MAYO CLINIC

eMERGE Data Dictionary Harmonization and Best Practices for Standardized Phenotype Data Representation

Jyoti Pathak

April 13th, 2010

MAYO CLINIC

Acknowledgment

- Chris Chute
- Dan Masys
- Janey Wang
- Sudha Kashyap
- Melissa Basford

MAYO CLINIC

Overall Objective

- Without terminology standards:
  - Health data is non-comparable
  - Health systems cannot meaningfully interoperate
  - Secondary uses of data for research and applications (e.g., clinical decision support) is not possible
- Our goal: Standardized and consistent representation of eMERGE network-wide phenotype data submitted to dbGaP

MAYO CLINIC

Standardized Resources

- NCI caDSR (Cancer Data Standards Repository)
- CDISC SDTM (Study Data Tabulation Model)
- NCI Thesaurus
- SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms)

MAYO CLINIC

Background: Clinical Terminology Standards and Resources

- NCI Cancer Data Standards Repository
  - Metadata registry based on ISO/IEC 11179 standard for storing common data elements (CDEs)
  - Allows creating, editing, deploying, and finding of CDEs
  - Provides the backbone for NCI's semantic-computing environment, including caBIG (Cancer Biomedical Informatics Grid)
  - Approx. 40,000 CDEs

MAYO CLINIC

Background: Clinical Terminology Standards and Resources

- CDISC Terminology
  - To define and support terminology needs of the CDISC models across the clinical trial continuum
  - Used as part of the Study Data Tabulation Model: an international standard for clinical research data, approved by the FDA as a standard electronic submission format
  - Comprises approx. 2300 terms covering demographics, interventions, findings, events, trial design, units, frequency, and ECG terminology

MAYO CLINIC

Background: Clinical Terminology Standards and Resources

- NCI Thesaurus
  - Reference terminology for clinical care, translational and basic cancer research
  - Comprises approx. 70,000 concepts representing information for nearly 10,000 cancers and related diseases
  - NCI Enterprise Vocabulary Services (LexEVS) provides the terminology infrastructure for caBIG, NCBO etc.

MAYO CLINIC

Background: Clinical Terminology Standards and Resources

- SNOMED-CT
  - Systematized Nomenclature of Medicine Clinical Terms is a comprehensive terminology covering most areas of clinical information including diseases, findings, procedures, microorganisms, pharmaceuticals etc.
  - Comprises approx. 370,000 concepts
  - Acquired by International Health Terminology Standards Development Organization (IHTSDO) in 2007

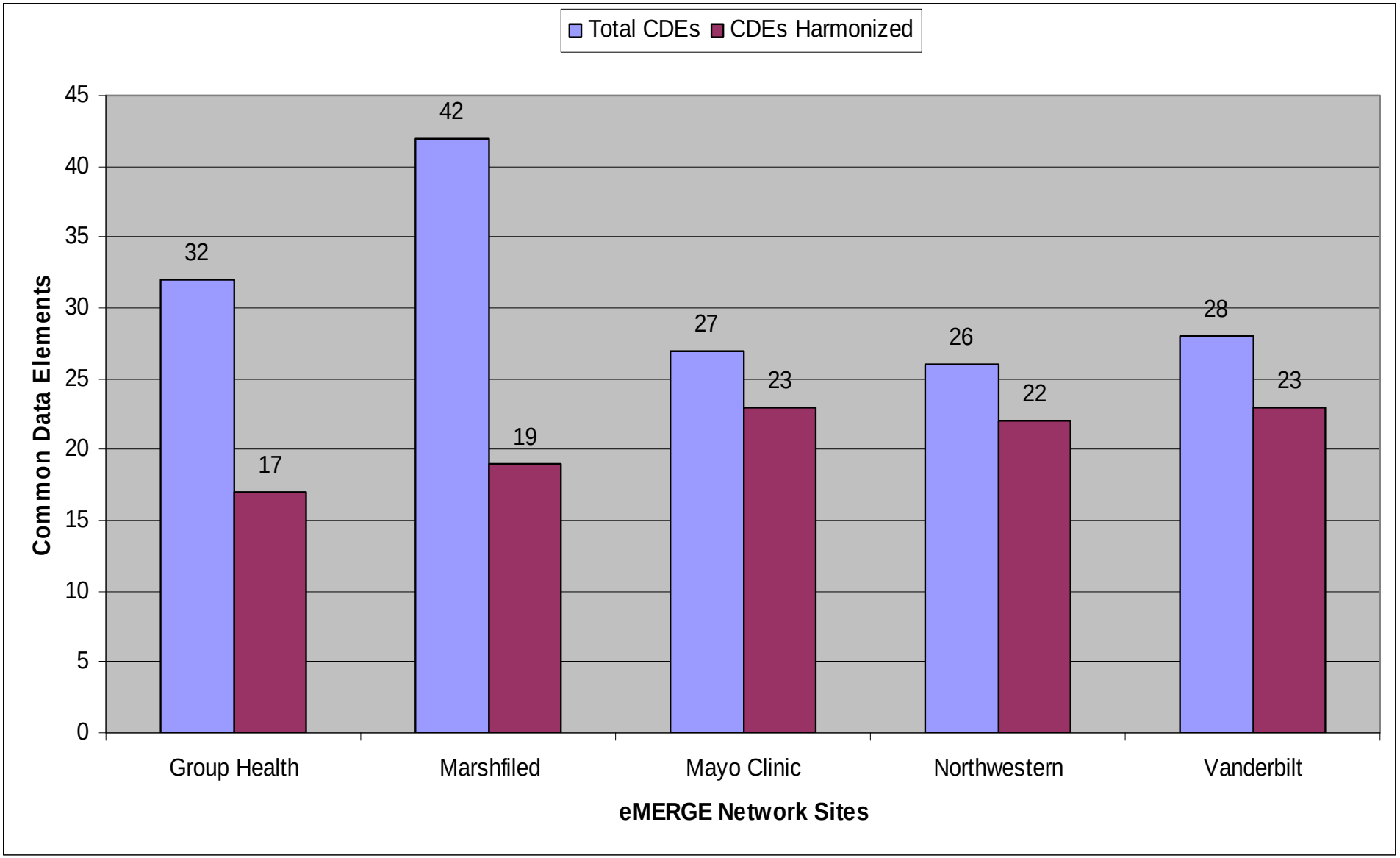
MAYO CLINIC

Methods: DD Harmonization

- Collected the latest data dictionaries from all the eMERGE sites
- Preliminary cleaning
  - Uniform variable names (e.g., BMI vs. Body_Mass_Index)
  - Added new variables as required (e.g., Observation_Age)
- Manual mapping of DD variables and permissible values
  - String search
  - Pre-existing mappings

MAYO CLINIC

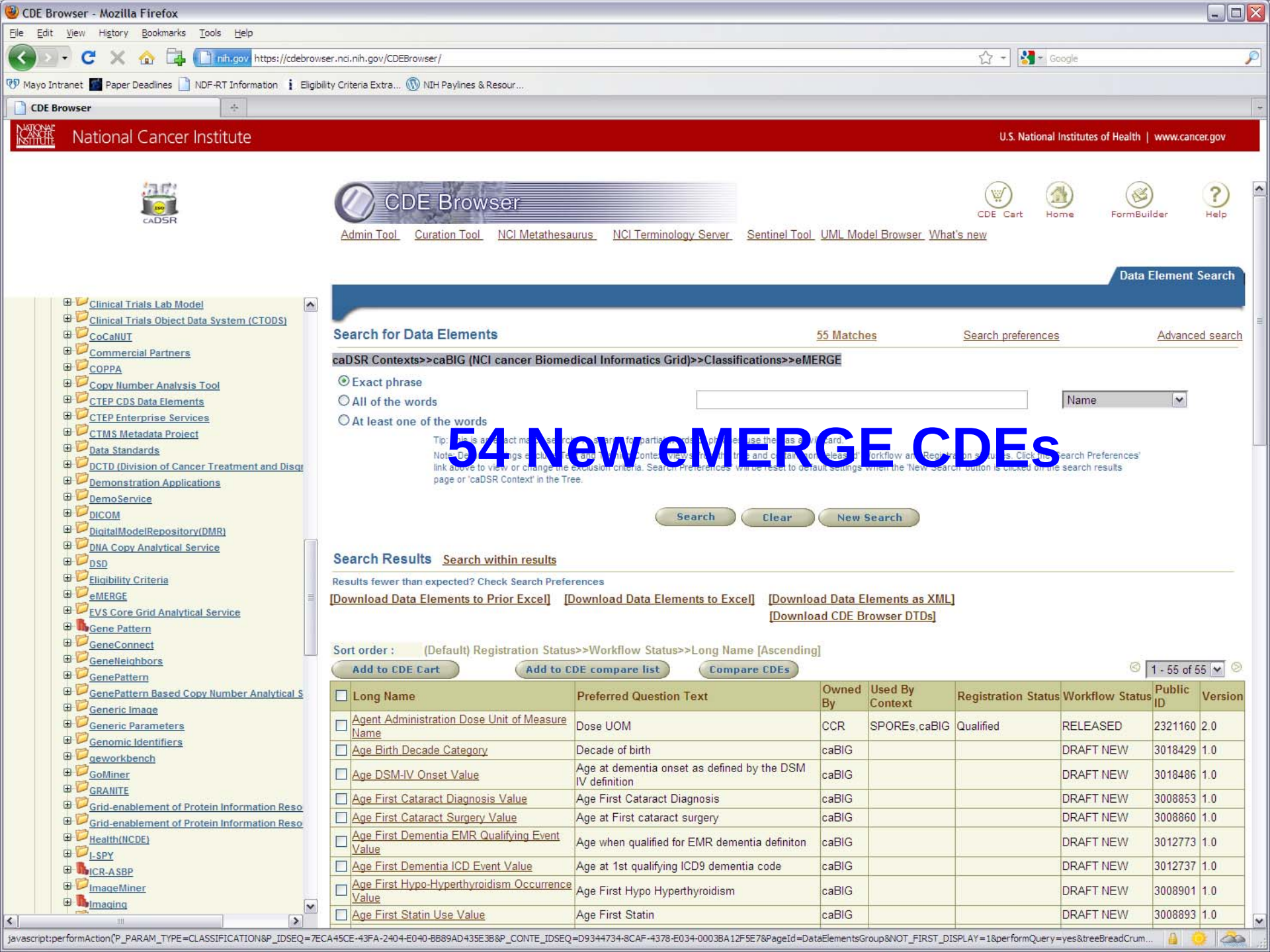
Results: Preliminary Mapping (11/2009)



MAYO CLINIC

Observations

- One size does not fit all; coverage is not uniform across all the standards
- High degree of overlap for commonly used enumerated variables (e.g., Race)
- Additional curation is required to improve coverage for eMERGE data elements
  - Communicated with the caDSR and NCI Thesaurus teams
  - 54 new eMERGE-specific data elements in caDSR were created during 11/2009 and 03/2010
    - Released under “draft” status
  - 9 new NCI Thesaurus concepts


CDE Browser - Mozilla Firefox
https://cdebrowser.nci.nih.gov/CDEBrowser/
National Cancer Institute
U.S. National Institutes of Health | www.cancer.gov
CDE Browser
Admin Tool Curation Tool NCI Metathesaurus NCI Terminology Server Sentinel Tool UML Model Browser What's new
CDE Cart Home FormBuilder Help
Data Element Search
Search for Data Elements
55 Matches
Search preferences
Advanced search
caDSR Contexts>>caBIG (NCI cancer Biomedical Informatics Grid)>>Classifications>>eMERGE
Exact phrase
All of the words
At least one of the words
Name
54 New eMERGE CDEs
Search
Clear
New Search
Search Results
Search within results
Results fewer than expected? Check Search Preferences
[Download Data Elements to Prior Excel] [Download Data Elements to Excel] [Download Data Elements as XML]
[Download CDE Browser DTDs]
Sort order : (Default) Registration Status>>Workflow Status>>Long Name [Ascending]
Add to CDE Cart
Add to CDE compare list
Compare CDEs
1 - 55 of 55
Long Name
Preferred Question Text
Owned By
Used By Context
Registration Status
Workflow Status
Public ID
Version
Agent Administration Dose Unit of Measure Name
Dose UOM
CCR
SPOREs,caBIG
Qualified
RELEASED
2321160
2.0
Age Birth Decade Category
Decade of birth
caBIG
DRAFT NEW
3018429
1.0
Age DSM-IV Onset Value
Age at dementia onset as defined by the DSM IV definition
caBIG
DRAFT NEW
3018486
1.0
Age First Cataract Diagnosis Value
Age First Cataract Diagnosis
caBIG
DRAFT NEW
3008853
1.0
Age First Cataract Surgery Value
Age at First cataract surgery
caBIG
DRAFT NEW
3008860
1.0
Age First Dementia EMR Qualifying Event Value
Age when qualified for EMR dementia definiton
caBIG
DRAFT NEW
3012773
1.0
Age First Dementia ICD Event Value
Age at 1st qualifying ICD9 dementia code
caBIG
DRAFT NEW
3012737
1.0
Age First Hypo-Hyperthyroidism Occurrence Value
Age First Hypo Hyperthyroidism
caBIG
DRAFT NEW
3008901
1.0
Age First Statin Use Value
Age First Statin
caBIG
DRAFT NEW
3008893
1.0
Clinical Trials Lab Model
Clinical Trials Object Data System (CTODS)
CoCaNUT
Commercial Partners
COPPA
Copy Number Analysis Tool
CTEP CDS Data Elements
CTEP Enterprise Services
CTMS Metadata Project
Data Standards
DCTD (Division of Cancer Treatment and Disgr
Demonstration Applications
DemoService
DICOM
DigitalModelRepository(DMR)
DNA Copy Analytical Service
DSD
Eligibility Criteria
eMERGE
EVS Core Grid Analytical Service
Gene Pattern
GeneConnect
GeneNeighbors
GenePattern
GenePattern Based Copy Number Analytical S
Generic Image
Generic Parameters
Genomic Identifiers
geworkbench
GoMiner
GRANITE
Grid-enablement of Protein Information Reso
Grid-enablement of Protein Information Reso
Health(NCDE)
I-SPY
ICR-ASBP
ImageMiner
Imaging
javascript:performAction('P_PARAM_TYPE=CLASSIFICATION&P_IDSEQ=7ECA45CE-43FA-2404-E040-BB89AD435E3B&P_CONTE_IDSEQ=D9344734-8CAF-4378-E034-0003BA12F5E7&PageId=DataElementsGroup&NOT_FIRST_DISPLAY=1&performQuery=yes&treeBreadCrum...

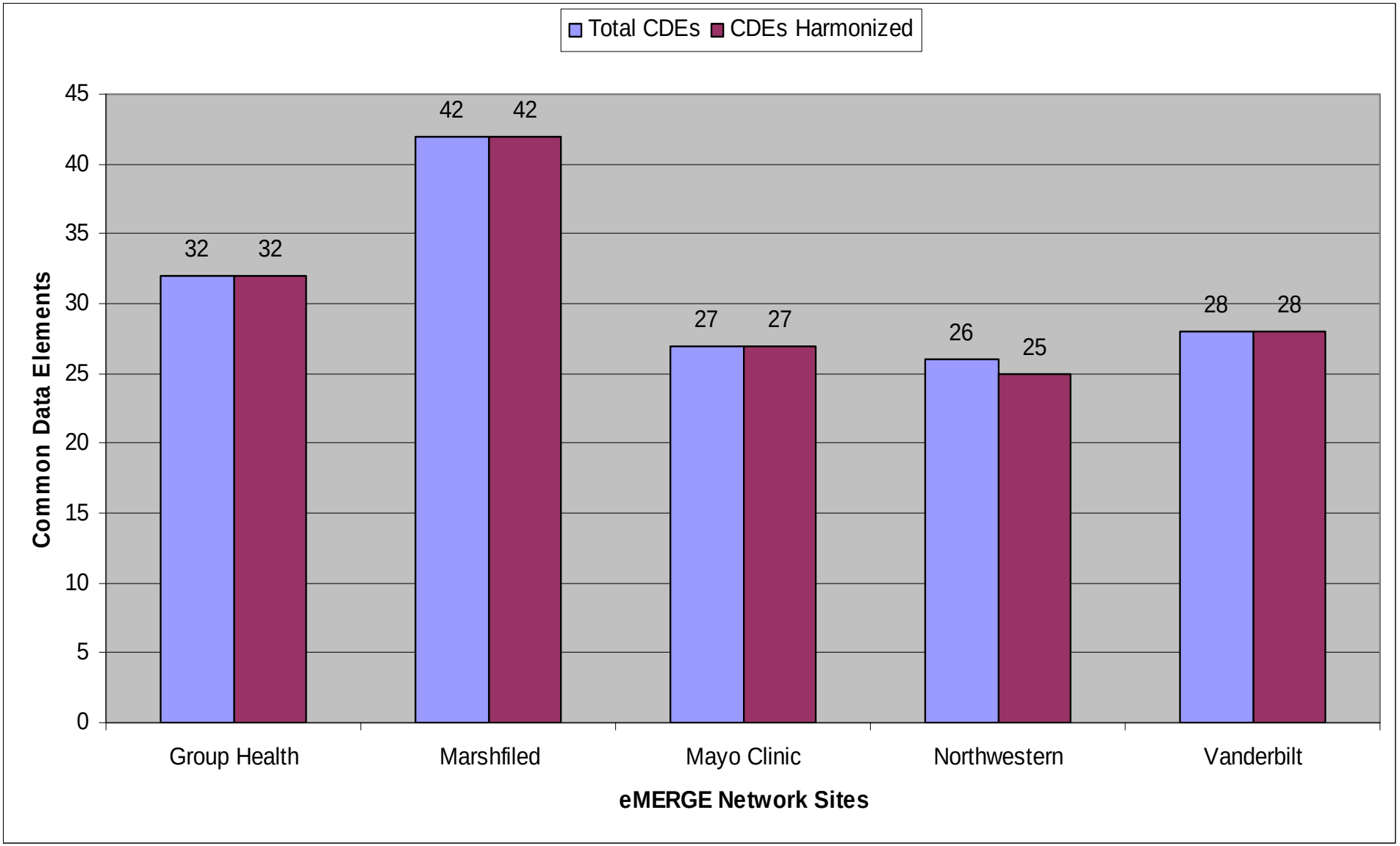
MAYO CLINIC

9 new NCI Thesaurus Concepts

- Ankle-Brachial Index (ABI) (C87304)
- Decade (C87556)
- Current Procedural Terminology (C87308)
- Cognitive Abilities Screening Instrument (C87307)
- Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (C86966)
- Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition (C86967)
- Fulfill (87531); Synonym “meets”
- NINCDS-ADRDA Criteria for Alzheimer's Disease (86983)
- Quartile (C87306)

MAYO CLINIC

Results: Revised Mapping (04/2010)



MAYO CLINIC

eleMAP Data Harmonization Tool

- Demo….http://www.gwas.net/eleMAP

MAYO CLINIC

Discussion: How to Proceed/Next Steps

- Is caDSR our “default” mapping?
  - Almost 100% mapping to caDSR CDEs
- Implication to dbGAP submissions
  - Individual sites going to “re-map” their phenotype data based on harmonized data dictionaries?
- Feedback on eleMAP implementation
  - Thanks to Luke from Marshfield!
- Publications
  - AMIA poster on eleMAP submitted
  - JAMIA manuscripts: (1) harmonization [draft] (2) SNOMED post-coordination [under preparation]
- Collaboration with PhenX

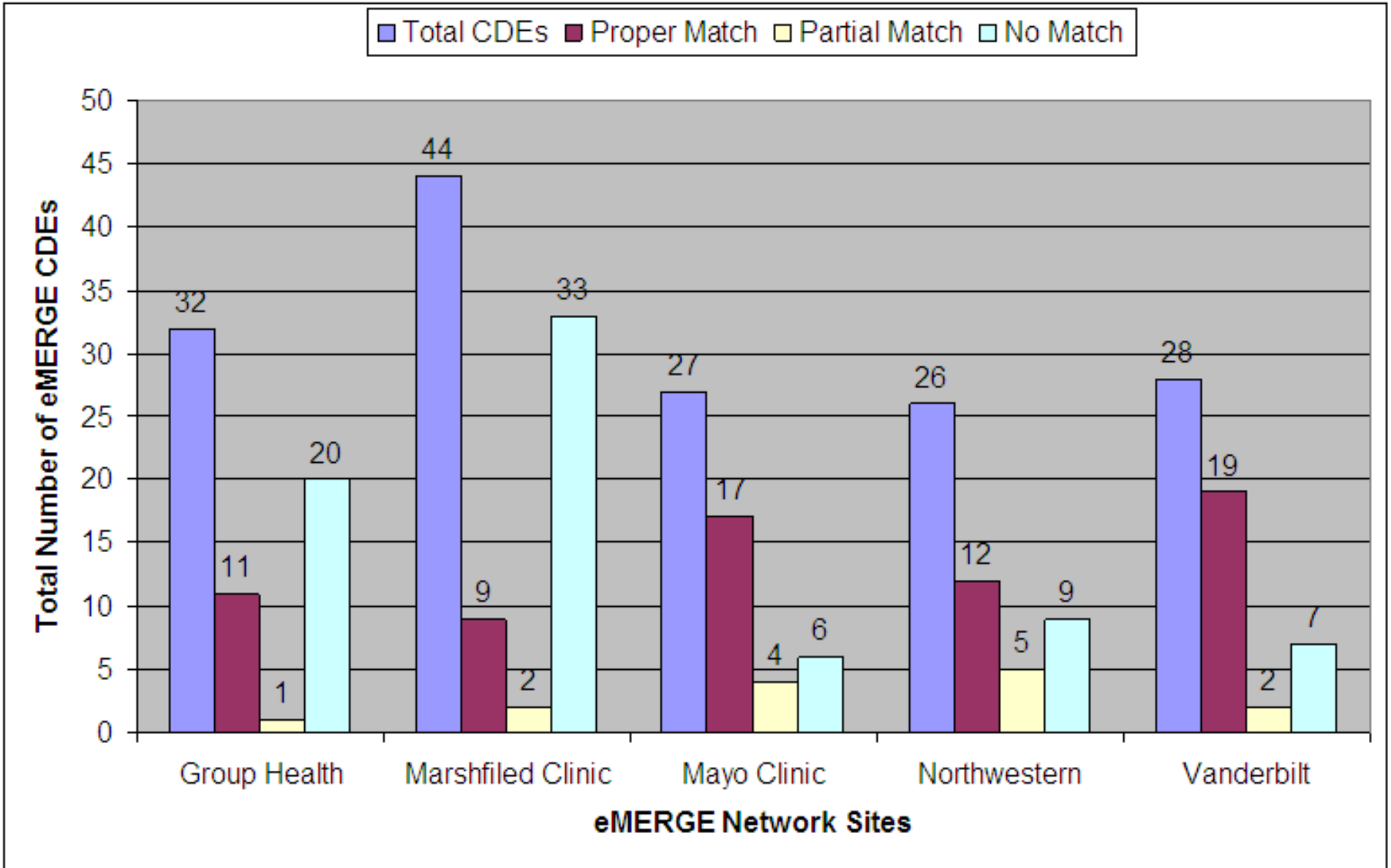
MAYO CLINIC

Thank You!

Q & A

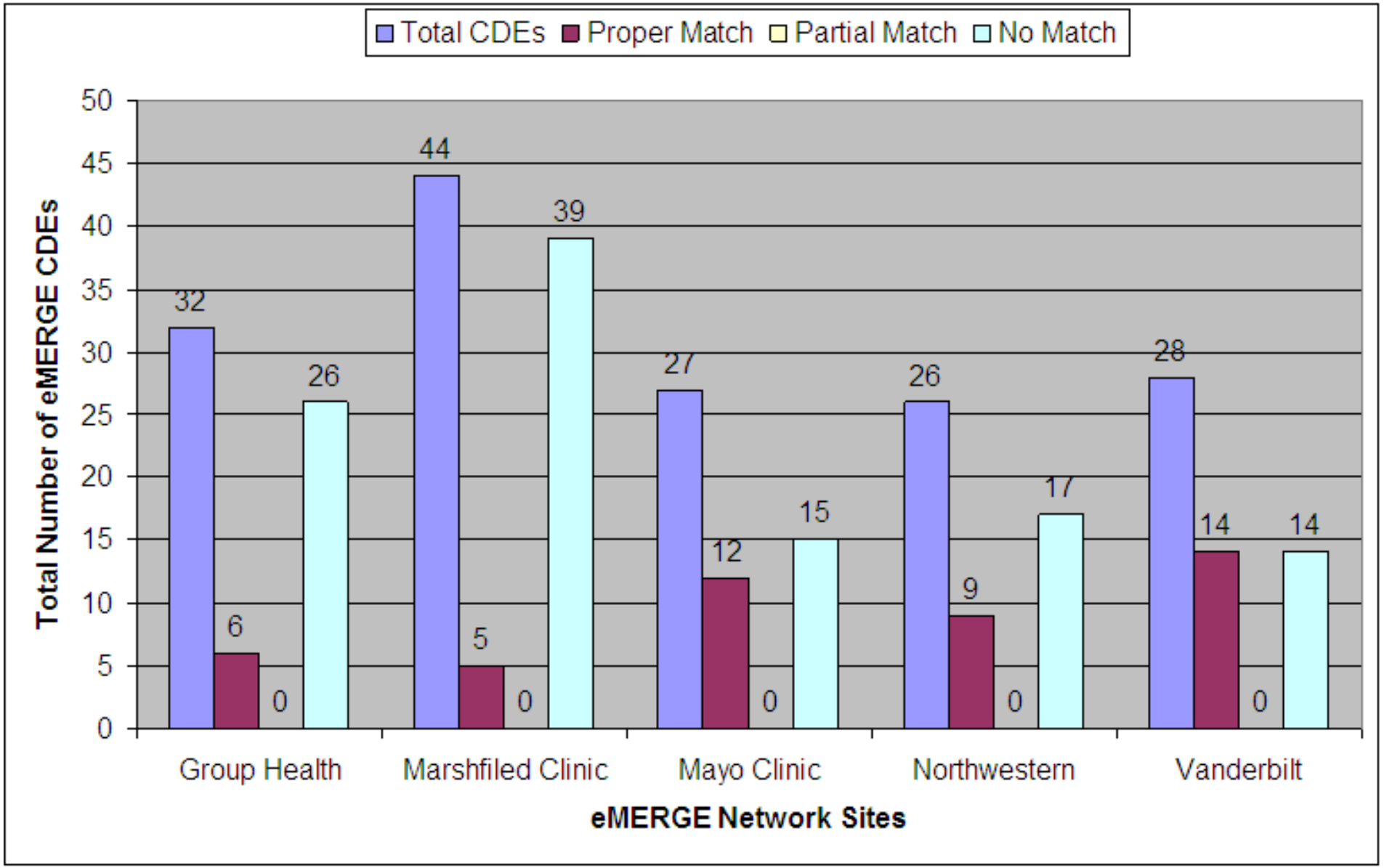
MAYO CLINIC

Results: NCI Thesaurus



MAYO CLINIC

Results: SDTM Terminology



MAYO CLINIC

Results: SNOMED (pre-coordination)