

The Vanderbilt Synthetic Derivative

a continuously updated de-identified
image of clinical data for research

Daniel R. Masys, MD
Professor and Chair,
Department of Biomedical Informatics
Professor of Medicine
Vanderbilt University School of Medicine

Synthetic Derivative

- A **Derivative**
 - information content reduced by 'scrubbing' identifiers
- With a **Synthetic** 'disinformation' component
 - inserted systematically shifted event dates



BioVU: clinically derived samples and data



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

Biobank design



- Extract DNA from leftover blood samples that have been de-identified
- Samples linked to Synthetic Derivative db
- Enables DNA sample retrieval based on clinical queries
- Enables clinical data retrieval based on genetics

Project features

- Both text and tissues are de-identified, and no link is retained to identities
- Considered non-human subjects research by 45 CFR 46 (“Common Rule”)
 - As non-human subjects research, does not include consent
 - Includes IRB, ethics and other oversight
 - Includes option for individuals to opt-out and extensive public education component



Project features, cont'd

Advantages

- Capable of generating 250,000 samples within 5 years
 - Allows searching of rare events
- Chronology of events preserved
- Rich in phenotypic attributes
- Low cost
- “Billions of observations as a byproduct of healthcare” (Kohane)

Disadvantages

- Implementation complexity
- No re-contact
- No family structure
- No information other than that contained within EMR



Opt Out Procedure on Consent to Treatment Form

I understand that any research using these leftover specimens or tissues will be done in a way that will not identify me or my medical information.

I also understand that if I do not want DNA research to be done using my leftover blood, I need to check the box shown below. If you have questions, please call 1-866-436-4710.

Do not use my leftover blood for the DNA Databank

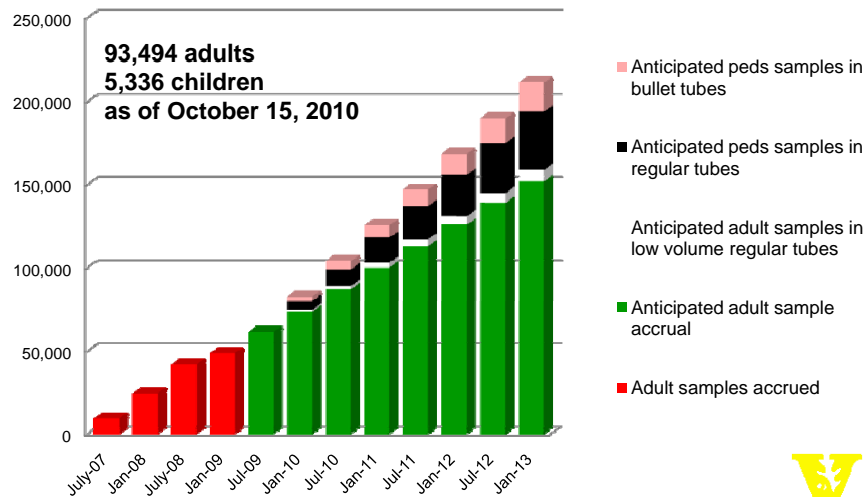
PLEASE READ THIS ENTIRE AUTHORIZATION PRIOR TO SIGNING.

Patient/
Legal Representative _____ Date _____ Time _____ A.M. P.M.

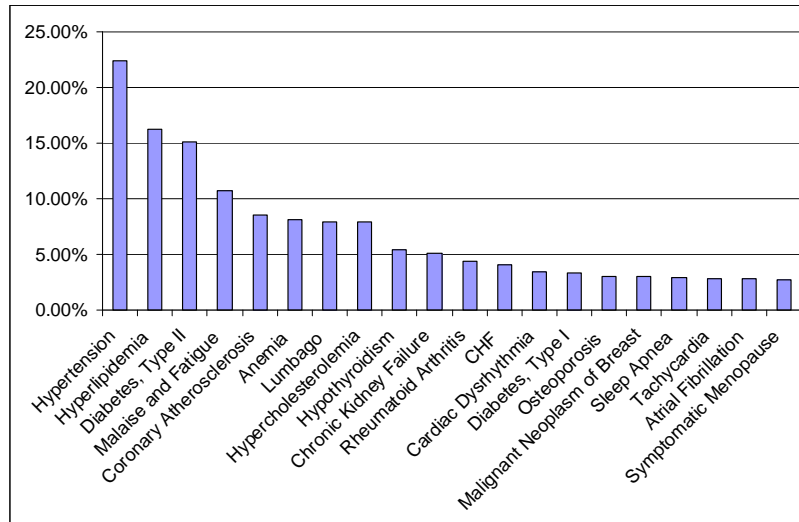
- On average 4.5 - 5% of patients opt out



Sample Accrual: Current and Future



ICD-9 codes (% in BioVU subset of SD)



Creation of the de-identified Synthetic Derivative database

- Uses both structured and full text components of electronic medical record
- Application of (commercial) DE-ID software with pre and post processing
 - Linked naming sources: US Census name list, Regional phone books and street directories, Voter lists, Vanderbilt faculty/staff/student directory
 - Dates shifted so dates are 'wrong' but date intervals correct
- Output is analogous to HIPAA limited dataset – de-id'd, but retains residual re-id potential



Vanderbilt's EMR: StarChart

- Designed, built and maintained by faculty-led teams
- Data on 1.8 million individuals extending back to 70's; comprehensive data for past ten years
- A document-centric architecture with both structured and unstructured elements
- Includes order entry data on inpatients since 1994
- Content converted to RDBMS extract (Oracle) for Synthetic Derivative



The screenshot displays the StarChart EMR interface for patient Hellen Smith. The top navigation bar includes links for 'Pt. Chart', 'StarVisit', 'StarNotes', 'Forms', 'Panels', 'Orders', 'MsgBasket', 'ViewResults', 'SignDrafts', and 'Miscellaneous'. The patient header shows 'SMITH, HELLEN (02/01/1949 - 56YO F) <999-99-9999> (655) 565-5555' with an alert for 'PCP: Mary, Johanson'. A list of orders is visible, including '2004/09/28 Notes' by 'Carter, Meredith' and '2004/09/28 Orders Medication Orders' by 'Carter, Meredith'. The main content area shows an 'Oncology Clinic Note' dated '2004/09/28 14:09' by 'Meredith Carter-Grant, M.D.'. The note text includes a diagnosis of 'Stage II invasive mammary breast cancer "T2 N0 M0."' and an oncologic history section describing the patient's medical background, including a mastectomy in August 2004 and subsequent treatment decisions.



De-Identification Process

Pre-processing: → Licensed DE-ID software: → Post-processing:

- Convert records to DE-ID standard format
- Take off words "cc:" and "sincerely"
- Add **PROTECTED[begin] and **PROTECTED[end] tags to protect all contents should not be scrubbed
- Convert first letter of names all in lower-cases to upper case

- Scrubbed HIPAA identifiers
- Names → **NAME[XXX,YYY]
 - Geographical → **PLACE, **INSTITUTION, **STREET-ADDRESS, **ZIP-CODE
 - Dates → **DATE, **AGE
 - Phone/Fax → **PHONE
 - Email → **EMAIL
 - SSN/MRN//Other IDs → **ID-NUM
 - Device → **DEVICE-ID
 - URL/IP → **WEB-LOC
 - Pathology Specimen # → **PATH-NUMBER

- Take off **PROTECTED[begin] and **PROTECTED[end] tags
- Replace **NAME tag with fake but consistent name randomly picked from common name dictionary
- Replace **ID-NUM tag with fake 9-digit number (under evaluation)
- Replace **PHONE tag with fake phone number (under evaluation)
- Offset **DATE



De-Identification results

200 'large record' test set

	# removed
Names	416,180
Address info more detailed than state or 3 digit zip	29,195
Dates	444,627
Telephone numbers or fax numbers	19,552
E-mail addresses	78
Social security numbers	325
Other numbers (Medical record numbers, Health plan beneficiary numbers, account numbers, certificate/license numbers, vehicle identifiers)	492,009
Device identifiers and serial numbers	165
Web URLs and IP addresses	1,026
Institutions	26,817

Examples under-marked

Pre-scrub	After scrub	Error Type
Rx for Lortab 10, #60 w/ one refill 12/8/4	Rx for Lortab 10, #60 w/ one refill 12/8/4	Date (Complete but malformed)
SOCIAL HISTORY: He currently lives at 77 Spruce Loop; Crossville, Tennessee	SOCIAL HISTORY: He currently lives at 77 Spruce Loop; **PLACE, Tennessee	Street Address
DATE OF BIRTH: 02/22/1912	DATE OF BIRTH: **DATE[Jun 22 1912].	Age Over 90
number of the ventilator is 98141 Patient being monitored with oximetry The	number of the ventilator is 98141 Patient being monitored with oximetry The	Device ID
Severe Left Thigh Hematoma (Traumatic) 6/00	Severe Left Thigh Hematoma (Traumatic) 6/00	Partial date



Examples over-marked

Pre-scrub	After scrub
GI: soft, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly	GI: **PLACE, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly
with iron, 40 gm protein daily, and 1500-2000 calories daily.	with iron, 40 gm protein daily, and **ID-NUM calories daily.
Standardized Balance Tests: BERG Total score: 34 Pt required frequent rest	Standardized Balance Tests: **NAME[XXX: WWW] score: 34 Pt required frequent rest
An attending Cardiologist was present throughout the diagnostic study.	An attending **NAME[SSS] was present throughout the diagnostic study.
filled through the Easter Seals. The patient is also requesting additional	filled through the The patient is also requesting additional

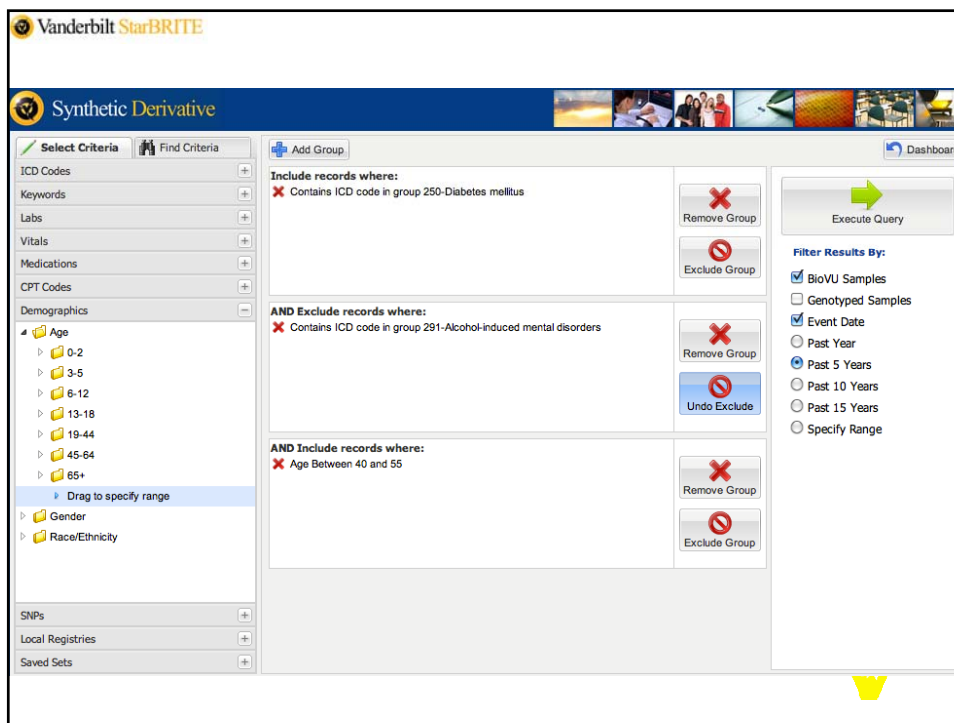
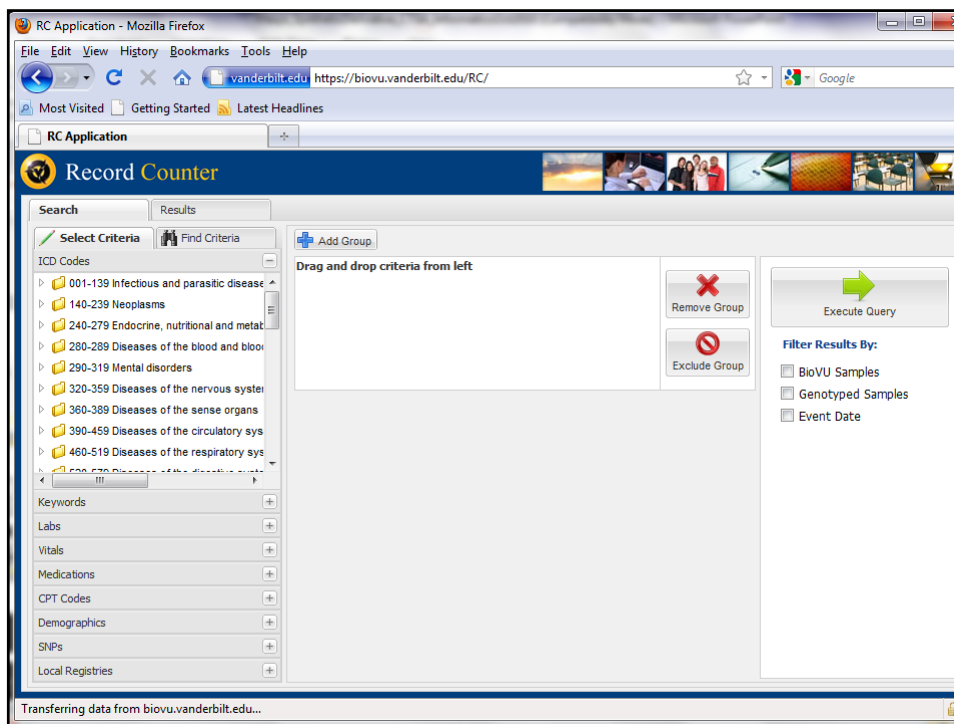


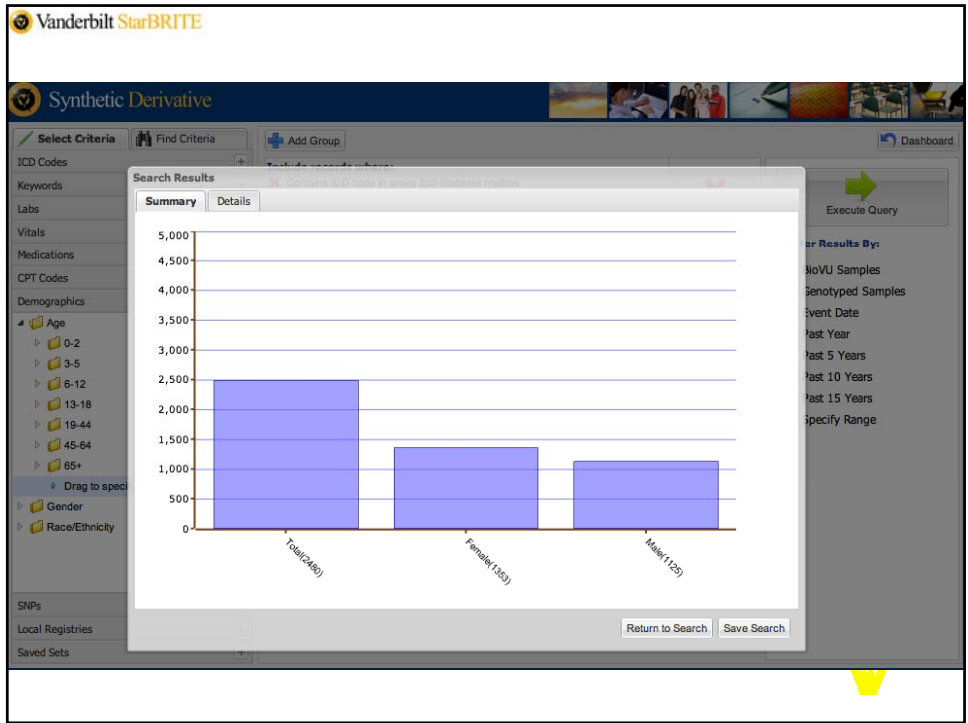
Data security: a combination of technology and policy

- De-identified records access restricted to VU employees; not a public resource
- Databank users sign Data Use Agreement that prohibits use of data for re-identification (similar to HIPAA Limited Data Set use terms)
- Most likely naming source would be clinical information system; if used for re-identification would be HIPAA violation
- Access approved on project-specific basis by Operations Advisory Board (OAB) and IRB
- Project-specific user ID and password; all data access logged and audited by OAB



The screenshot shows a web browser window with the address bar displaying 'https://starbrite.vanderbilt.edu/biovu/'. The page features the Vanderbilt StarBRITE logo and a search bar. A navigation menu includes links for My Research, Governance Dashboard, My Profile, Quick Links, Contact Us, and Logout. Below the menu, there are several content tiles: StarBRITE Home, Planning, Recruitment & Implementation, Research on Practice and Policy, Funding Support, Data Management, Educational Resources, and BioVU & Synthetic Derivative. The main content area is titled 'BioVU & Synthetic Derivative' and contains a sidebar with links to BioVU Home, Application Instructions, Support, Application Status, and Record Counter. A 'Synthetic Derivative Access' button is also present. A statistics box shows '# OF ADULTS in BioVU' as 93,494. The main text area is titled 'BioVU and the Synthetic Derivative' and provides a detailed description of the BioVU resource, stating that it is a centralized resource for making associations between genetic information and disease susceptibility or variable drug responses across populations. It also mentions that genotyping on approximately 7,800 samples at 100 loci (SNPs) has been completed and is now available.





Vanderbilt StarBRITE

Synthetic Derivative

Diabetes_No_Alcohol_Insulin ALL(2480) Not Reviewed(2477) Included(2) Not Included(1)

Previous Record: 4 of 2480 Next Go to Record: Go Go to RUID: Go Dashboard

RUID: Age: Gender: Race/Ethnicity: Deceased: NA Not Reviewed

ICD9 Codes

Date	Code	Description	Age At Event
Code Group : 06			
	06.4	COMPLETE ...	52
Code Group : 226			
	226	BENIGN NE...	52
Code Group : 242			
	242.00	TOX DIF GO...	51
	242	THYROTOXI...	51
	242.00	TOX DIF GO...	51
	242.00	TOX DIF GO...	51
	242.00	TOX DIF GO...	52
	242.00	TOX DIF GO...	52
Code Group : 244			
	244.0	POSTBURGI...	52
	244.0	POSTBURGI...	52
Code Group : 246			
	246.1	DYSHORMO...	51
	246.8	DISORDERS...	52

Problem Lists

Date	Sub Type	Contents
2008-04-01	PROBLEM LIST	---- Known Adverse and Allergic Drug Reactions (if none, enter NKA): PCN (rash)
	PROBLEM LIST	
	PROBLEM LIST	
	PROBLEM LIST	
	PROBLEM LIST	
	PROBLEM LIST	
	PROBLEM LIST	

Page 1 of 2 Displaying 1 - 10 of 18 Expand All

Reports

Date	Sub Type	Contents
		path this am for op
		Date Procedure Performed:
		Ordering Physician:
		**NAME[XXX]
		Attending Physician:
		Source:
		Patient Height:
		Patient Weight:
		Age:

Page 1 of 1 Displaying 1 - 3 of 3 Expand All

Use to date

- Accessible via StarBRITE portal
 - Record Counter requires only university network ID
 - Full Synthetic Derivative access requires project-specific ID granted after project proposal review
- “Record Counter”: 461 users, 3788 saved searches
- Full Synthetic Derivative: 166 users and 828 saved searches



Building a Better De-identification Process

- 2 de-identification models
 - Current: Removal of Identifiers
 - *WHITE-TO-BLACK*
 - Remove Unsafe Terms
 - Favors Recall
 - Most approaches to data scrubbing use this
 - Future: Retention of non-identifiers
 - *BLACK-TO-WHITE*
 - Retain Safe Terms
 - Favors Precision

WHITE-TO-BLACK

Marjorie Long, M.D. St. John's Hospital Huntington 18 Boston, MA 02151	RE: Virginia Townsend CH#32-841-09787 DOB 05/26/86
---	--

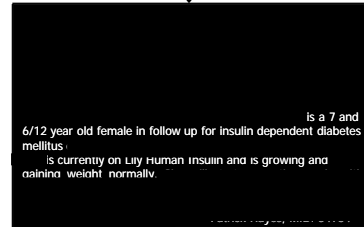
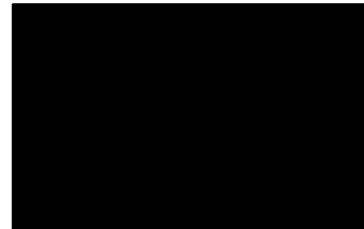
Dear Dr. Lang:
I feel much better after seeing Virginia this time. Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins at Marina Corp 473-1214 for a follow-up visit for her daughter.
Patrick Hayes, M.D. 34764

Dear Dr. [REDACTED]
I feel much better after seeing [REDACTED] this time. [REDACTED] is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. [REDACTED] at [REDACTED]. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact [REDACTED] at [REDACTED] for a follow-up visit for her daughter.

Building a Better De-identification Process

- 2 de-identification models
 - Current: Removal of Identifiers
 - *WHITE-TO-BLACK*
 - Remove Unsafe Terms
 - Favors Recall
 - Most approaches to data scrubbing use this
 - Future: Retention of non-identifiers
 - *BLACK-TO-WHITE*
 - Retain Safe Terms
 - Favors Precision

BLACK-TO-WHITE



Acknowledgements: the SD team

Faculty

- DBMI Faculty
 - Paul Harris, PhD
 - Josh Denny, MD, MS
 - Brad Malin, PhD
 - Hua Xu, PhD
 - Dan Masys, MD
- Other Dept Faculty
 - Dan Roden, MD
 - Jill Pulley, MBA

Staff

- Melissa Basford, MBA
- Jay Cowan
- Sunny Wang
- Nik Nik Hassan

Supported by 3UL1RR024975-03 Vanderbilt Institute for Clinical and Translational Research (VICTR)



