Overview of the Synthetic Derivative

April 22, 2009

Melissa Basford, MBA

Program Manager, BioVU and Synthetic Derivative

What is BioVU?

Vanderbilt BioVU

- A biobank intended to support a broad view of biology
- Currently contains de-identified DNA extracted from leftover blood after clinically-indicated testing of Vanderbilt patients who have not opted out
  - Future expectation of other tissue types: serum proteomics, possibly surgical tissues
- Link to Synthetic Derivative

The Synthetic Derivative

- A **Derivative** of the EMR - information content reduced by ‘scrubbing’ identifiers
- With a **Synthetic** ‘disinformation’ component – inserted systematically shifted event dates
- Can be searched restricting to records for which DNA is available
- A research resource in its own right, without biobank
  - Systematic query of clinical events
  - Detection of previously unsuspected comorbidities and causal relationships (e.g., automated Vioxx-like event detection)

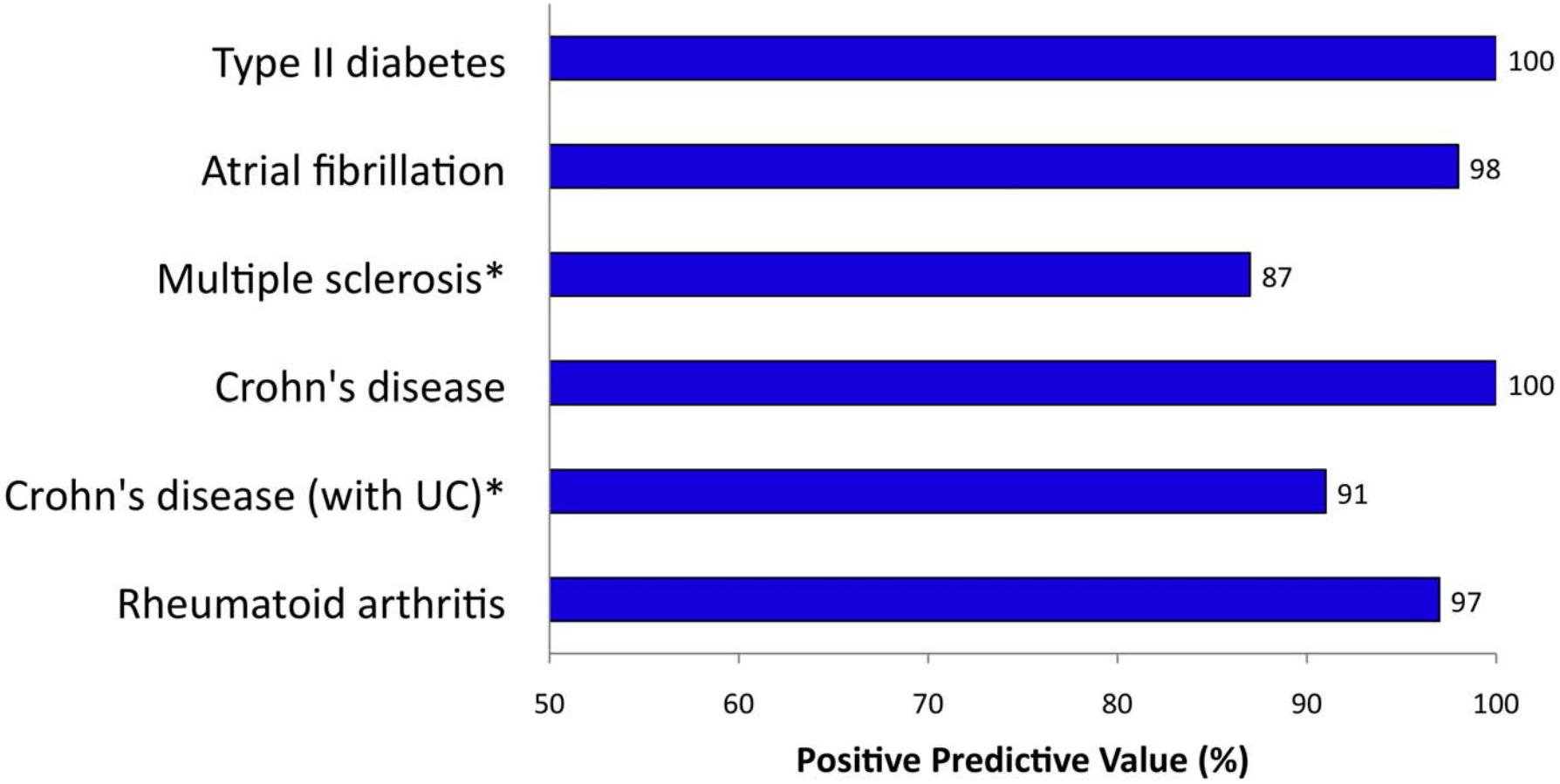
The “demonstration project”

- Assess previously reported genotype-phenotype association for SNPs associated with disease
- Genotype SNPs in the first 10,000 samples accrued
  - 38 loci in common phenotypes

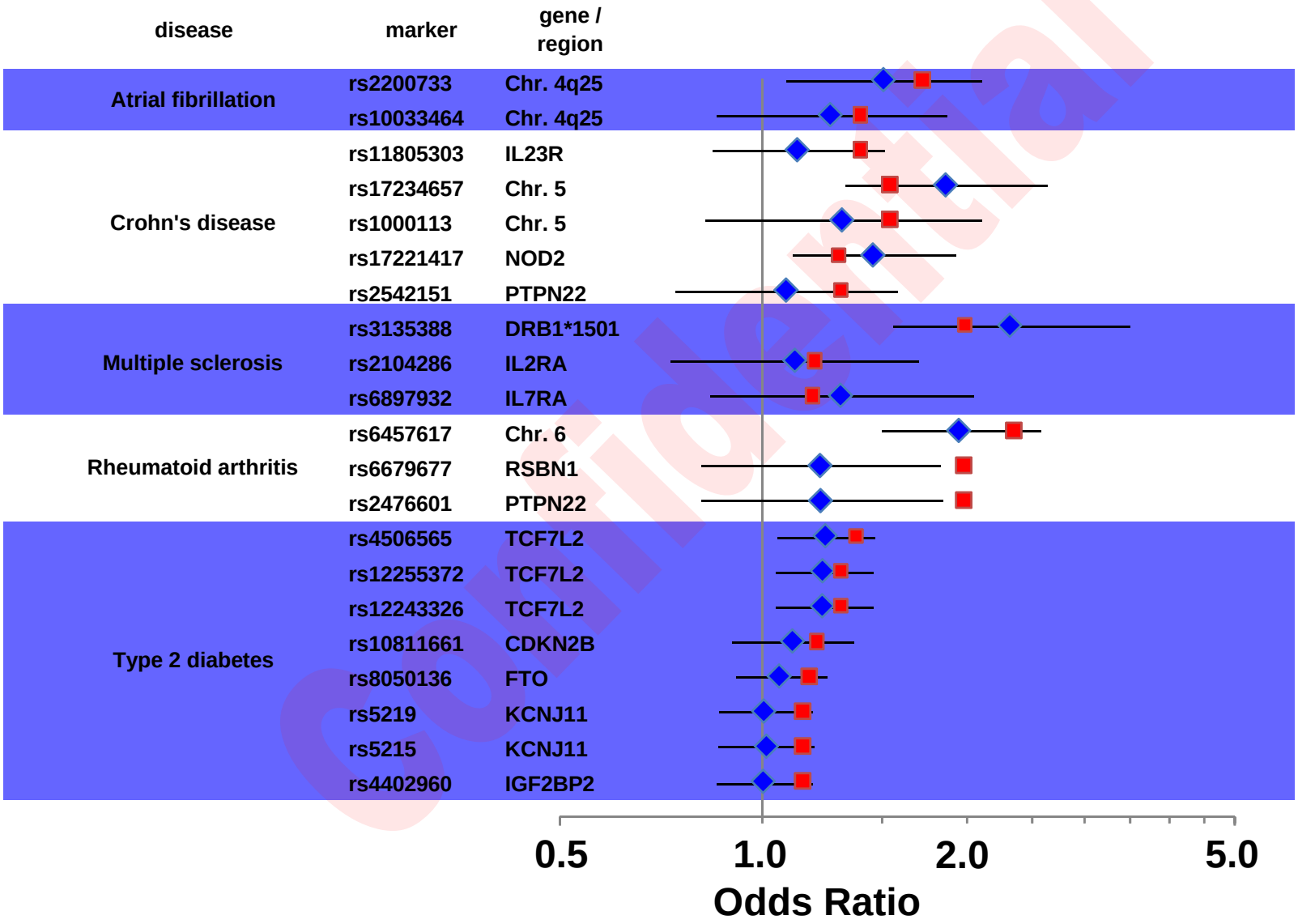
Atrial fibrillation	Alzheimer’s Disease
Bipolar disorder	Breast cancer
Crohn’s disease	**Multiple Sclerosis**
MI at age <50	Prostate cancer
Rheumatoid arthritis	Type I Diabetes
Type II Diabetes	QT interval

- Develop automated phenotype identification methods for both cases and controls

Algorithm accuracy



Demonstration project results



Conclusions

- In each disease studied, at least one previously associated SNP was replicated
  - SNPs that did not replicate were generally statistically underpowered given the small sample sizes
- In some diseases, phenotypes extracted from the EMR showed stronger effects
- **Shows that EMR derived phenotypes can be used for genetic association studies**

Synthetic Derivative resource overview

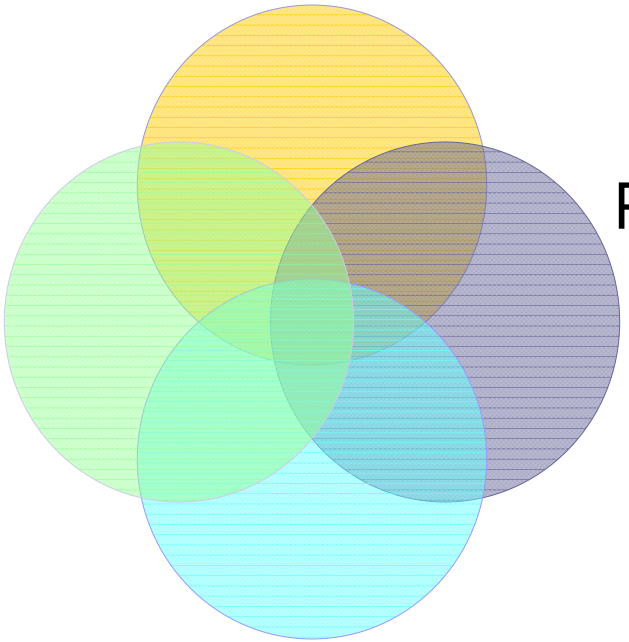
- Rich, multi-source database of de-identified clinical and demographic data
- Contains ~1.7 million records
  - ~1 million with detailed longitudinal data
  - averaging 100k bytes in size
  - an average of 27 codes per record
- Records updated over time and are current through 12/31/08

Research use cases assumed in resource development

(either alone, or with DNA samples)

Retrospective chart reviews

Hypothesis generation



Rapid preliminary data for grant submissions

Feasibility assessment

Data Types (so far)

- Narratives, such as:
  - Clinical Notes
  - Discharge Summaries
  - History & Physicals
  - Problem Lists
  - Surgical Reports
  - Progress Notes
  - Letters
- Diagnostic codes, procedural codes
- Forms (intake, assessment)
- Reports (pathology, ECGs, echocardiograms)
- Clinical Communications
- Lab values and vital signs
- Medication orders
- TraceMaster (ECGs)

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

Date shift feature

- Our algorithm shifts the dates within a record by a time period that is consistent within each record, but differs *across* records
  - up to 364 days backwards
  - e.g. if the date in a particular record is April 1, 2005 and the randomly generated shift is 45 days in the past, then the date in the SD is February 15, 2005)

What the SD can’t do

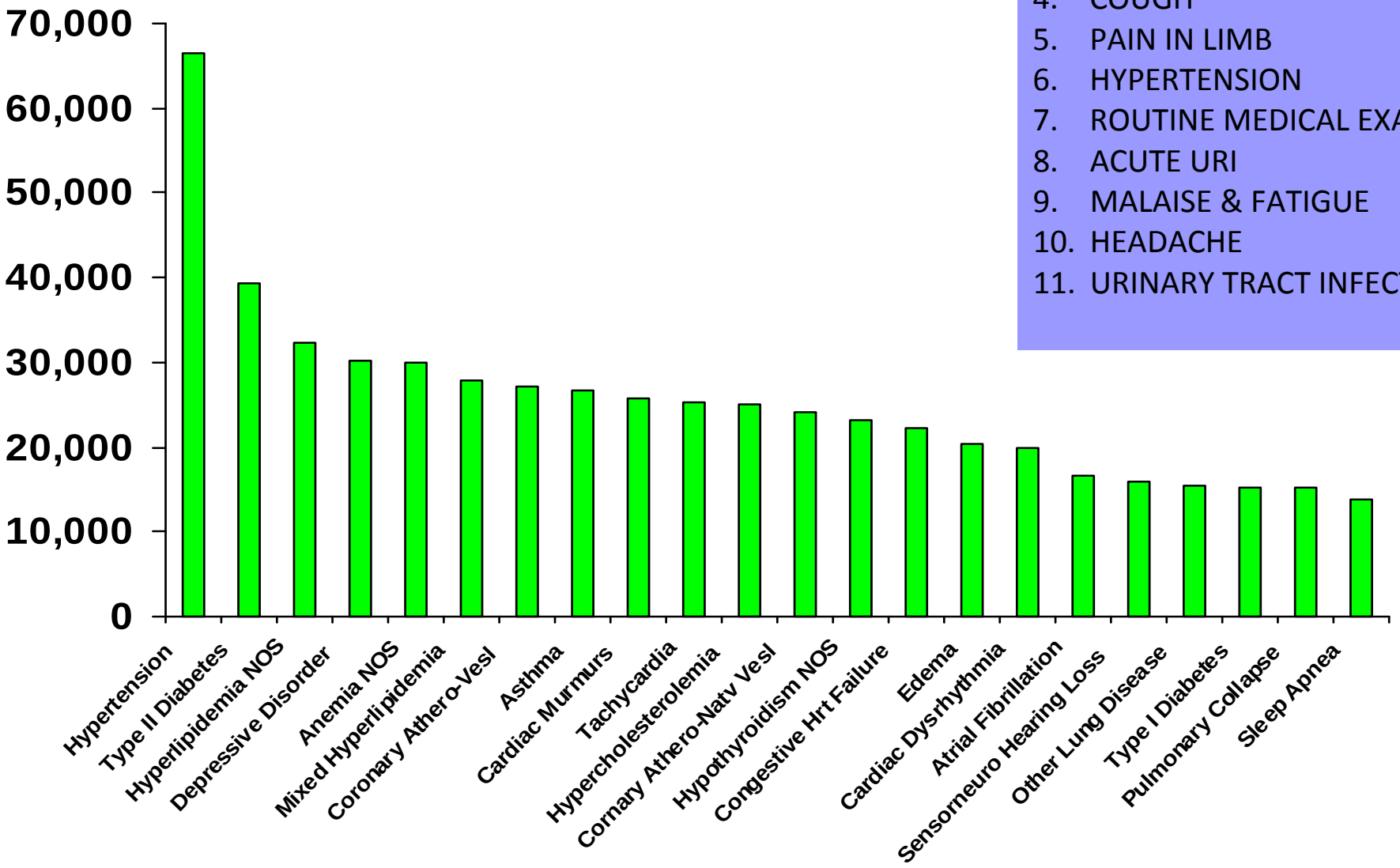
- Support research on date-specific events (outbreaks, seasonal patterns, catastrophes, etc)
- Find a specific patient (e.g. to contact)
- Replace large scale epidemiology research (e.g. TennCare database)
- Temporal search capabilities limited (but under development)
  - “First this, than that” study designs require significant manual effort
  - Expect “timeline” views and searching Q1-Q2

Demographic Characteristics

	SD	Davidson County	Tennessee	United States
N	1,716,085	578,698	6,038,803	299,398,484
Gender (%)				
Female	55.2	51.3	51.1	50.7
Male	44.6	48.7	48.9	48.3
Unknown	0.2	-	-	-
Race/Ethnicity* (%)				
Afr American	14.3	27.9	16.9	12.8
Asian / Pacific	1.2	3.0	1.4	4.6
Caucasian	80.5	60.1	77.5	66.4
Hispanic	2.6	7.1	3.2	14.8
Indian American	0.1	0.4	0.3	1.0
Others	1.4	-	-	-
Multiple Races	0	1.5	1.0	1.6

*A significant number of SD records are of unknown race/ethnicity. Multiple efforts are underway to better classify these records including NLP on narratives.

Examples of frequent diagnoses in total SD

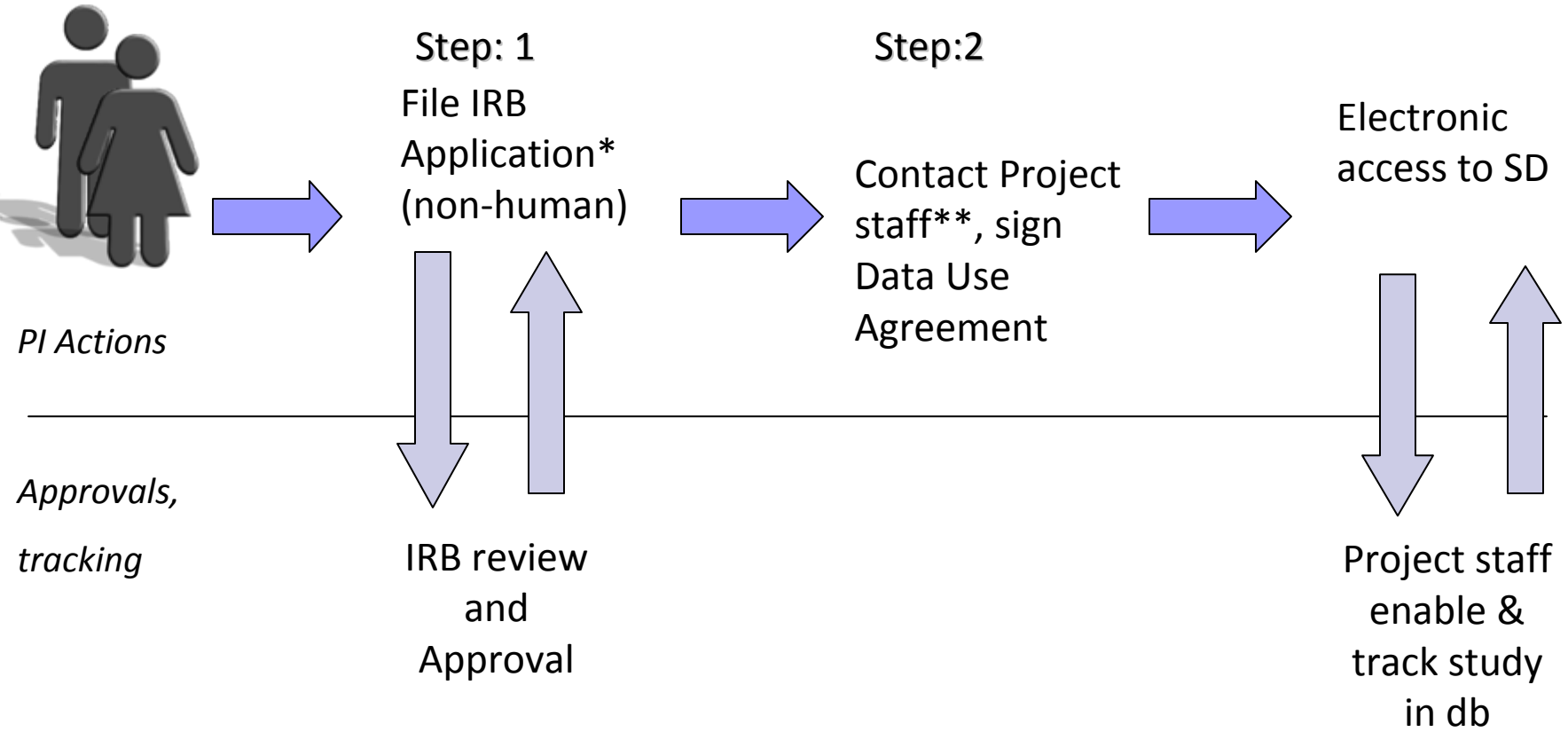


Top diagnosis codes overall:

1. FEVER
2. CHEST PAIN
3. ABDOMINAL PAIN
4. COUGH
5. PAIN IN LIMB
6. HYPERTENSION
7. ROUTINE MEDICAL EXAM
8. ACUTE URI
9. MALAISE & FATIGUE
10. HEADACHE
11. URINARY TRACT INFECTION

Using the SD resource

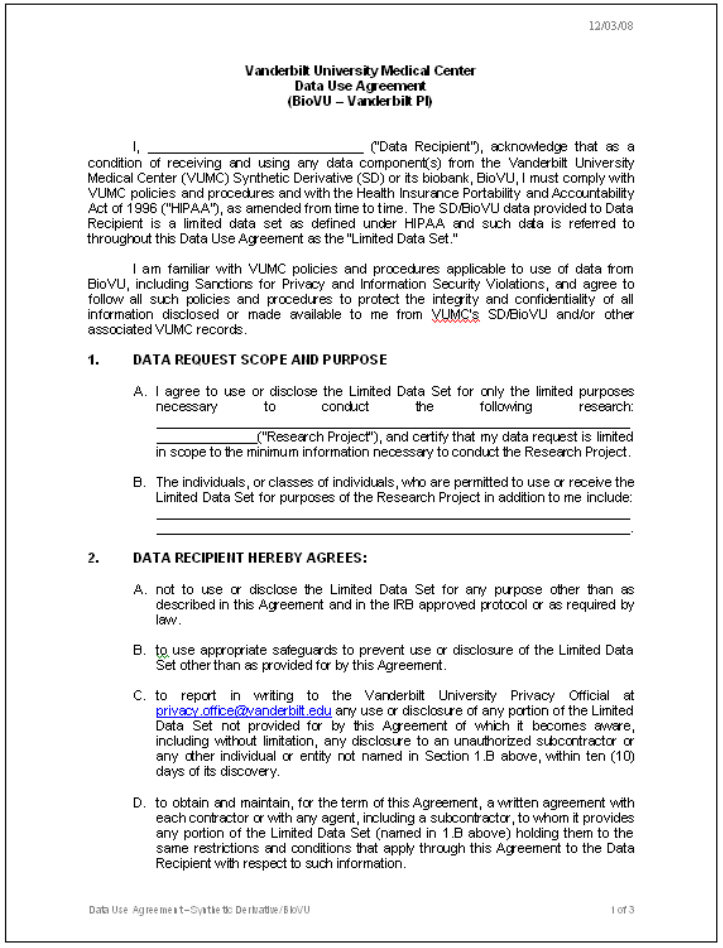
DNA Databank Access Request Process



*Form 1122 – Non Human Non Research

**https://www.mc.vanderbilt.edu/starbrite/medicaldata/

Data Use Agreement Components



12/03/08

Vanderbilt University Medical Center
Data Use Agreement
(BioVU – Vanderbilt PI)

I, _____ ("Data Recipient"), acknowledge that as a condition of receiving and using any data component(s) from the Vanderbilt University Medical Center (VUMC) Synthetic Derivative (SD) or its biobank, BioVU, I must comply with VUMC policies and procedures and with the Health Insurance Portability and Accountability Act of 1996 ("HIPAA"), as amended from time to time. The SD/BioVU data provided to Data Recipient is a limited data set as defined under HIPAA and such data is referred to throughout this Data Use Agreement as the "Limited Data Set."

I am familiar with VUMC policies and procedures applicable to use of data from BioVU, including Sanctions for Privacy and Information Security Violations, and agree to follow all such policies and procedures to protect the integrity and confidentiality of all information disclosed or made available to me from VUMC's SD/BioVU and/or other associated VUMC records.

1. DATA REQUEST SCOPE AND PURPOSE

A. I agree to use or disclose the Limited Data Set for only the limited purposes necessary to conduct the following research: _____ _____("Research Project"), and certify that my data request is limited in scope to the minimum information necessary to conduct the Research Project.

B. The individuals, or classes of individuals, who are permitted to use or receive the Limited Data Set for purposes of the Research Project in addition to me include: _____ _____.

2. DATA RECIPIENT HEREBY AGREES:

A. not to use or disclose the Limited Data Set for any purpose other than as described in this Agreement and in the IRB approved protocol or as required by law.

B. to use appropriate safeguards to prevent use or disclosure of the Limited Data Set other than as provided for by this Agreement.

C. to report in writing to the Vanderbilt University Privacy Official at privacy.office@vanderbilt.edu any use or disclosure of any portion of the Limited Data Set not provided for by this Agreement of which it becomes aware, including without limitation, any disclosure to an unauthorized subcontractor or any other individual or entity not named in Section 1.B above, within ten (10) days of its discovery.

D. to obtain and maintain, for the term of this Agreement, a written agreement with each contractor or with any agent, including a subcontractor, to whom it provides any portion of the Limited Data Set (named in 1.B above) holding them to the same restrictions and conditions that apply through this Agreement to the Data Recipient with respect to such information.

Data Use Agreement–Synthetic Derivative/BioVU 1 of 3

E. not to identify the information contained i
including using Star Panel or other inform
contact any individual whose information i

F. in the event the Data Recipient become
health information unintentionally missed
that the personally identifiable health in
received by Data Recipient, to report al
program staff and/or the privacy office for
any third party.

G. to immediately notify Vanderbilt's Office c
receipt of any request or subpoena for an
information related to this Agreement. T
assume responsibility for challenging the
will cooperate fully with Vanderbilt in any s

H. If genomic sequence data are generated,
any BioVU dataset(s) and DNA samples
year of when access was granted. This w
a robust resource continually enriched by

I. to retain control over the Limited Data Set,
Limited Data Set in any form to any entity
pursuant to the restrictions set forth in
research team member(s) who agree
Agreement, subject to applicable law.
responsibility for ensuring appropriate use

J. that if he/she changes institutions any d
returned and/or destroyed and that da
institution.

K. to acknowledge the Synthetic Derivati
presentations, disclosures, and publicatic
SD/BioVU datasets. A sample statement
"The dataset(s) used for the analyses de
University Medical Center's [INSERT: Bi
supported by institutional funding ar
1UL1RR024975-01 from NCRR/NIH."

3. DATA DISCLAIMER

VUMC disclaims all warranties as to the accuracy of t
performance or fitness of the data for any particular
acknowledges that VUMC does not and cannot warrant the results that may be obtained by using data included in the Data Set.

4. DEFINITIONS

Terms used but not otherwise defined in this Agreement shall have the same meaning as those terms in the Privacy Rule.

Data Use Agreement–Synthetic Derivative/BioVU 2 of 3

12/03/08

B. Individual shall have the same meaning as the term "individual" in 45 CFR Sect. 164.501 of the Privacy Rule and shall include a person who qualifies as a personal representative in accordance with 45 CFR Sect. 164.502(g) of the Privacy Rule.

C. Limited Data Set shall refer to the data requested and used by Data Recipient. The term "limited data set" in 45 CFR 164.514(e) of the Privacy Rule means protected health information that excludes the following direct identifiers of the individual or of relatives, employers, or household members of the individual: names, postal address information (other than town or city, state, and zip code), telephone numbers, fax numbers, electronic mail addresses, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate/license numbers, vehicle identifiers and serial numbers (including license plate numbers), device identifiers and serial numbers, web universal resource locators (URLs), internet protocol (IP) address numbers, biometric identifiers (including finger and voice prints), full face photographic images, and any comparable images.

D. Privacy Rule shall mean the Standards for Privacy of Individually Identifiable Information at 45 CFR Part 160 and Part 164, Subparts A and E, as amended from time to time.

E. Protected Health Information or PHI shall have the same meaning as the term "protected health information" in 45 CFR Sect. 164.501 of the Privacy Rule, to the extent such information is created or received by Data Recipient from Vanderbilt.

F. Required by Law shall have the same meaning as the term "required by law" in 45 CFR Sect. 164.501 of the Privacy Rule.

AGREED TO AND ACCEPTED BY:

Principal Investigator: **Research Project Approval:** _____

_____

IRB #_____
Print Name: _____
Title: _____
Date: _____ Date: _____

(Original to be filed with BioVU)
(PI to retain copy for research file)

Data Use Agreement–Synthetic Derivative/BioVU 3 of 3

The problem with ICD9 codes

- ICD9 give both false negatives and false positives
- False **negatives**:
  - Outpatient billing limited to 4 diagnoses/visit
  - Outpatient billing done by physicians (e.g., takes too long to find the unknown ICD9)
  - Inpatient billing done by professional coders:
    - omit codes that don't pay well
    - can only code problems actually explicitly mentioned in documentation
- False **positives**
  - Diagnoses evolve over time -- physicians may initially bill for suspected diagnoses that later are determined to be incorrect
  - Billing the wrong code (perhaps it is easier to find for a busier clinician)
  - Physicians may bill for a different condition if it pays for a given treatment
    - Example: Anti-TNF biologics (e.g., infliximab) originally not covered for psoriatic arthritis, so rheumatologists would code the patient as having rheumatoid arthritis

Assume that record review will be critical part of any SD study

- A number of incorrect ICD9 codes for RA and MS assigned to patients
- Evolving disease
  - “Recently diagnosed with Susac’s syndrome - prior diagnosis of MS incorrect.” (Notes also included a thorough discussion of MS, ADEM, and Susac’s syndrome.)
- Difference between two doctors:
  - Presurgical admission H&P includes “rheumatoid arthritis” in the past medical history
  - Rheumatology clinic visits notes say the diagnosis is “dermatomyositis” - never mention RA
- Sometimes incorrect diagnoses are propagated through the record due to cutting-and-pasting / note reuse

Phenotype Consult Service

- A CTSA supported service
- Faculty: Josh Denny, MD
  - Program manager/contact: Melissa Basford, MBA
  - Supported by technical staff
- Services:
  - Studios and other expert consults
  - Data mining techniques and strategies
  - Assistance with electronic algorithm development

Live Demo