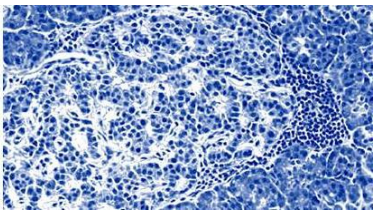
nPOD
Network for Pancreatic Organ Donors with Diabetes



DataView: A Data Science Platform Integrating 10+ years of nPOD enabled T1D research

AUTHORS

Anh Nguyet Vu,[1,2] Dave Ko,[1,2] John S. Kaddis[1,2]

Department of Diabetes Immunology
Department of Diabetes and Cancer Discovery Science
Diabetes and Metabolism Research Institute
City of Hope, Duarte, CA

PURPOSE

Since 2008, investigators in nPOD have utilized the biobank to generate and publish a diverse amount of data. Toolkits are needed to connect and incorporate this information by donor across multiple experiments, laboratories, and assay types. However, there has been no comprehensive framework to aggregate, curate, and integrate these results in ways that may be used to drive new hypotheses, research questions, and collaborations. The DataView platform demonstrates an approach to how donor-connected heterogeneous data can be integrated and made available for exploration and new applications.

METHODS

Relevant data sources for curation were defined as manuscripts and datasets (which may not have an associated paper) published between 2008 and June of 2019 and include nPOD donor-derived primary data. Data was obtained through manual or computational extraction from the publication/public repository or directly requested from the authors. The information was then cleaned, record-linked with nPOD CaseIDs, transformed, and augmented with ontology and other metadata. Additional donor data was obtained from nPOD DataShare. Interactive modules for viewing, exploring, and mining data were built with R and the Shiny web application framework.

SUMMARY OF RESULTS

A total of 78 data sources were curated. Data integration was possible for only 50% of the sources (n=39) and yielded ~400 small-throughput measurement features. A total of 4 data sources included medium/high-throughput and complex-type datasets. Analysis can be performed across assay/sources and donor/cohorts. Data can be discovered through a friendly user interface or through SPARQL endpoints as well as downloaded for use in other data science applications.

CONCLUSIONS

DataView is available as a web-based application, Docker image, and R package. Three major challenges remain: availability and access to raw or processed data, accommodation of complex data types, and implementation of mining tools for heterogeneous data.