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.

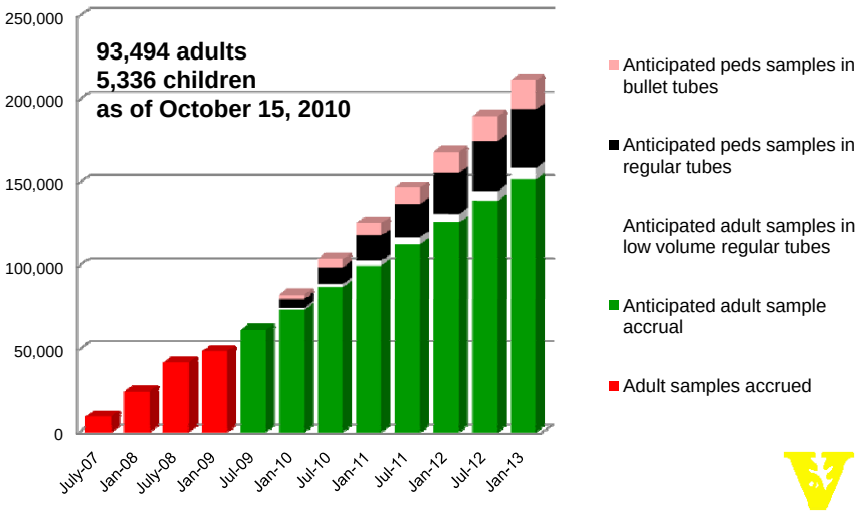
☐ ***Do not use my leftover blood for the DNA Databank***

PLEASE READ THIS ENTIRE AUTHORIZATION PRIOR TO SIGNING.

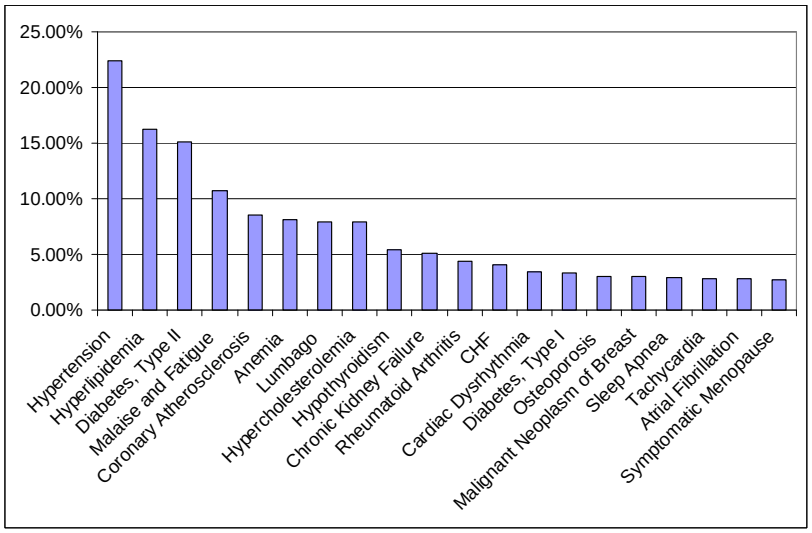
Patient/
Legal Representative _____ Date _____ Time _____ A.M. P.M.

- On average 4.5 - 5% of patients opt out

Sample Accrual: Current and Future



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

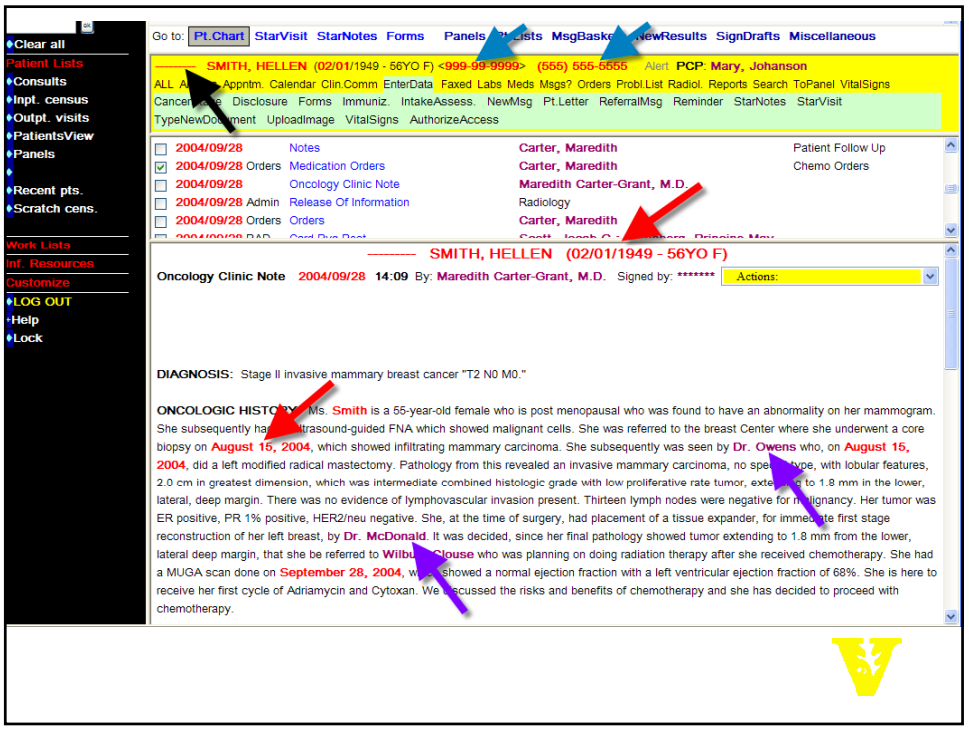


Creation of the de-identified Synthetic Derivative database

- 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



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

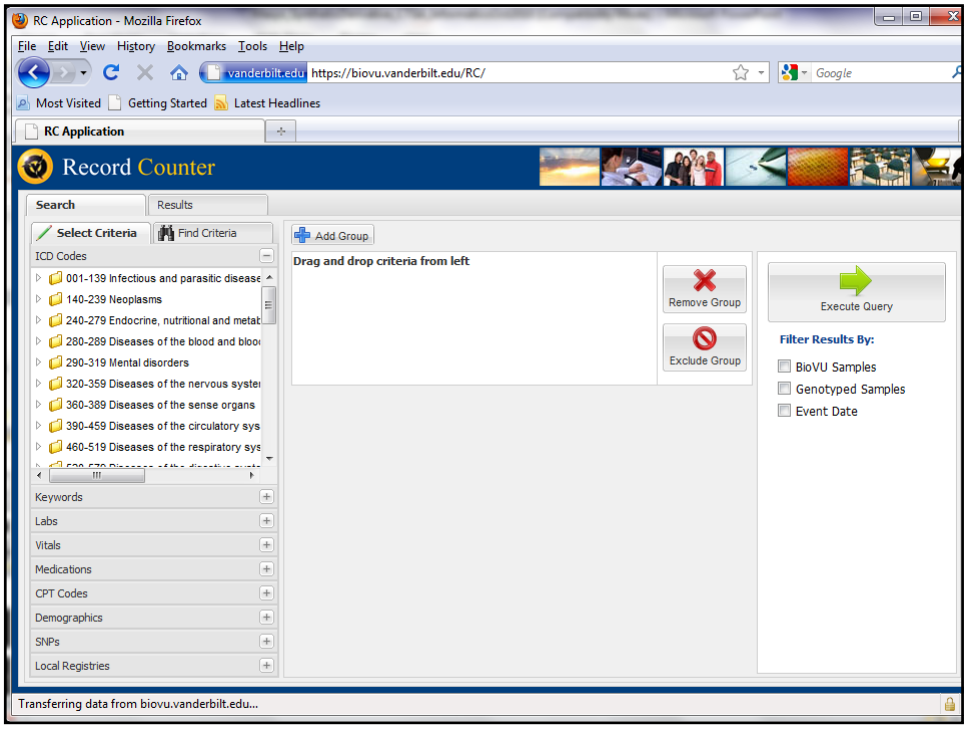
Pre-scrub	After scrub	Error Type
Rx for Lortab 10, #60 w/ one refill 12/8/4	Rx for Lortab 10, #60 w/ one refill 12/8/4	Date (Complete but malformed)
SOCIAL HISTORY: He currently lives at 77 Spruce Loop; Crossville, Tennessee	SOCIAL HISTORY: He currently lives at 77 Spruce Loop; **PLACE, Tennessee	Street 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

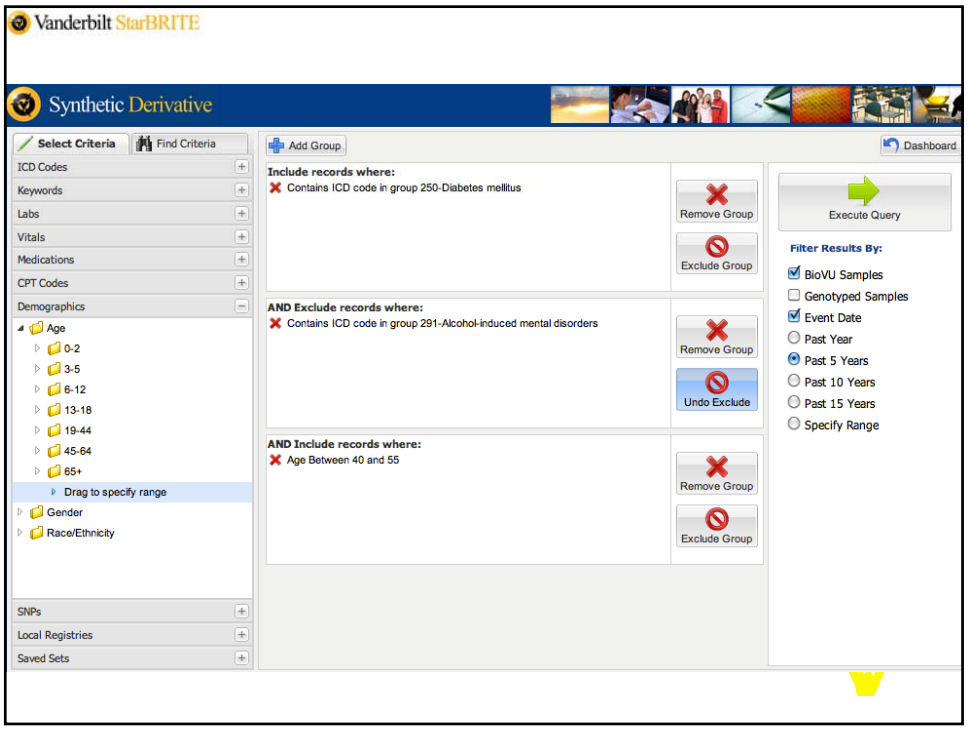
Examples over-marked

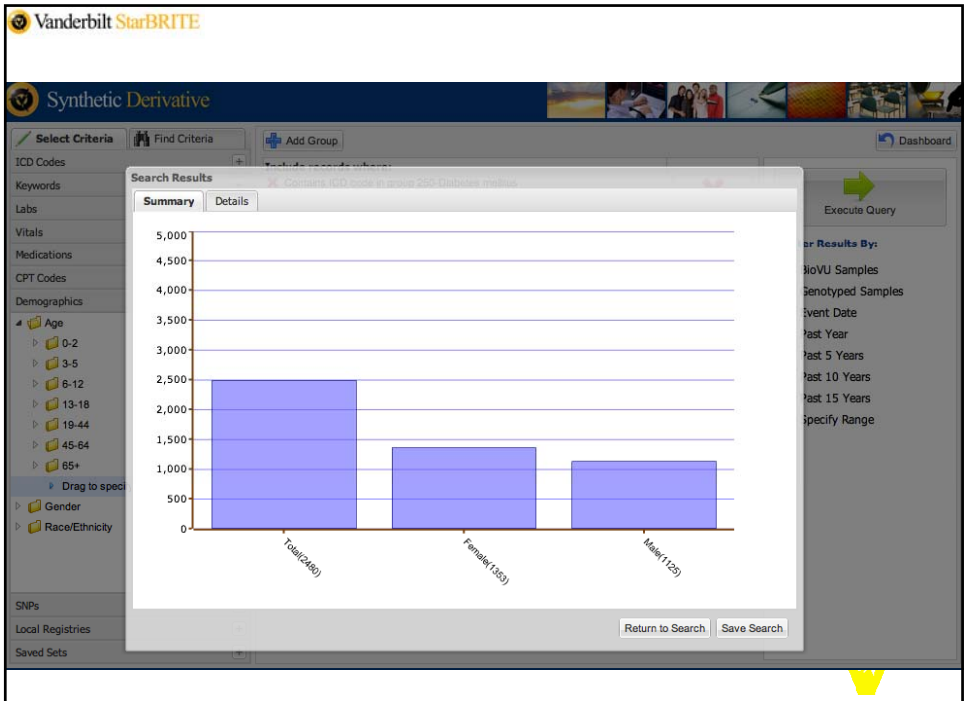
Pre-scrub	After scrub
GI: soft, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly	GI: **PLACE, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly
with iron, 40 gm protein daily, and 1500-2000 calories daily.	with iron, 40 gm protein daily, and **ID-NUM calories daily.
Standarized Balance Tests: BERG Total score: 34 Pt required frequent rest	Standarized 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

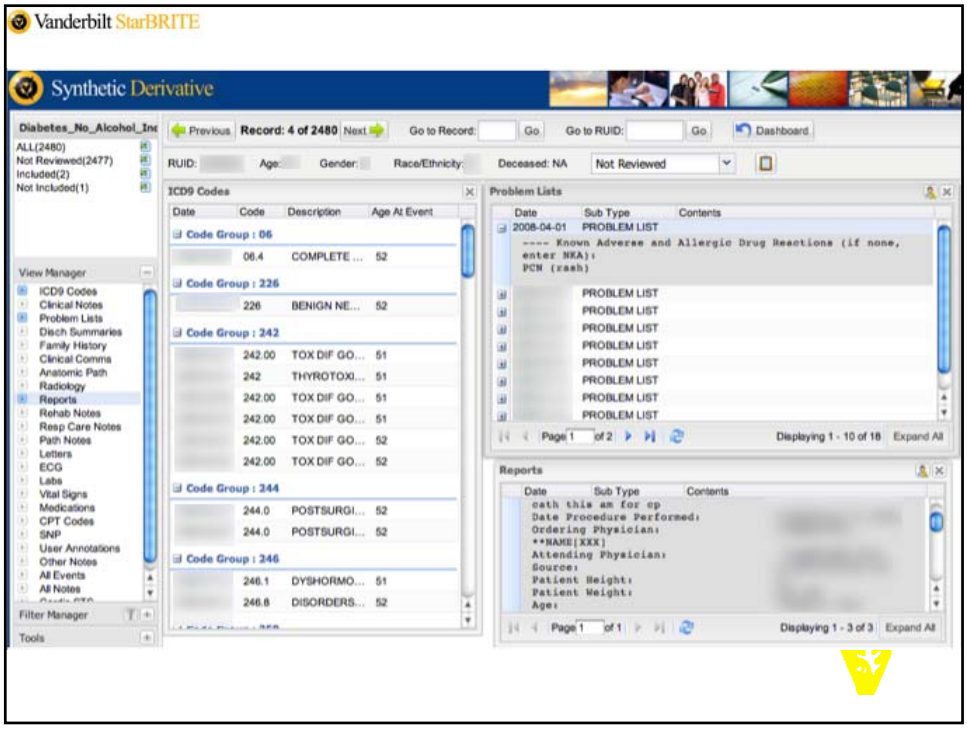
Data security: a combination of technology and policy

- 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


RC Application - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://biovu.vanderbilt.edu/RC/
Most Visited Getting Started Latest Headlines
RC Application
Record Counter
Search
Results
Select Criteria
Find Criteria
ICD Codes
001-139 Infectious and parasitic disease
140-239 Neoplasms
240-279 Endocrine, nutritional and metat
280-289 Diseases of the blood and bloo
290-319 Mental disorders
320-359 Diseases of the nervous syste
360-389 Diseases of the sense organs
390-459 Diseases of the circulatory sys
460-519 Diseases of the respiratory sys
Keywords
Labs
Vitals
Medications
CPT Codes
Demographics
SNPs
Local Registries
Add Group
Drag and drop criteria from left
Remove Group
Exclude Group
Execute Query
Filter Results By:
BioVU Samples
Genotyped Samples
Event Date
Transferring data from biovu.vanderbilt.edu...


Vanderbilt StarBRITE
Synthetic Derivative
Select Criteria
Find Criteria
Add Group
Dashboard
ICD Codes
Keywords
Labs
Vitals
Medications
CPT Codes
Demographics
Age
0-2
3-5
6-12
13-18
19-44
45-64
65+
Drag to specify range
Gender
Race/Ethnicity
SNPs
Local Registries
Saved Sets
Include records where:
Contains ICD code in group 250-Diabetes mellitus
AND Exclude records where:
Contains ICD code in group 291-Alcohol-induced mental disorders
AND Include records where:
Age Between 40 and 55
Remove Group
Exclude Group
Undo Exclude
Execute Query
Filter Results By:
BioVU Samples
Genotyped Samples
Event Date
Past Year
Past 5 Years
Past 10 Years
Past 15 Years
Specify Range


Vanderbilt StarBRITE
Synthetic Derivative
Select Criteria
Find Criteria
Add Group
Dashboard
ICD Codes
Keywords
Labs
Vitals
Medications
CPT Codes
Demographics
Age
0-2
3-5
6-12
13-18
19-44
45-64
65+
Gender
Race/Ethnicity
SNPs
Local Registries
Saved Sets
Search Results
Summary
Details
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Total(2480)
Female(1353)
Male(1125)
Return to Search
Save Search
Execute Query
Past Year
Past 5 Years
Past 10 Years
Past 15 Years
Specify Range


Vanderbilt StarBRITE
Synthetic Derivative
Diabetes_No_Alcohol_Inc
ALL(2480)
Not Reviewed(2477)
Included(2)
Not Included(1)
Previous
Record: 4 of 2480
Next
Go to Record:
Go to RUID:
Dashboard
RUID:
Age:
Gender:
Race/Ethnicity:
Deceased: NA
Not Reviewed
ICD9 Codes
Problem Lists
Reports
View Manager
Filter Manager
Tools

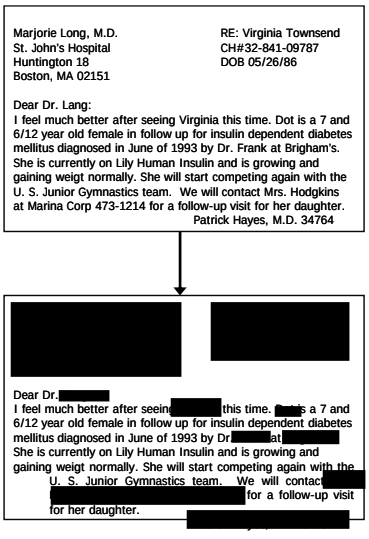
Use to date

- 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 weigt 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

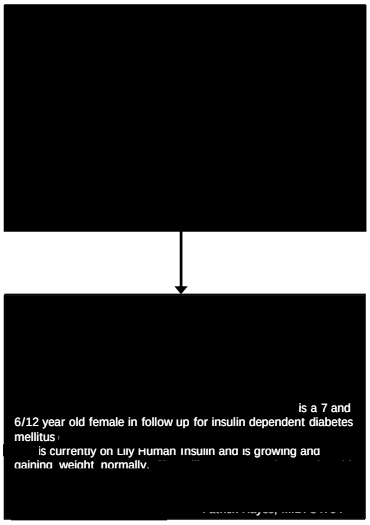
[redacted]

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 weigt normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact [redacted] for a follow-up visit for her daughter.
[redacted]

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



Acknowlegements: 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)