

Genomic Data Sharing: Academic Model Lessons Learned from the eMERGE network

Daniel Masys, MD

Professor and Chair

Department of Biomedical Informatics

Professor of Medicine

Vanderbilt University School of Medicine

Principal Investigator

electronic Medical Records and Genomics (eMERGE)

Coordination Center

Topics

- The eMERGE consortium: on the frontier of genomes and phenotypes derived from clinical sources
- Lessons in progress regarding data sharing
- Data de-identification and re-identification

RFA HG-07-005: Genome-Wide Studies in Biorepositories with Electronic Medical Record Data

- 2007 NIH Request for Applications from the National Human Genome Research Institute
- “The purpose of this funding opportunity is to provide support for **investigative groups affiliated with existing biorepositories to develop necessary methods and procedures for, and then to perform, if feasible, genome-wide studies in participants with phenotypes and environmental exposures derived from electronic medical records**, with the aim of widespread sharing of the resulting individual genotype-phenotype data to accelerate the discovery of genes related to complex diseases.”

eMERGE: an NHGRI-funded consortium for Biobanks linked to EMR data

- Consortium members
 - Group Health of Puget Sound
 - Marshfield Clinic
 - Mayo Clinic
 - Northwestern University
 - Vanderbilt University

The eMERGE Network

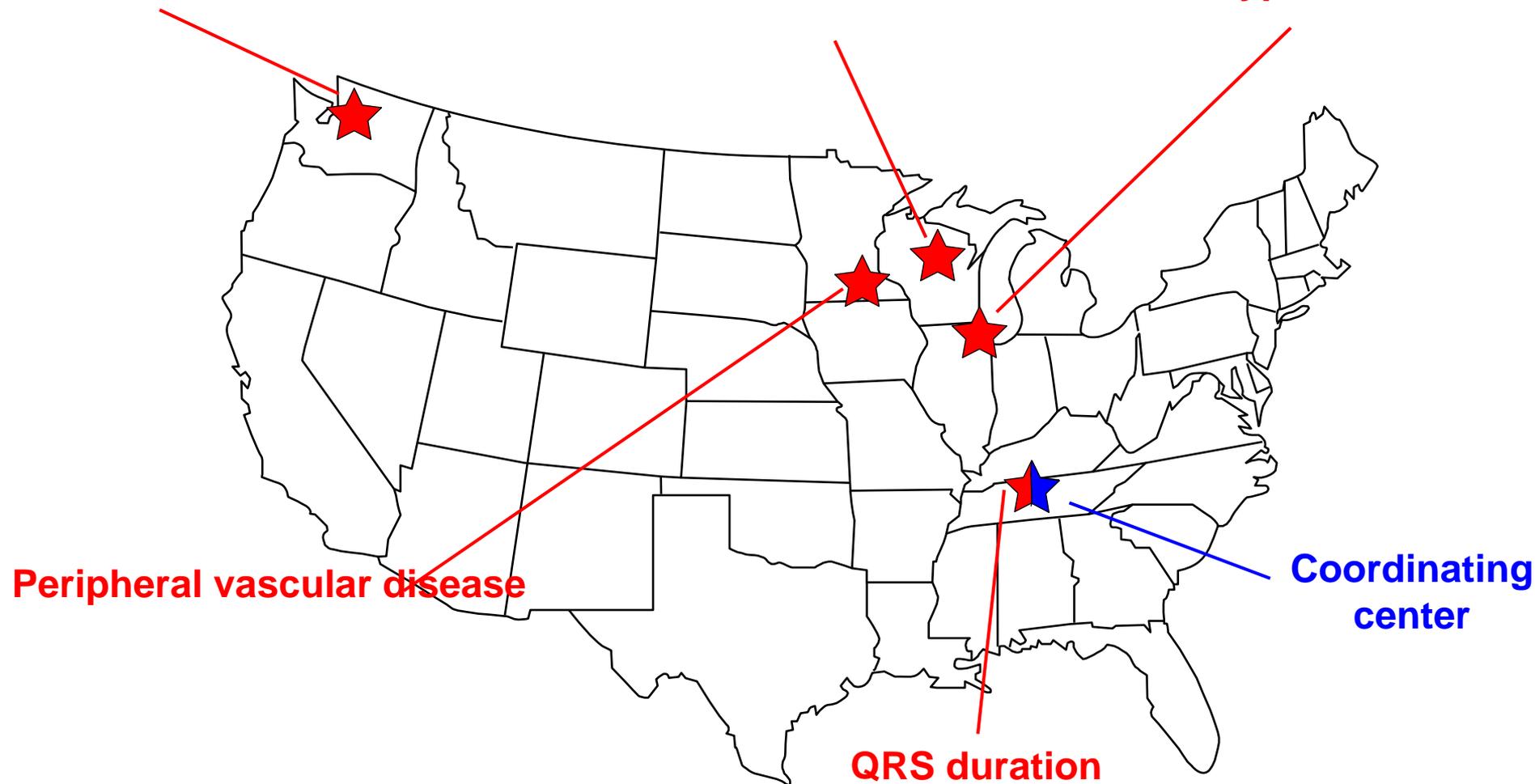
electronic Medical Records & Genomics

*A consortium of biorepositories linked to electronic medical records data
for conducting genomic studies*

Dementia

Cataracts

Type II diabetes



The eMERGE Network

electronic Medical Records & Genomics

*A consortium of biorepositories linked to electronic medical records data
for conducting genomic studies*

- Each site includes DNA linked to electronic medical records
- Each project includes community engagement, genome science, natural language processing capability for EMR data
- Research plans include identifying a phenotype of interest in 3,000 subjects and conduct of a genome-wide association study at each center: $\Sigma=18,000$
- Supplemental funding provided for cross-network phenotypes
- *Condition of NIH funding: contribute genomic and EMR-derived phenotype data to dbGAP database at NCBI*

Supplement phenotypes using genotyped samples from primary phenotypes*

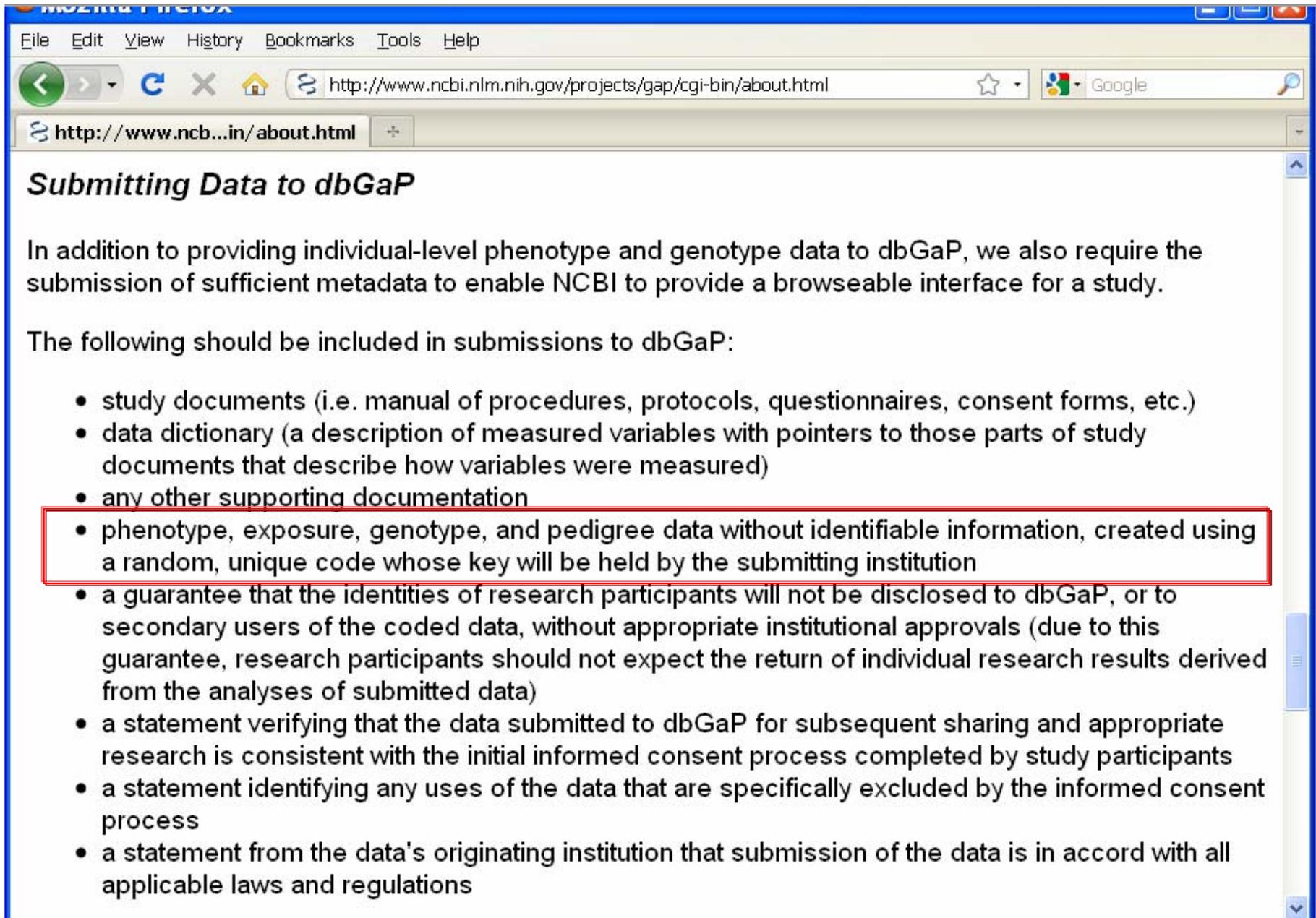
*The benefit of data from routine clinical testing results recorded in EHR

	RBC/WBC	Diabetic Retinopathy	Lipid Levels & Height	GFR
GHC	3,579	230	3,114	1,713
Marshfield	3,865	213	3,693	3,929
Mayo	3,346	806	3,175	3,340
NU	2,484	139	2,816	1,485
VU	2,650	1,449	1,631	2,679

Informatics Issues in eMERGE

- Determination of comparability of patient populations across institutions
- Data exchange standards for phenotype data
- Representation of change over time (repeated measures) and 'clinical uncertainty' for EMR-derived observations (definite – probable – possible for assertion and negation)
- Re-identification potential of clinical data and associated samples: maximizing scientific value while complying with federal privacy protection policies

dbGAP Data Submission Policy



The image is a screenshot of a web browser window. The title bar at the top reads "Mozilla Firefox". The menu bar includes "File", "Edit", "View", "History", "Bookmarks", "Tools", and "Help". The address bar shows the URL "http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html". Below the address bar, there is a search bar with the Google logo and the text "http://www.ncb...in/about.html". The main content area of the browser displays the following text:

Submitting Data to dbGaP

In addition to providing individual-level phenotype and genotype data to dbGaP, we also require the submission of sufficient metadata to enable NCBI to provide a browseable interface for a study.

The following should be included in submissions to dbGaP:

- study documents (i.e. manual of procedures, protocols, questionnaires, consent forms, etc.)
- data dictionary (a description of measured variables with pointers to those parts of study documents that describe how variables were measured)
- any other supporting documentation
- phenotype, exposure, genotype, and pedigree data without identifiable information, created using a random, unique code whose key will be held by the submitting institution
- a guarantee that the identities of research participants will not be disclosed to dbGaP, or to secondary users of the coded data, without appropriate institutional approvals (due to this guarantee, research participants should not expect the return of individual research results derived from the analyses of submitted data)
- a statement verifying that the data submitted to dbGaP for subsequent sharing and appropriate research is consistent with the initial informed consent process completed by study participants
- a statement identifying any uses of the data that are specifically excluded by the informed consent process
- a statement from the data's originating institution that submission of the data is in accord with all applicable laws and regulations

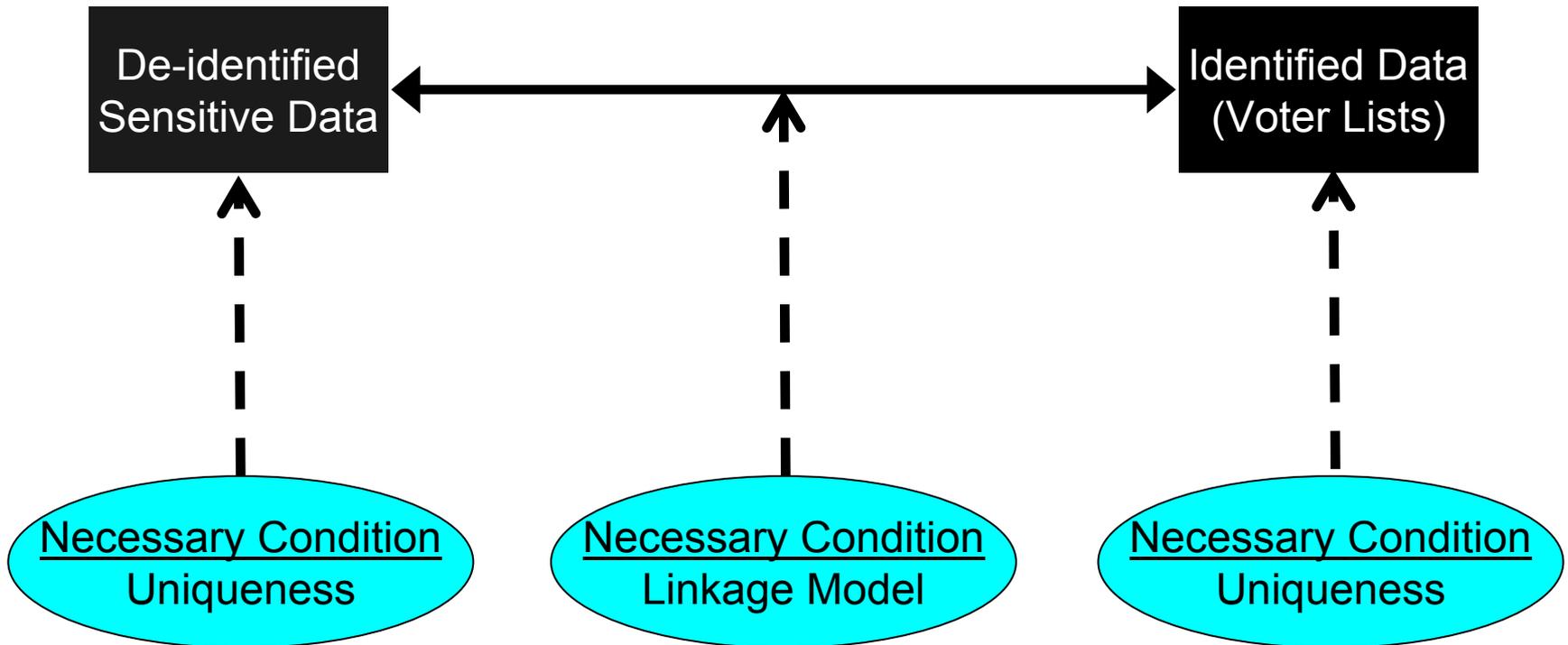
Anonymous vs. De-identified

- **Anonymous** = not identifiable or traceable to an individual
 - A concept prevalent from 5000 BC to about 2000 AD
 - A dichotomous variable: either data is anonymized or it is not
- **De-identified**
 - A concept that has largely replaced anonymous
 - Recognizes that biological data is so inherently rich in attributes that re-identification potential never goes to zero
 - A continuous variable whose properties can be calculated for some (but not all) types of health data

Re-identification of de-identified information

- Requires:
 1. Establishment of *uniqueness* of a collection of data/attributes (“logical unit record”) associated with an individual
 2. A *naming source* that is part of or linkable to the collection described in 1.
- As a result, de-identification methods are generally aimed at either preventing isolation of unique records, blockage of links to naming sources, or both.

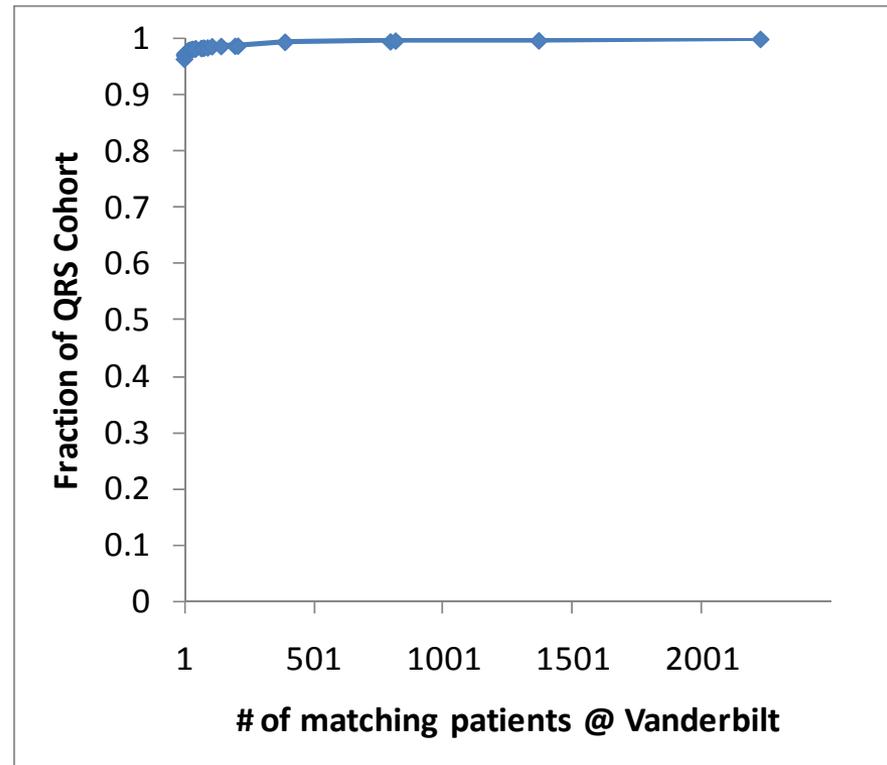
Central Dogma of Re-identification



*Malin B, Kantarcioglu M, Cassa C. A survey of challenges and solutions for privacy in clinical genomics data mining. Chapter in *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*. CRC Press. 2010.

Steps to re-identification: Leveraging Diagnosis Codes to establish uniqueness*

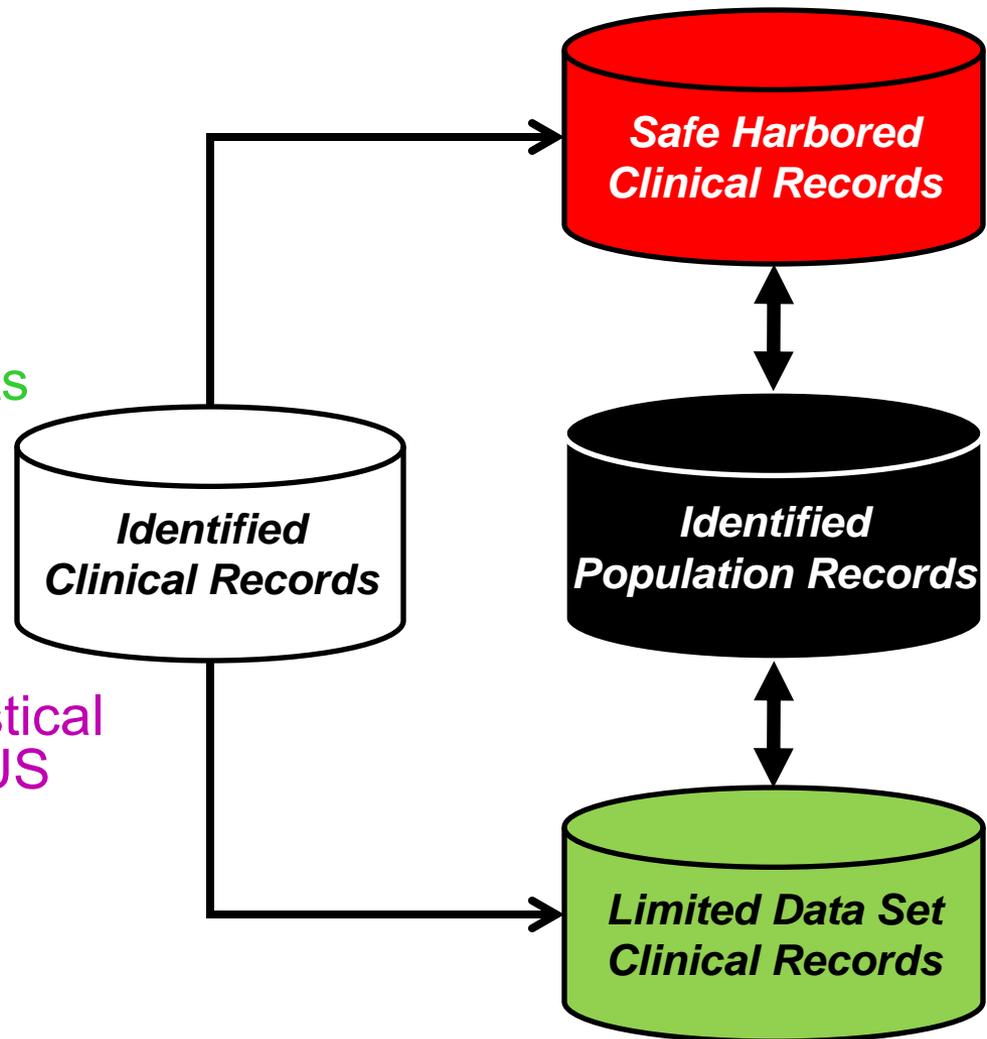
- Cohort: ~2500 Vanderbilt patients in a GWAS
- Each individual in the cohort has set of ICD-9 codes
- Evaluated for “distinctiveness” with respect to entire Vanderbilt population (1.5 million)
- ~97% of individuals are unique



*Loukides G, Denny J, & Malin B. Do clinical profiles constitute privacy risks for research participants? *AMIA Fall Symposium*. 2009.

Re-identification Risk via Demographics: Consider the HIPAA Policies

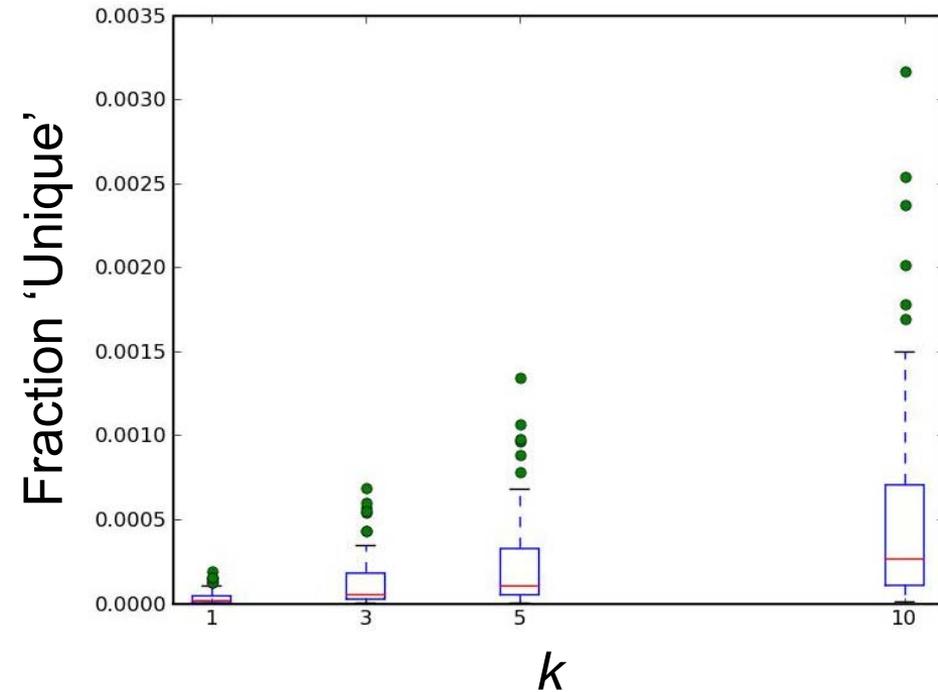
- **HIPAA Safe Harbor**
 - Race
 - Gender
 - Year of Birth
 - State
- **HIPAA Limited Data Sets**
 - Race
 - Gender
 - Date of Birth
 - County
- **Analysis based on statistical approx.* through 2000 US Census**



All States combined (US Census data)

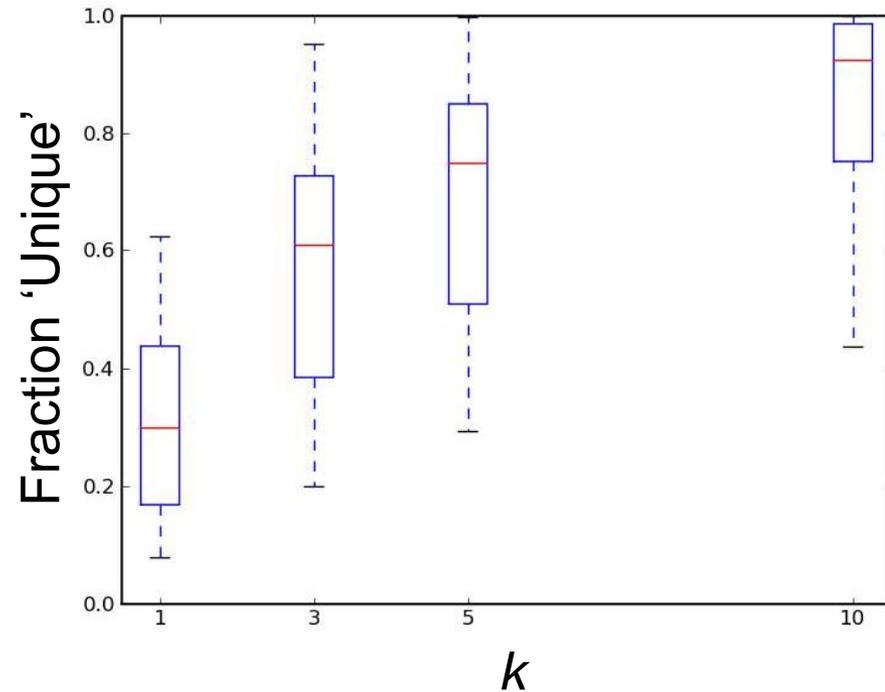
HIPAA Safe Harbor

[Year of Birth, Sex, Race]



HIPAA Limited Data Set

[Date of Birth, Sex, Race, County]



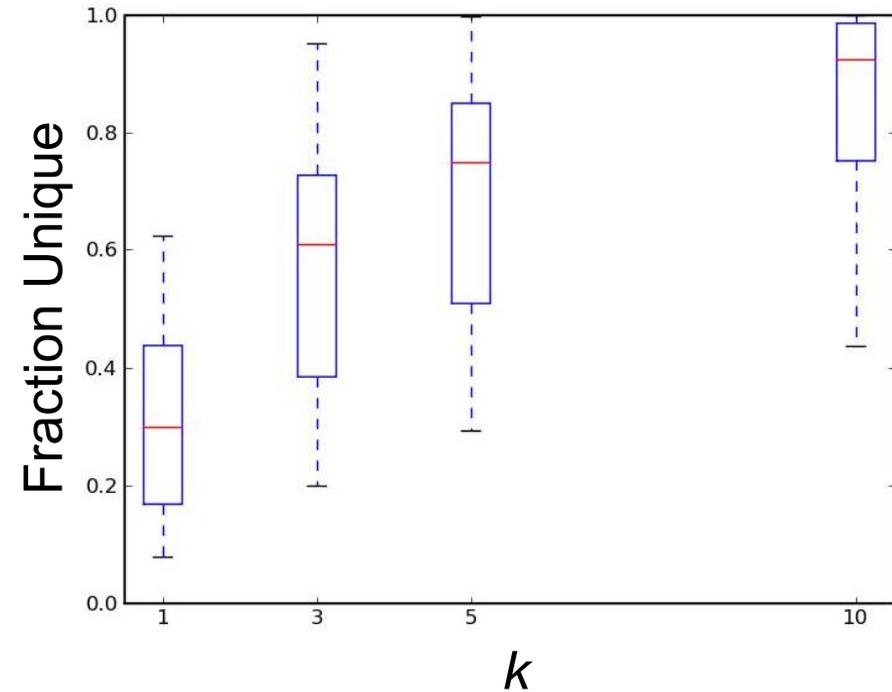
Sources of demographic identifiers

	IL	MN	TN	WA	WI
Authorized Users	Registered Political Committees (ANYONE – In Person)	MN Voters	Anyone	Anyone	Anyone
Format	Disk	Disk	Disk	Disk	Disk
Cost	\$500	\$46; “use ONLY for elections, political activities, or law enforcement”	\$2500	\$30	\$12,500
Voter ID	●	●	●	●	●
Name	●	●	●	●	●
Address	●	●	●	●	●
Voter Status	●	●	●	●	●
District Information	●	●	●	●	●
Election History	●	●	●	●	●
Date of Birth	●	○	●	●	
Date of Registration	●	●	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

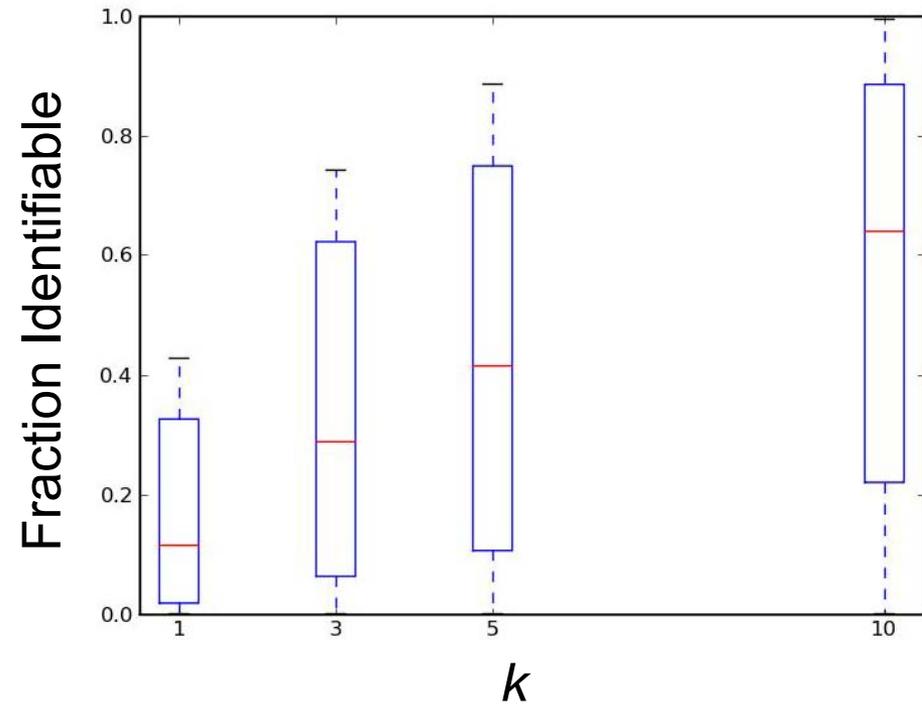
Linking unique records to naming sources

Limited Data Set

[Date of Birth, Sex, Race, County]



Limited Data Set linked to Voter Reg.



eMERGE data sharing procedures

- Clinical data (e.g., ICD9 diagnosis codes) shared with dbGAP are a subset of those present in the full clinical record, with uncommon codes that support elevated re-identification risk removed.
- eMERGE coordination center provides data privacy consultation to network members, including quantitative determination of re-identification risk of each submitted dataset.
- Generalizable tools and methods for determining re-identification risk are in development and testing, and will be freely available

The eMERGE Network

electronic Medical Records & Genomics

*A consortium of biorepositories linked to electronic medical records data
for conducting genomic studies*

Main Menu

[Home](#)

[Member Sites](#)

[About eMERGE Network](#)

[Links](#)

[Contact Us](#)

User Menu

[Log in](#)

The eMERGE Network

The eMERGE Network is a national consortium formed to develop, disseminate, and apply approaches to research that combine DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research.

The mapping of the human genome has enabled new exploration of how genetic variations contribute to health and disease. To better realize this promise, researchers must now determine ways in which genetic make-up gives some individuals a greater chance of becoming sick with chronic conditions such as diabetes, Alzheimer's, or heart disease, in order to ultimately improve patient care.

There are a number of studies conducted routinely to uncover the association between disease and a person's genetic make-up, but they are typically costly and take a long time to complete. This consortium will use data from the EMR – clinical systems that represent actual health care events, an alternative methodology, which is highly cost and time-efficient, to propel this research. Electronic medical records are one of the most exciting potential resources for research data.

Each center participating in the consortium, organized by the National Human Genome Research Institute



URL:

www.gwas.net