Automated Processing of Phenotypic Data Submissions

R. Mayani[1], K. Vahi[1], JL. Ambite[1], S. Sharma[1], M. Azaro[2], S. Wilson[2], B. Ruocco[2], G. Davis[2], M. Romanella[2], L. Brzustowicz[2], E. Deelman[1], Y. Arens[1]

1 = University of Southern California, Information Sciences Institute, 2 = Rutgers University, Genetics Department

Motivation

- The NIMH Repository & Genomics Resource (NRGR) is a sharing repository that maintains biomaterials, genetic data, and clinical data of individuals with a range of psychiatric illnesses, their family members, and unaffected controls.
  - Goal is to provide data & samples to researchers to accelerate psychiatric genetics research.
- NRGR receives demographic data, summary and detailed clinical data from NIMH funded studies and makes this data accessible in a secure, access-controlled fashion.
  - NRGR hosts data from **226** studies, and **350+** requests to access these datasets have been approved.
- Different NIMH funded studies collect data in ad-hoc formats, so same information could be collected with varying representations, For example, studies could collect Gender information as 0 or 1, M or F, Male or Female.
  - It is difficult for researchers to consume data from multiple studies with differing formats.
- Data can be even more valuable, if standardized.
  - Previously, the curation efforts were largely manual with ad-hoc standardization procedures in place that led to inconsistencies and variance across studies and disorders.
  - Automated quality control system (AutoQC) streamlines and formalizes phenotypic data submissions ensuring stringent data quality requirements.

Design Goals

- **Web Based** –
  - No need to install any complex software.
  - Reduces learning curve.
- **Easy & Flexible Data Dictionaries** –
  - Support **system-defined** requirements for all studies. For example, Gender specified as **M** or **F**.
  - Allow **study-defined** data dictionaries for data unique to the study.
- **Fire and Forget** –
  - Users can submit their data and monitor the progress in real time.
  - Users can submit their data to the system, and log out without interrupting the curation processes.
- **Scalable and Responsive** –
  - System to be responsive and scalable, so multiple simultaneous submissions do not overwhelm the system and impact response times.

Basic Checks using Data Dictionary

- **Descriptive Fields** –
  - *name*, *unit*, *description* fields describe data being collected in the clinical instrument.
- **Type Checks** –
  - *type* field defines the data type of the value expected in submission, i.e. integer, string, etc.
- **Range Checks** –
  - *min* and *max* fields define the range of valid values.
- **Length Checks** –
  - *min_length* and max_length fields define length restrictions for textual values.
- **Relational Database styled Checks –**
  - *primary_key*, *unique*, and *mandatory* fields are used to represent database styled primary key, unique key, and NULL constraints respectively.
- **Conditionally Required Fields –**
  - *mandatory* field also allows for an expression, i.e. value is required if the expression evaluates to true.
- **Both System defined and Study defined requirements are expressed using this data dictionary. Examples below.**
  - **Standard Demographic Data**
  - **Race Ethnicity Data**
  - **Detailed Diagnosis Data**

name	unit	type	min	max	mandatory	values
age	years	integer	0.0	120.0	y	
twins		string,fixed_set			c[“subject_type”] != “DUMMY”	Monozygotic |Dizygotic

Ancestry Checks on Family Data

- **Ancestry information is important factor in psychiatric genetic research.**
  - Ensure individuals identified as fathers are also Males.
  - Ensure individuals identified as mothers are also Females.
  - Ensures individuals identified as parents are older than their children.
  - Ensure every family has at least one individual identified as a proband.

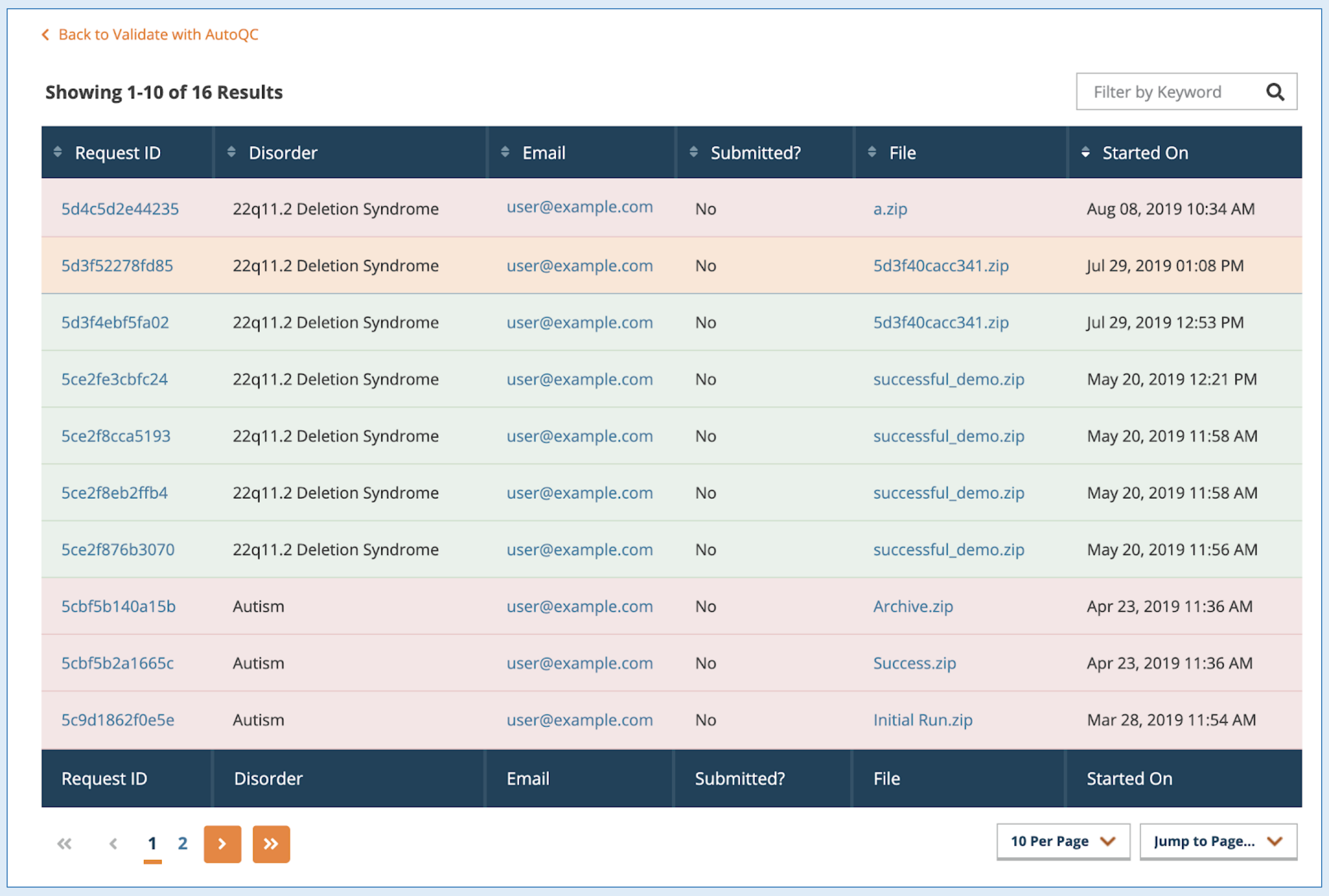
Advanced Checks

- **Semantic Checks** –
  - Ensure there are no mismatches: **age >= age of onset**, **current_year - year of birth >= age**.
  - Ensure year of death is specified if an individual is marked as deceased.
  - Ensure consent value is specified for non-dummy individuals.
- **Correlate data with NRGR records** –
  - NRGR collects both, electronic clinical data and physical bio-samples (blood, saliva, etc.) from individuals.
  - System ensures clinical data and bio-samples can be correlated with each other.
  - Ensures study and site id specified in submission are valid and registered with NRGR.
- **Other Checks** –
  - Ensures diagnosis codes specified are valid within the specified diagnosis systems (DSM, etc.).
  - Ensures all subjects with clinical interview data also have a record in standard demographic and diagnostic file.
- **Suggested Corrections** –
  - AutoQC suggests valid values for common issues, for example
    - Change diagnosis code to **238.00** from **238.0**.

Scalability with Pegasus WMS

- AutoQC runs submissions using Pegasus WMS managed scientific workflows.
- Pegasus WMS allows user submissions to be queued and run on a cluster in a distributed fashion using HTCondor.
- Pegasus WMS ensures the system doesn’t get overwhelmed by too many submissions running simultaneously, and impacting system response times.
- Pegasus WMS allows adding or removing more computing nodes to the cluster without impacting running curation processes.

Screenshots

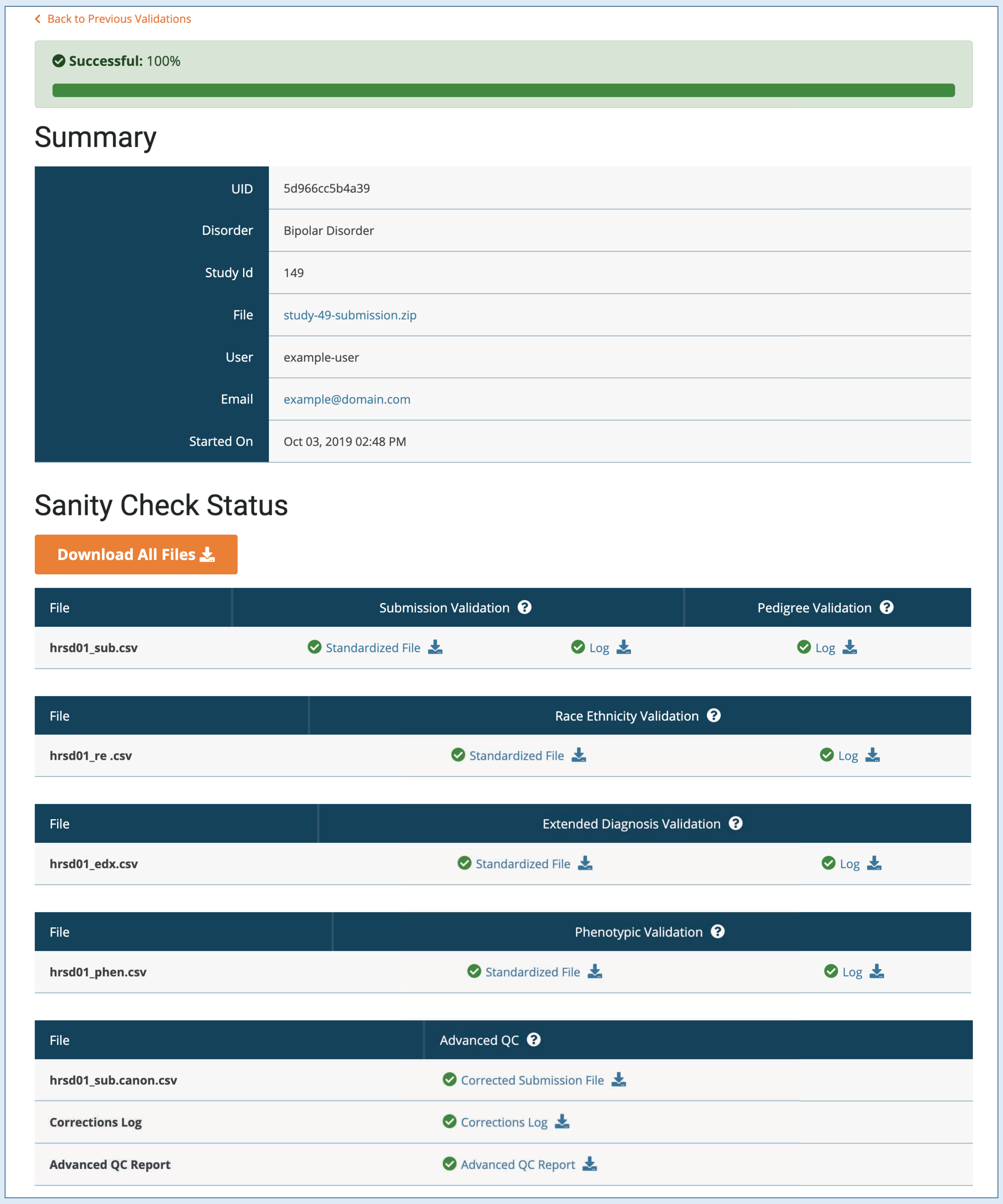


‹ Back to Validate with AutoQC

Showing 1-10 of 16 Results

Request ID	Disorder	Email	Submitted?	File	Started On
5d4c5d2e44235	22q11.2 Deletion Syndrome	user@example.com	No	a.zip	Aug 08, 2019 10:34 AM
5d3f52278fd85	22q11.2 Deletion Syndrome	user@example.com	No	5d3f40cacc341.zip	Jul 29, 2019 01:08 PM
5d3f4ebf5fa02	22q11.2 Deletion Syndrome	user@example.com	No	5d3f40cacc341.zip	Jul 29, 2019 12:53 PM
5ce2fe3cbfc24	22q11.2 Deletion Syndrome	user@example.com	No	successful_demo.zip	May 20, 2019 12:21 PM
5ce2f8cca5193	22q11.2 Deletion Syndrome	user@example.com	No	successful_demo.zip	May 20, 2019 11:58 AM
5ce2f8eb2ffb4	22q11.2 Deletion Syndrome	user@example.com	No	successful_demo.zip	May 20, 2019 11:58 AM
5ce2f876b3070	22q11.2 Deletion Syndrome	user@example.com	No	successful_demo.zip	May 20, 2019 11:56 AM
5cbf5b140a15b	Autism	user@example.com	No	Archive.zip	Apr 23, 2019 11:36 AM
5cbf5b2a1665c	Autism	user@example.com	No	Success.zip	Apr 23, 2019 11:36 AM
5c9d1862f0e5e	Autism	user@example.com	No	Initial Run.zip	Mar 28, 2019 11:54 AM
Request ID	Disorder	Email	Submitted?	File	Started On

Figure 1: Dashboard listing users’ past curation runs.



‹ Back to Previous Validations

Successful: 100%

Summary

UID	5d966cc5b4a39
Disorder	Bipolar Disorder
Study Id	149
File	study-49-submission.zip
User	example-user
Email	example@domain.com
Started On	Oct 03, 2019 02:48 PM

Sanity Check Status

File	Submission Validation		Pedigree Validation
hrsd01_sub.csv	✓ Standardized File	✓ Log	✓ Log

File	Race Ethnicity Validation	
hrsd01_re .csv	✓ Standardized File	✓ Log

File	Extended Diagnosis Validation	
hrsd01_edx.csv	✓ Standardized File	✓ Log

File	Phenotypic Validation	
hrsd01_phen.csv	✓ Standardized File	✓ Log

File	Advanced QC
hrsd01_sub.canon.csv	✓ Corrected Submission File
Corrections Log	✓ Corrections Log
Advanced QC Report	✓ Advanced QC Report

Figure 2: Monitoring submission progress.
Phenotypic Validation is done using Study defined data-dictionaries.

Important Links –

AutoQC – www.nimhgenetics.org/submit-your-data/overview
Pegasus WMS – pegasus.isi.edu

www.nimhgenetics.org