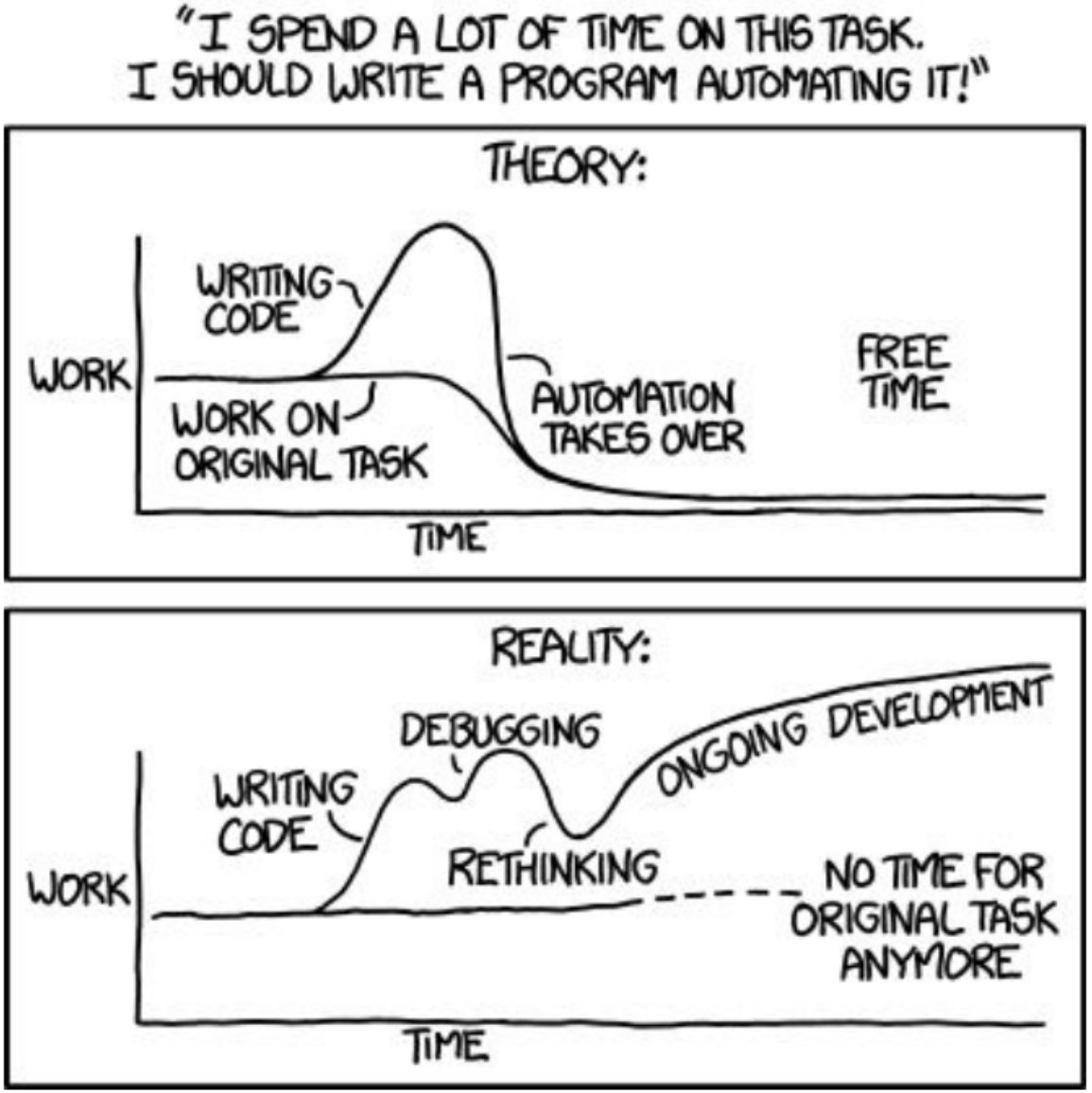
Missing day 180 anthropometry

Show 10 entries Search:

record_id	fu180_visdate	site	fu180_weight	fu180_height1	fu180_height2	fu180_muac1	fu180_muac2	fu
No data available in table								

Showing 0 to 0 of 0 entries Previous Next

- Seamless data extraction and processing with no human intervention.
- Run automated cleaning scripts and monitor data quality checks in real time.
- Up-to-minute access to standardized study reports by sites' PI and study staff.
- Real time study progress reports helped planning of recruitment and other activities.
- Improved productivity of the data team and reduced over-reliance on the CHAIN central team


"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"
THEORY:
WORK
WRITING CODE
WORK ON ORIGINAL TASK
AUTOMATION TAKES OVER
FREE TIME
TIME
REALITY:
WORK
WRITING CODE
DEBUGGING
RETHINKING
ONGOING DEVELOPMENT
NO TIME FOR ORIGINAL TASK ANYMORE
TIME

Everything never went smoothly as projected, we faced various challenges

- Dataset grew exceedingly large, slowing down the app. load time.
- Server downtime, making the app. inaccessible sometimes.
- Processing data from different sources concurrently for reporting purposes within the app.

The framework presented here can easily be extended to produce robust and scalable Shiny applications:-

- Spatial data analysis and interactive visualization.
- Data products for statistical analysis and modelling.
- Highly engineered data cleaning and anomaly detection tools

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