DATA SOURCE DESCRIPTION

Nashville
Biosciences®

Leveraging Vanderbilt University Innovation™

DATA SOURCE

Nashville Biosciences

Nashville Biosciences (NashBio), a wholly owned subsidiary of Vanderbilt University Medical Center (VUMC, Tennessee, US), is a data and analytics provider to clients within the healthcare and biotechnology industries. NashBio harnesses VUMC's extensive genomic and bioinformatics resources, including, but not limited to BioVU®. Created at VUMC and Vanderbilt University, BioVU® is one of the world's most comprehensive biorepositories and genetic databases linked to de-identified medical records with years of longitudinal clinical data, established in order to propagate drug, diagnostics and other therapeutic discovery and development.

Synthetic Derivative (SD)

NashBio constructs patient cohorts and extracts clinical phenotype data from VUMC's Synthetic Derivative (SD) database[1,2], a de-identified version of VUMC's electronic medical record (EMR). The SD contains de-identified clinical data covering more than 3.6 million patients.

VUMC's SD database is derived from the front line EMR system used at VUMC hospitals and a network of owned primary-care clinics using the same platform. VUMC operates these healthcare facilities in Tennessee, as well as a small number of clinics in Kentucky. VUMC first began using an EMR in 2001, with the implementation of StarPanel, a custom, longitudinal EMR system developed in-house by Vanderbilt engineers. De-identification of the EMR to generate the SD began in 2007. VUMC transitioned to an Epic Systems platform effective November 1, 2017, with no impact on the continuity of clinical data in the SD. VUMC handles the transformation of these clinical data to a common data model (Observational Health Data Sciences & Informatics' OMOP standard, version 5.3.1) and subsequent de-identification through an extensive and automated pipeline for ultimate storage in the SD, as further described below. Patients in the SD have an average of 10 years of EMR data available, and data in the SD are refreshed on a bi-monthly basis.

BioVU®

BioVU® is a collection, or biobank, of de-identified DNA samples linked to the SD database[1,2]. During routine patient care, VUMC clinicians may order blood tests to monitor or help treat a patient. Leftover portions of any sample collected for routine testing are typically discarded. However, for patients that have consented to participate, (https://victr.vumc.org/biovu-consent/)[3] BioVU® retains these residual samples, and extracts and banks the patient's germline DNA from them. DNA samples are de-identified and only linked to the SD through a randomly assigned unique identifier, not back to the patient or their underlying medical record. Germline DNA samples from more than 307,000 unique VUMC patients are currently available to Vanderbilt academic researchers and to commercial researchers through NashBio to study the genome's role in human health and disease. SNP genotyping and next generation sequencing (whole exome and whole genome) have been completed on various sub-cohorts of BioVU®, and are also available for study. Blood plasma may be derived from the same residual blood samples, and may be collected on demand in a phenotype-specific manner for additional 'omic studies.

Data Integrity & Privacy

NashBio is not required to seek study-specific patient authorization for the use of the SD or BioVU® since all samples and patient records are de-identified under the safe harbor provisions of the Health Insurance Portability and Accountability Act of 1996, as amended (HIPAA). The 18 different fields considered identifying under HIPAA are removed from the structured part of the clinical data, and unstructured clinical notes are de-identified using natural language processing tools to remove possible identifiers while preserving linguistic context. These notes are not included in the structured clinical data made available to NashBio clients, though their review and analysis is readily available through established processes with NashBio's clinical or data teams.

Once the 18 identifiers are removed, patients are subsequently assigned a unique identifier via a secure, one-way hash algorithm (SHA-512) developed by the National Security Agency. The unique identifier is not tied to any of the stripped identifiers and all analysis and updates to the SD thereafter utilize this unique identifier as the patient index. Additionally, each patient's record is time-shifted a random number of days backwards (up to 364 days) from their actual time index that is consistent within each record but differs across patients. The relative temporal structure of each record is maintained, but the random shifting further obscures patient identity. In keeping with strict adherence to HIPAA safe harbor regulations, VUMC removes access to clinical data for patients from the date of their 90th birthday onward in the SD. For example, if a person is in the VUMC system from age 65-96 years, clinical data will be accessible for years 65-89 only. Subjects with a current age > 89 years can be included in a cohort, but clinical data for these subjects will not reflect any diagnoses, medications, or procedures in their 90th year or beyond. In addition to these steps taken for safe harbor compliance, the structured clinical data has been confirmed as de-identified under the HIPAA expert determination method.

Similarly, NashBio is not required to seek study-specific consent for use of these datasets because the use of BioVU® is classified as non-human subject research by VUMC's Institutional Review Board (IRB). As noted, all patients consent to their samples and data being contributed to BioVU® and all samples and data are de-identified. A small fraction of DNA samples collected pursuant to consent are randomly excluded from inclusion in BioVU® as an additional privacy protection step. As part of the program design, participants cannot be re-contacted. Due in part to these measures, the Vanderbilt IRB has determined that creation and management of the biobank repository in this manner is not human subjects research, and that the use of residual de-identified samples and de-identified data from the repository resource does not constitute human subject research.

The overall biobanking program is reviewed annually by the IRB to maintain this determination and make decisions about patient protections, privacy and ethical issues. Each individual study seeking to use the SD database and BioVU® biobank is filed with VUMC's IRB to validate its non-human subject classification and appropriate use of the data. NashBio takes steps to protect client confidentiality during this submission (e.g., use of a client-specific coded identifier).

REFERENCES

1. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharm Ther. 2008;84:362-9.

2. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt out, de-identified biobank. Clin Transl Sci. 2010;3:42-8.

3. During the timeframe from 2007-2015, VUMC's standard consent for routine treatment process described the use of leftover blood for research to patients. During patient registration, patients were also presented with information about the biobank and given the option to decline participation through an opt-out checkbox. Starting in 2015, participants were consented under an affirmative biobank consent form, the present version of which is provided at the link above.