What Can We Learn from Vanderbilt's Synthetic Derivative?  
A Rich Clinical Data Resource Stripped of Personal Identifiers

Ian Brooks, PhD
Director, Office of Biomedical Informatics
The Vanderbilt Synthetic Derivative

- The Synthetic Derivative (SD) is a database containing clinical information derived from Vanderbilt’s electronic medical record.

- The SD is a set of records that is no longer linked to the identified medical record from which it is derived and has been altered to the point it no longer closely resembles the original record.

- The SD can be used as a stand-alone research resource, or can be used in conjunction with BioVU to identify record sets for genome-phenome analysis.

- The SD interface allows the user to search data extracted from most of the major health information databases at Vanderbilt and contains records for over 2.2 million unique individuals.

- User can search records via basic clinical and demographic information, (ICD 9 codes, CPT codes, meds, lab values, age and gender), returns de-identified data to the user for review and selection.
Information collected during clinical care

- Star Server
- EDW
- HEO

DE-IDENTIFICATION

Data Parsing

Restructuring for research

One way hash

Data export

SD Database

Access through secured online application
The Vanderbilt Synthetic Derivative

- What is it used for?

- Retrospective chart reviews
- Rapid preliminary data for grant submissions
- Hypothesis generation
- Feasibility assessment
The Vanderbilt Synthetic Derivative

- Opt-out consent system
- De-identified extracts from EMR & Data warehouse
- 128 RUI 1-way hash
  - Dates shifted within a record by a time period that is consistent within each record, but differs across records up to 364 days backwards
- Linked to blood in BioVU for Genome-Phenome studies
Technology + policy

- **De-identification**
  - Derivation of 128-character identifier (RUI) from the MRN generated by Secure Hash Algorithm (SHA-512)
    - RUI is unique to input, cannot be used to regenerate MRN
    - RUI links data through time and across data sources
  - HIPAA identifiers removed using combination of custom techniques and established de-identification software

- **Restricted access & continuous oversight**
  - Access restricted to VU; not a public resource
  - IRB approval for study (non-human)
  - Data Use Agreement
  - Audit logs of all searches and data exports
Figure 2 A descriptive example of a record in the synthetic derivative (SD) described in the text. The arrows indicate examples of scrubbing: the medical record number has been removed (black), the social security and phone numbers have been masked (blue), names have been changed (purple), and dates have been shifted (red) as described in Methods.
What the SD can’t do...

- Outbreaks and other date-specific studies (catastrophes, etc.)
- Find a specific patient (e.g. to contact)
- Replace large scale epidemiology research (e.g. TennCare database)
- Temporal search capabilities limited
  - “First this, then that” study designs require significant manual effort
The Vanderbilt Synthetic Derivative

- Benefits of a de-identified data registry
  - Generate large cohorts of controls
  - Rapid & inexpensive hypothesis testing
  - Clinical trial subject availability
  - Linked to GWAS/PheWAS – Biobank
    - High impact demonstration & validation papers being published
    - Rapidly generate preliminary data for funding applications
The Vanderbilt Synthetic Derivative

- **Technical considerations**
  - Computationally demanding
  - Requires broad & dedicated input from stakeholders
  - Requires extensive experience in NLP

- **Ethical considerations**
  - Requires extensive community buy-in
  - Requires extensive ethical oversight
  - Opt-In Consent is a significant burden on staff
The UTHSC Synthetic Derivative

Could we?...Should we?...Can we?...