Enhanced Data Querying For Neuroinformatics Databases

D. MacFarlane[1], S. Das[1], P. Kostopoulos[1], J. Kat[2], C. Rogers[1], C. Makowski[1], A.C. Evans[1]
1. Montreal Neurological Institute / McGill University, Montreal, QC, Canada
2. Douglas Institute Research Centre / McGill University, Montreal, QC, Canada

Introduction

One of the analysis challenges in large scale multimodal studies is for non-technical users to easily extract meaningful data, without requiring database expertise. Research databases have traditionally relied on a relational database model, which requires at least basic knowledge of SQL or other programming languages for data querying and extraction.

The Data Querying Tool (DQT) developed for the LORIS web-based neuroinformatics database (S. Das et al, 2012)[1] allows the combination of clinical, psychological, neuroimaging and biomarker data into a tabular display which can be organized either cross-sectionally or longitudinally and viewed or downloaded through a user-friendly web interface. Basic analysis of the data is also possible through this tool.

Although the tool was developed for LORIS, the data fed to the querying tool can come from any data source with the creation of an import script. Thus, this querying tool allows the combination of data from multiple sources into a format which is consistent, and more efficient than traditional SQL-based solutions.

Methods

CouchDB was selected as a backend for our new DQT. CouchDB's JSON-based "schema-less" design offers advantages in terms of incorporating data from diverse data sources compared with SQL's strict schema requirements.

The DQT was implemented as a "CouchApp" using only HTML5 and Javascript on the client side for loading and processing of data. This allowed us to implement basic data summarizing features, such as scatterplots, histograms, and column statistics without placing any additional burden on the server.

Results

A proof of concept using CouchDB instead of MySQL for the data querying tool has been implemented for IBIS (Wolff et al, 2012)[2], a multi-center longitudinal neurodevelopmental project consisting of approximately 60,000 tests administered, and PreventAD[3], a smaller study of approximately 7500 tests. Performance is compared below using PreventAD.

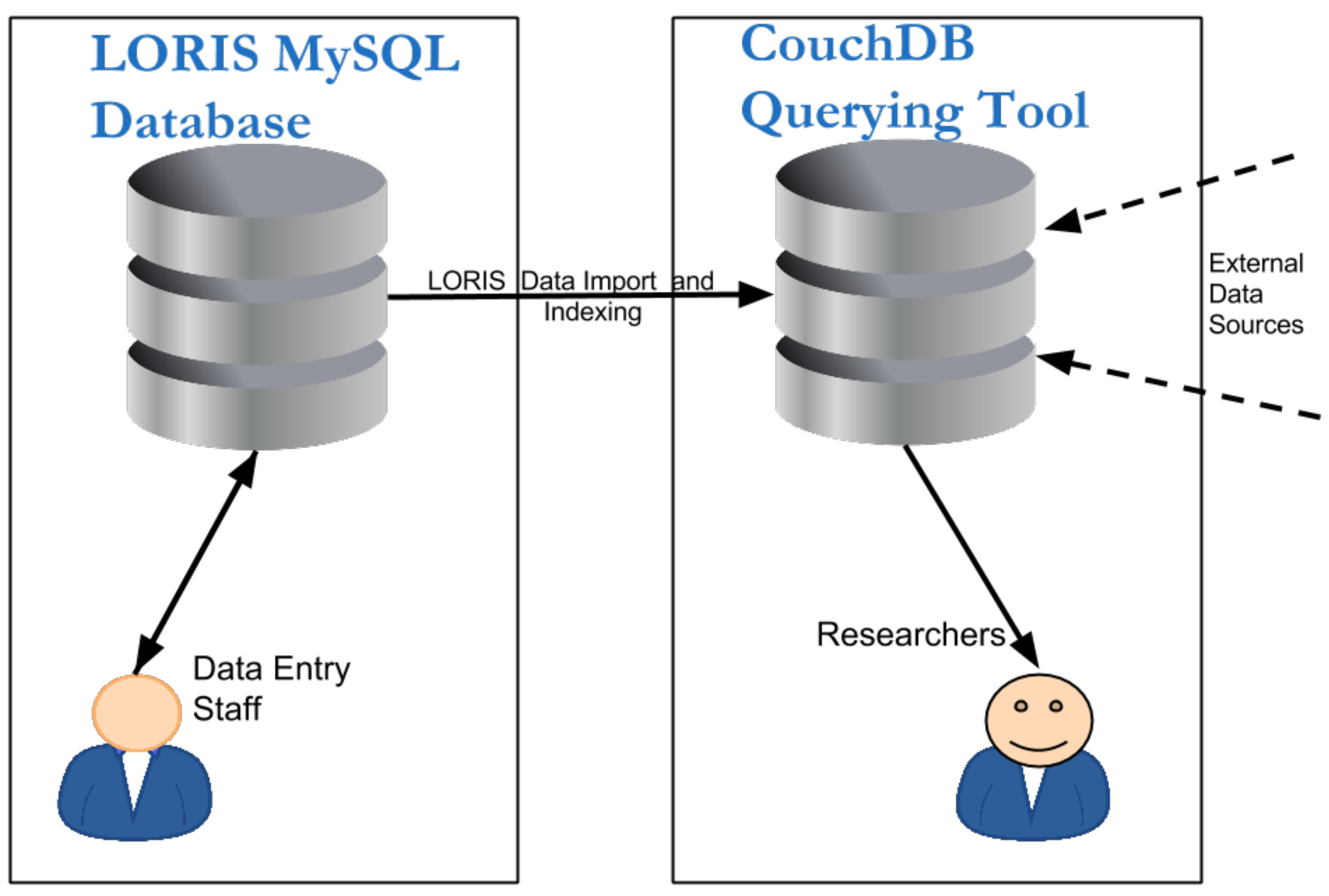
Action	MySQL	CouchDB
Initial Data Import	362 seconds	45 seconds
Query Time	115 seconds	31 seconds

Table 1: DQT Performance Comparison

In our experiments, building and indexing the DQT using CouchDB rather than a MySQL database resulted in ten-fold efficiency gains (Table 1), a benchmark that only increased with larger datasets. This added the critical benefit of allowing us to populate the DQT with greater frequency and less downtime. Data is indexed incrementally rather than completely redone on every import using CouchDB. In addition, during previous MySQL data imports, tables were locked and therefore inaccessible for data entry which is no longer the case with CouchDB.

Querying time also benefitted from the CouchDB platform, where querying for one clinical measure (including demographic data) was notably faster (Table 1).

Data Flow


LORIS MySQL Database

CouchDB Querying Tool

LORIS Data Import and Indexing

External Data Sources

Data Entry Staff

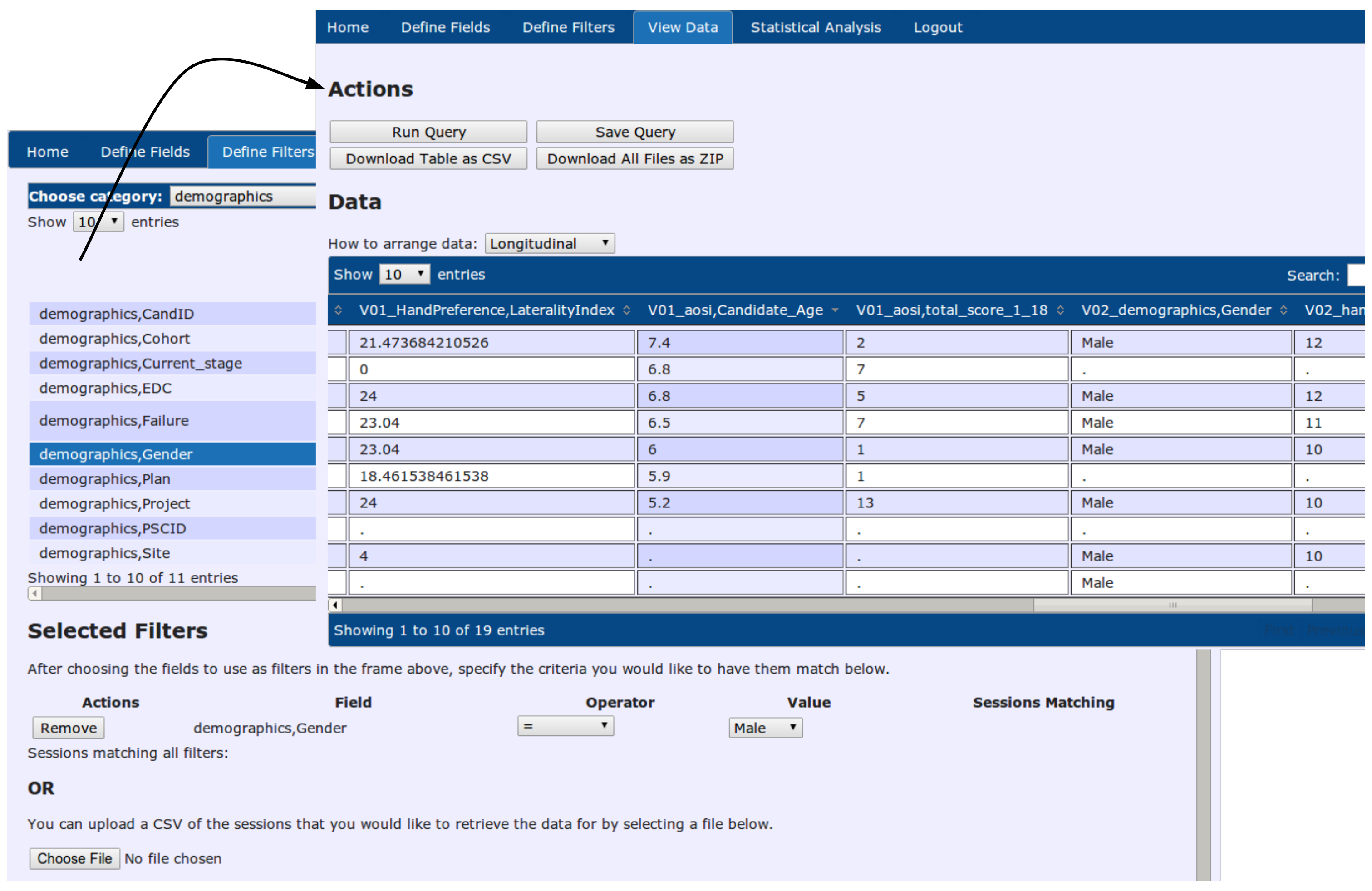
Researchers

Conclusion

We were able to create a faster, more efficient graphical querying tool for our data by switching from a MySQL backend to a CouchDB backend. By moving the processing of data to the web browser we were able to add more features with no significant impact on our server resources.

Data Extraction

The DQT can be used for combining and extracting various types of data related to the study. The following screenshots show an example of a query in a longitudinal study arranged as one row per candidate from our demo database.



Actions

Data

How to arrange data: Longitudinal

Show 10 entries Search:

	V01_HandPreference,LateralityIndex	V01_aosi,Candidate_Age	V01_aosi,total_score_1_18	V02_demographics,Gender	V02_han
	21.473684210526	7.4	2	Male	12
	0	6.8	7	.	.
	24	6.8	5	Male	12
	23.04	6.5	7	Male	11
	23.04	6	1	Male	10
	18.461538461538	5.9	1	.	.
	24	5.2	13	Male	10
	.	.	.	.	.
	4	.	.	Male	10
	.	.	.	Male	.

Showing 1 to 10 of 19 entries

Choose category: demographics

Show 10 entries

demographics,CandID
demographics,Cohort
demographics,Current_stage
demographics,EDC
demographics,Failure
demographics,Gender
demographics,Plan
demographics,Project
demographics,PSCID
demographics,Site

Showing 1 to 10 of 11 entries

Selected Filters

After choosing the fields to use as filters in the frame above, specify the criteria you would like to have them match below.

Actions	Field	Operator	Value	Sessions Matching
Remove	demographics,Gender	=	Male	

Sessions matching all filters:

OR

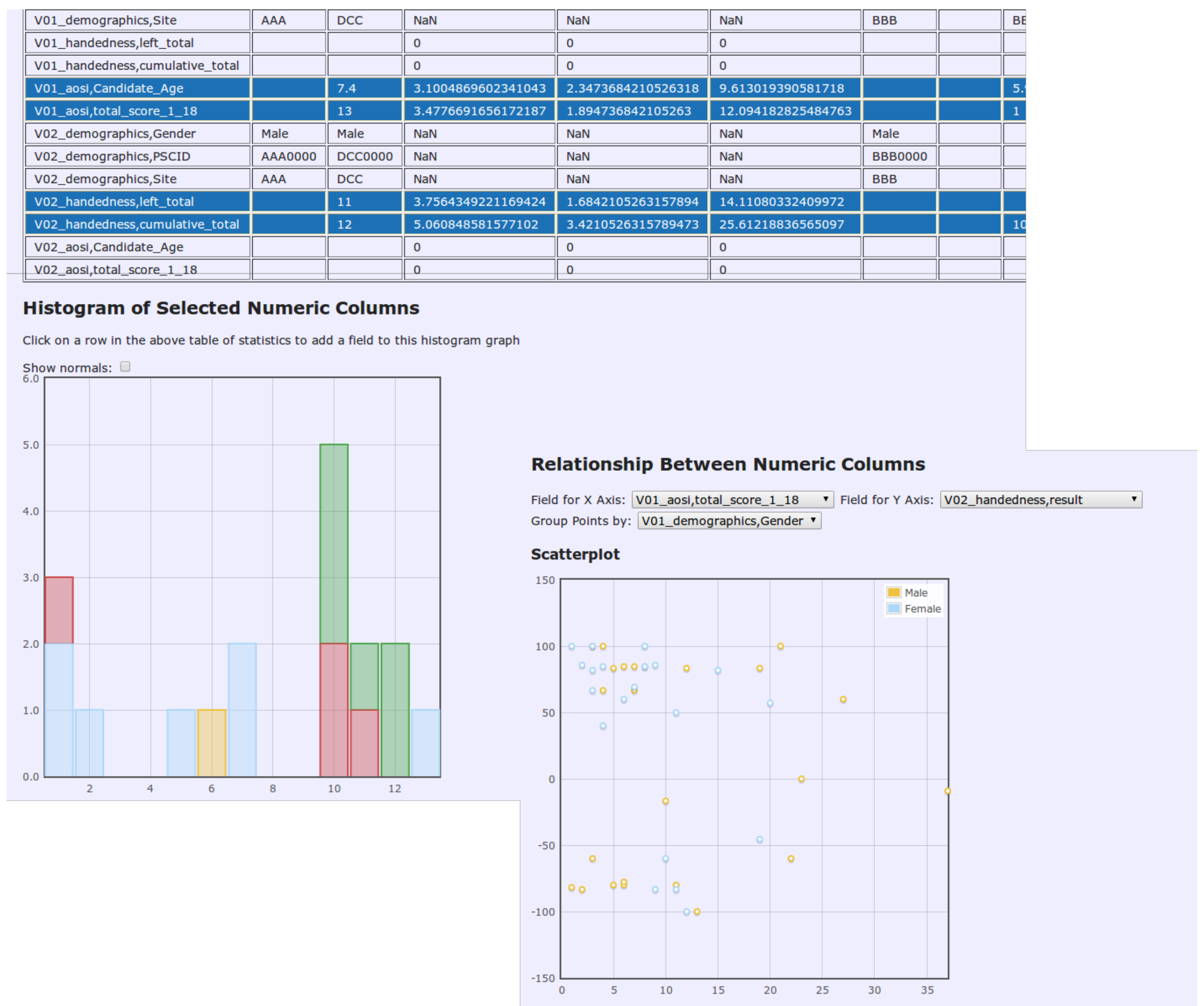
You can upload a CSV of the sessions that you would like to retrieve the data for by selecting a file below.

Choose File No file chosen

The table can also be arranged such that each row represents a single visit, with a candidate who has multiple timepoints containing multiple rows.

Basic Analysis

Basic statistical features are also incorporated directly into the DQT which allow you to quickly see summary statistics of your table columns at a glance. The following screenshots show a sample of the features available.



V01_demographics,Site	AAA	DCC	NaN	NaN	NaN	BBB		BE
V01_handedness,left_total			0	0	0			
V01_handedness,cumulative_total			0	0	0			
V01_aosi,Candidate_Age		7.4	3.1004869602341043	2.3473684210526318	9.613019390581718			5.
V01_aosi,total_score_1_18		13	3.477691656172187	1.894736842105263	12.094182825484763			1
V02_demographics,Gender	Male	Male	NaN	NaN	NaN	Male		
V02_demographics,PSCID	AAA0000	DCC0000	NaN	NaN	NaN	BBB0000		
V02_demographics,Site	AAA	DCC	NaN	NaN	NaN	BBB		
V02_handedness,left_total		11	3.7564349221169424	1.6842105263157894	14.11080332409972			
V02_handedness,cumulative_total		12	5.060848581577102	3.4210526315789473	25.61218836565097			10
V02_aosi,Candidate_Age			0	0	0			
V02_aosi,total_score_1_18			0	0	0			

Histogram of Selected Numeric Columns

Click on a row in the above table of statistics to add a field to this histogram graph

Show normals: □

Relationship Between Numeric Columns

Field for X Axis: V01_aosi,total_score_1_18 Field for Y Axis: V02_handedness,result

Group Points by: V01_demographics,Gender

Scatterplot

The first table shows basic statistical outputs such as minimum, maximum, standard deviation, and quartile values for each column queried. Visual tools such as histograms and scatterplots are also included and allow for additional benefits, such as quick data visualization of any two selected columns, either cross-sectionally or longitudinally organized. Scatterplots can optionally be segmented based on a third column and you can show the least square fit for each group.

Try it

You can try our DQT using a modern web browser at our demo database at https://demo.loris.ca/dqt/ using the username/password "demo"/"demo".

References

1. Das S, Zijdenbos AP, Harlap J, Vins D & Evans AC (2012) LORIS: a web-based data management system for multi-center studies. Front. Neuroinform. 5:37

2. Wolff et al (2012) Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. Am J Psychiatry

3. PreventAD: http://www.douglas.qc.ca/page/prevent-alzheimer-the-centre