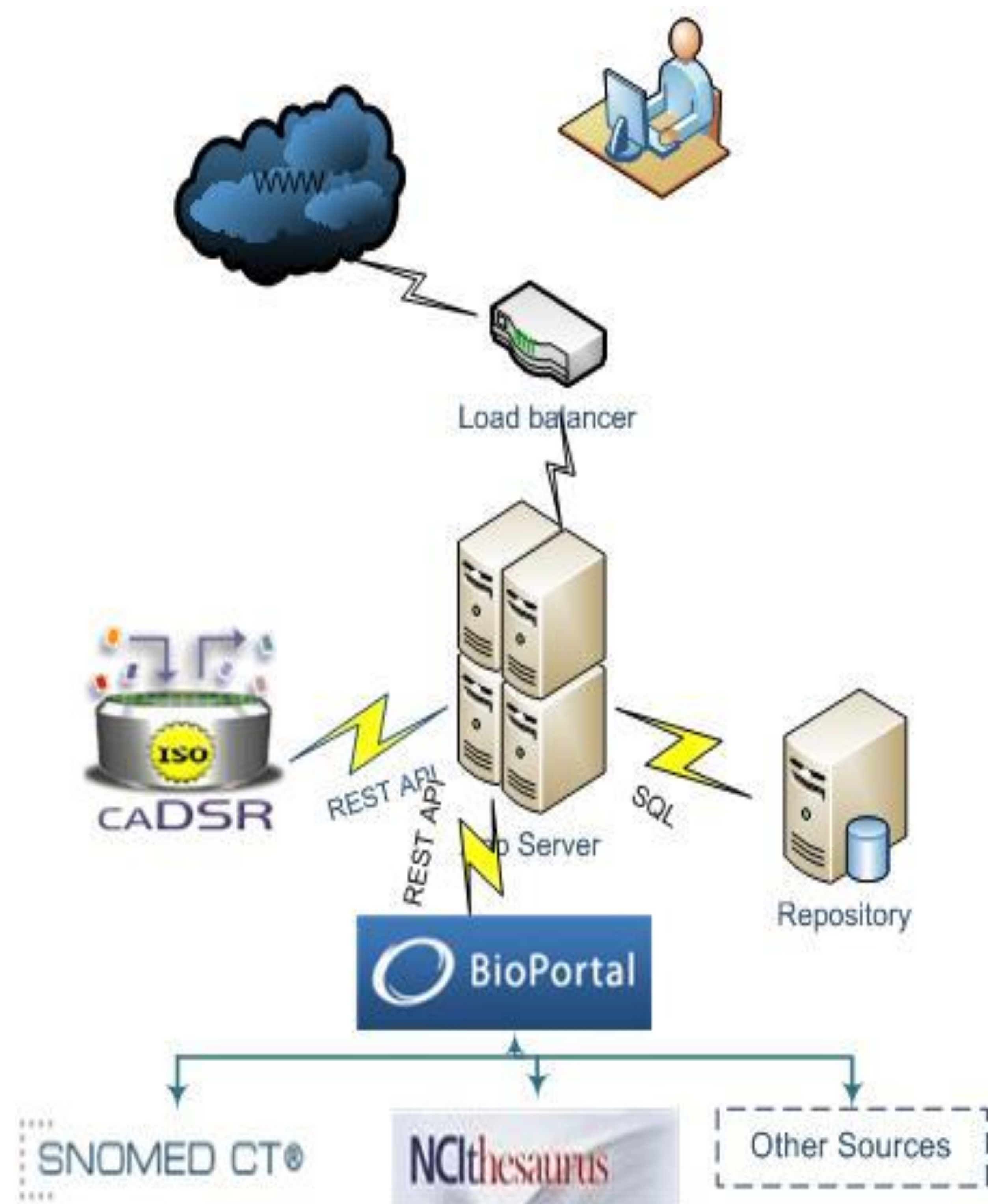


eleMAP: An Online Tool for Harmonizing Data Elements using Standardized Metadata Registries and Biomedical Vocabularies

Jyotishman Pathak, PhD¹ Janey Wang, MS² Sudha Kashyap² Rongling Li, MD, PhD³ Daniel R. Masys, MD² Christopher G. Chute, MD, DrPH¹
¹Mayo Clinic, Rochester, MN; ²Vanderbilt University, Nashville, TN; ³National Human Genome Research Institute, Bethesda, MD

The eMERGE Network
 electronic Medical Records & Genomics



Abstract. A key aspect in enabling high-throughput phenotyping studies requires standardized representation of the phenotype data using Common Data Elements (CDEs) and controlled biomedical vocabularies. In this abstract, we introduce eleMAP—an online tool that allows researchers to harmonize their local phenotype data dictionaries to existing metadata and terminology standards such as the caDSR (Cancer Data Standards Registry and Repository) and SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms).

Introduction and Background. With recent advances in genotyping technologies, to increase our ability to fully understand the genetic basis of common diseases, the NIH in 2007 funded a multi-site consortia called eMERGE (Electronic Medical Records and Genomics; <http://www.gwas.net>) for high-throughput phenotyping. In particular, the crux of eMERGE is the development of tools and algorithms for extracting phenotypic data, representing actual healthcare events, from the EMR systems at each institution in a consistent and comparable fashion. However, due to lack of common EMR systems or standardization of EMR data across the institutions, one of the goals of eMERGE is develop tools and methods to facilitate harmonization of phenotype data dictionaries and CDEs to terminological and metadata healthcare standards for interoperable representation of phenotype data. To address this requirement, we developed eleMAP—an online tool that allows researchers to harmonize their local phenotype data dictionaries to existing metadata and terminology standards such as the caDSR (Cancer Data Standards Registry and Repository [1]) and SNOMED-CT (Systematized Nomenclature of Medicine [2]).

Methods and Results. Our approach to mapping CDEs to pre-coordinated terms and concepts from standardized biomedical terminologies and metadata resources is as conservative as possible. We first try to find an exact string match for the CDE variable provided in the data dictionaries of several eMERGE studies (e.g., cataract, type 2 diabetes). If no match is found, we do an approximate search by normalizing the original search string (e.g., eliminating underscores, hyphen variations) as well as adding a wildcard (*) to the beginning and end of the string. The entire process is automated, and the search stops as soon as a match is found. Furthermore, if CDE has an enumerated list of permissible values (in the data dictionary), we repeat the above procedure to find corresponding terms for the CDE value set contents. We developed an online tool called eleMAP for mapping CDEs from several eMERGE studies to the caDSR and various biomedical vocabularies in the NCBO. For querying the caDSR, we use the caDSR HTTP API which provides BioPortal [3]. Various forms of functions for querying the CDEs. For querying biomedical vocabularies, we used RESTful Web services from BioPortal. While BioPortal contains approximately 200 biomedical terminologies and ontologies, our searches were restricted to SNOMED-CT and NCI Thesaurus. More information about eleMAP is available at: <http://www.gwas.net/eleMAP>.

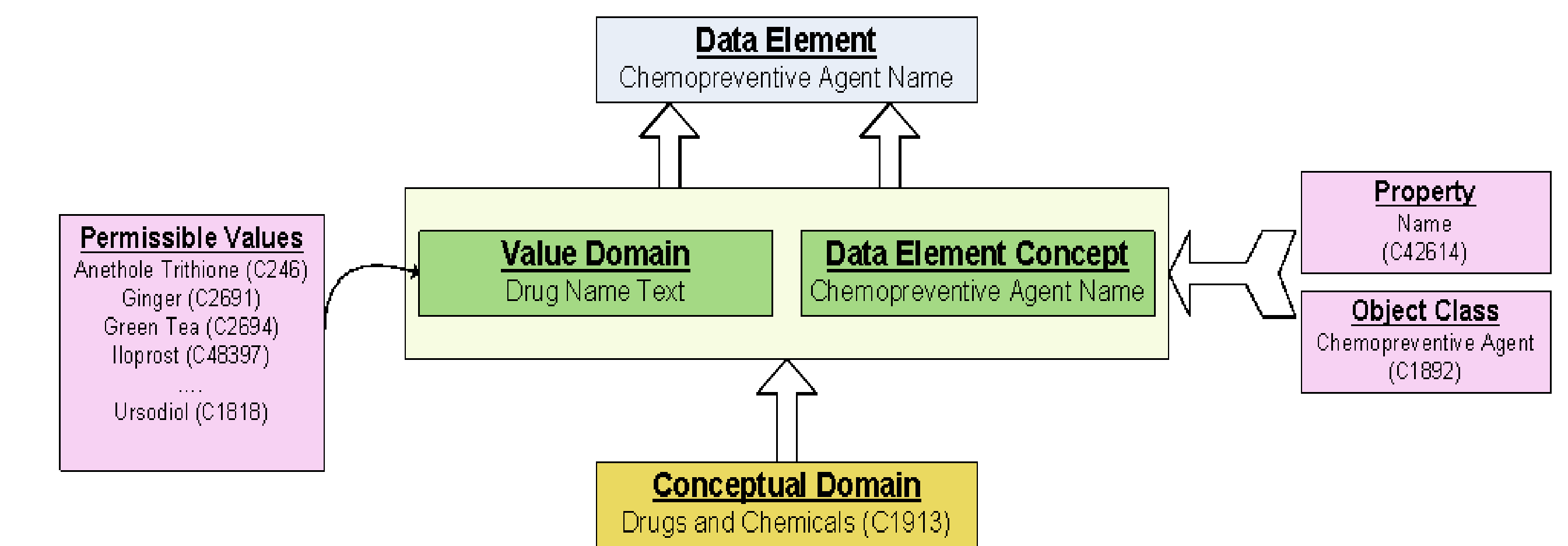


Figure 1: caDSR and ISO/IEC 11179 Model for Metadata Registries

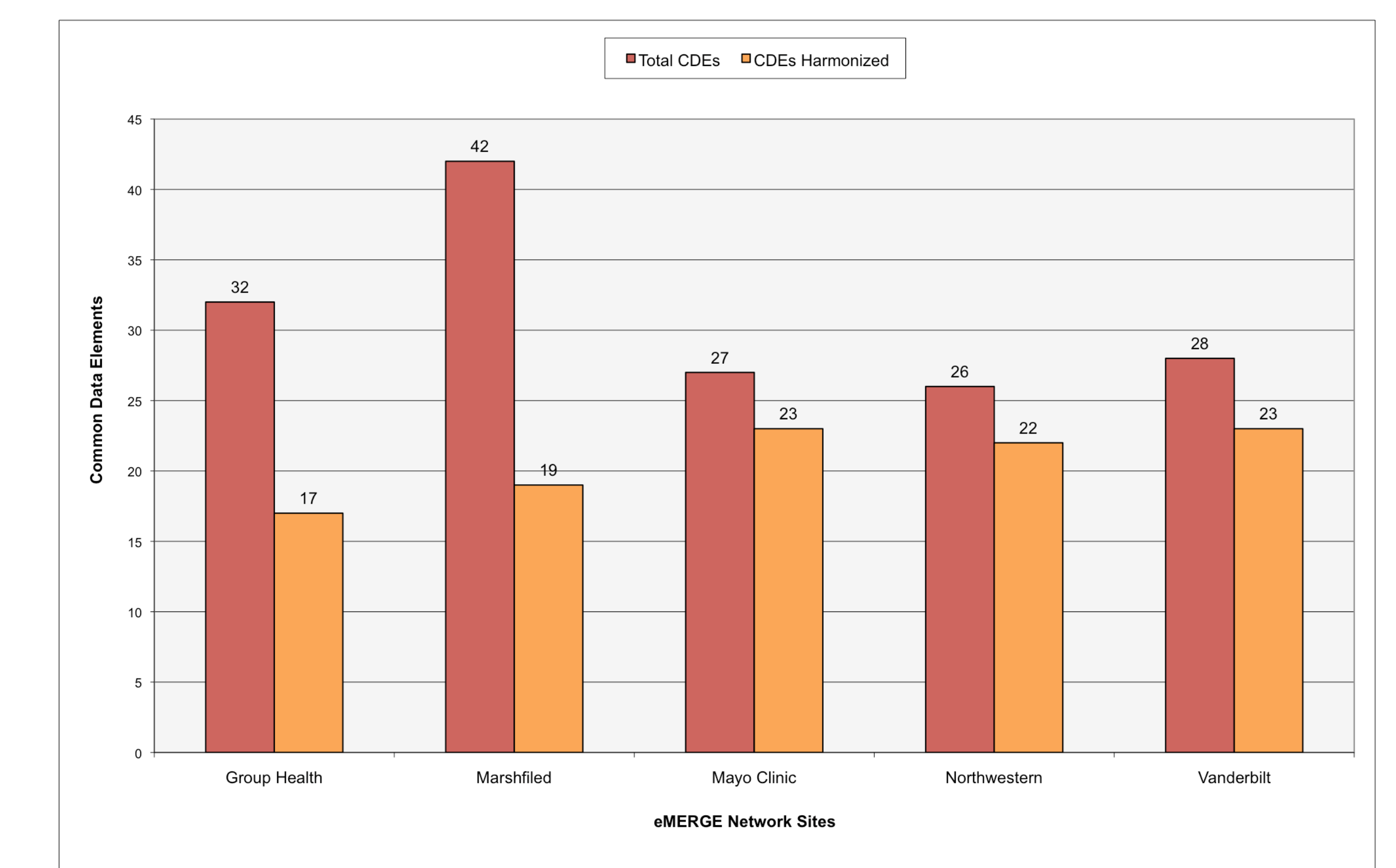


Figure 2: Preliminary results from eMERGE network CDE harmonization using eleMAP

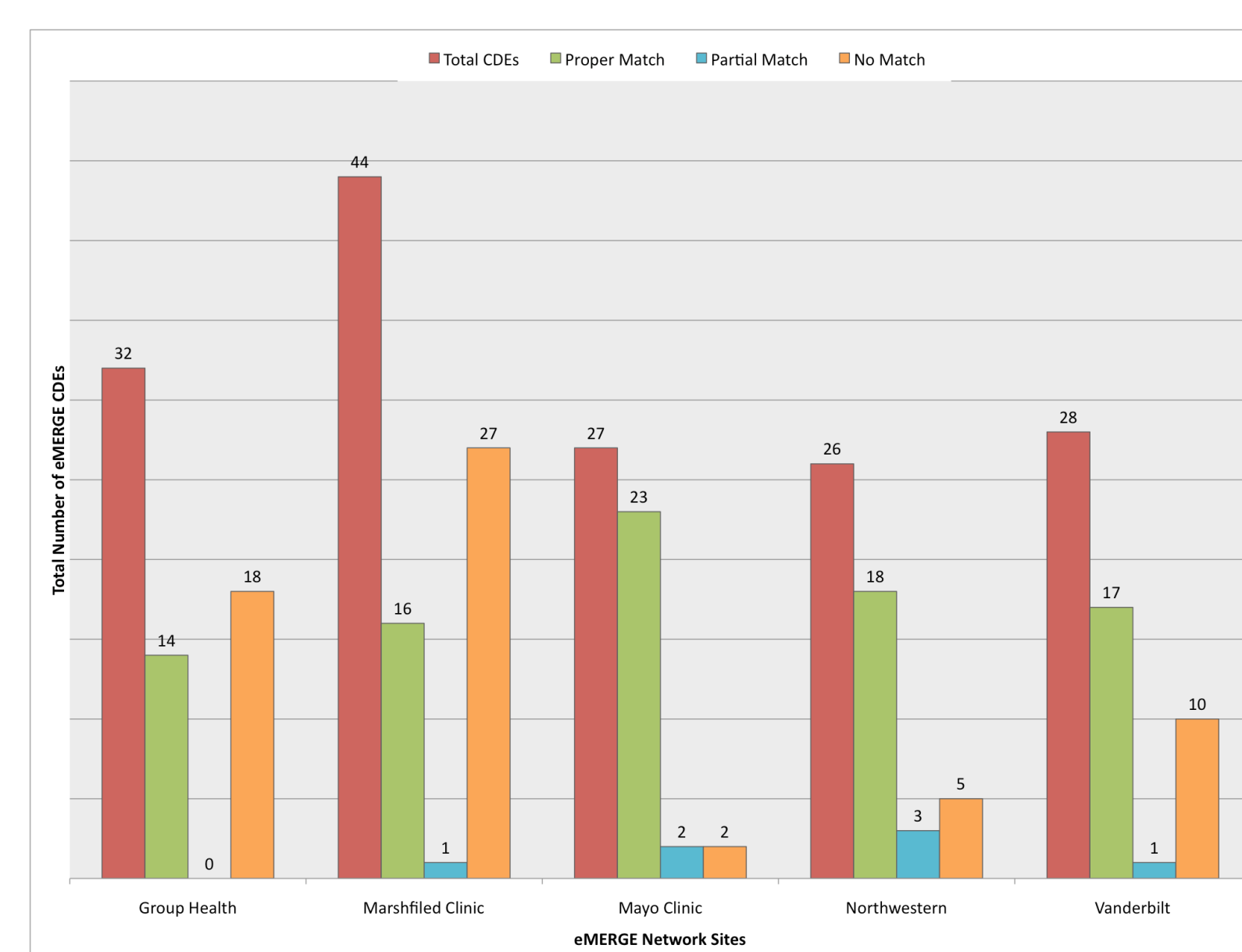


Figure 3a: eMERGE data elements mapped to caDSR (November 19th, 2009 caDSR release)

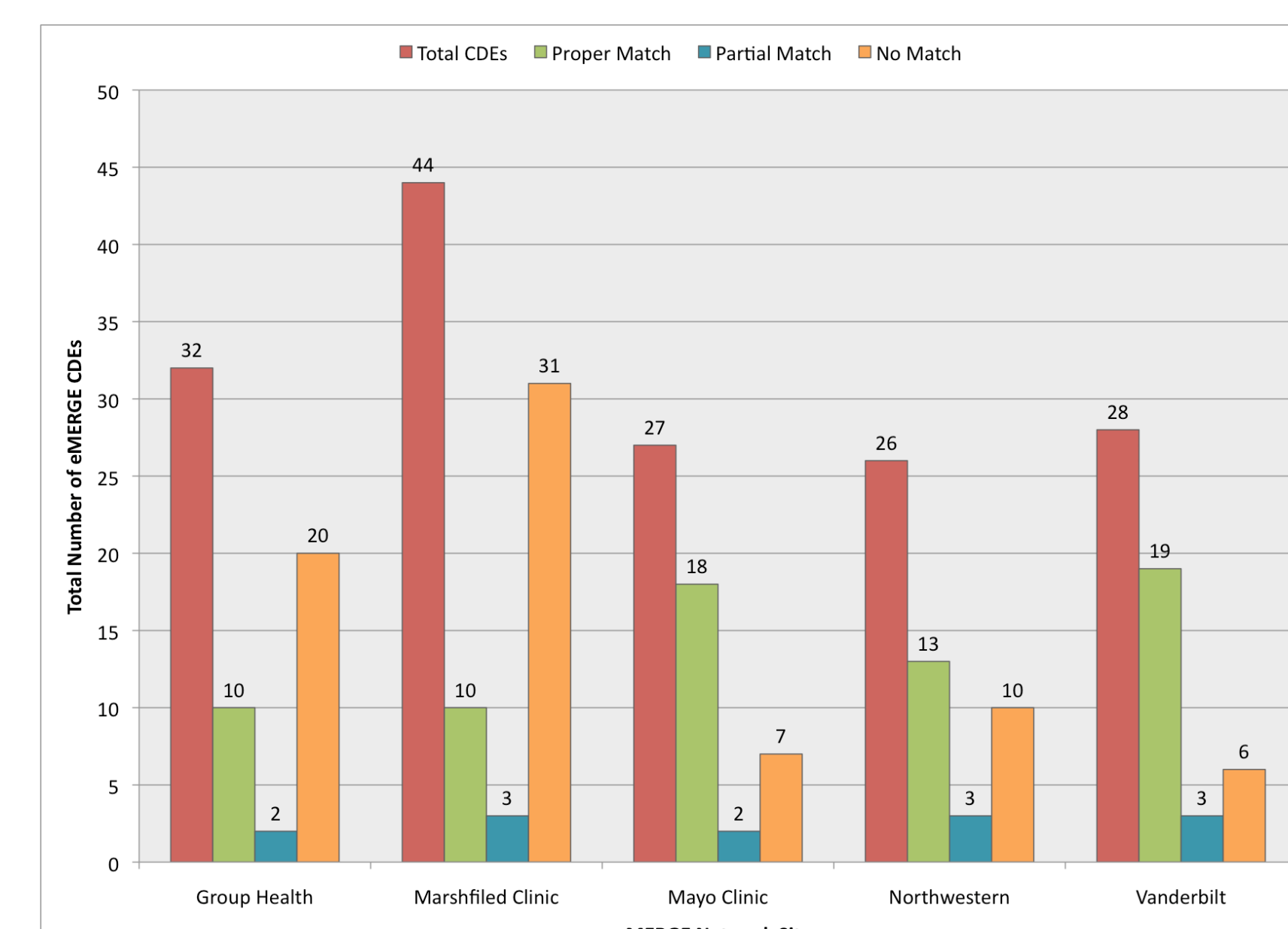


Figure 4a: eMERGE data elements mapped to pre-coordinated SNOMED-CT concepts (January 31st; 2010 SNOMED International Release)

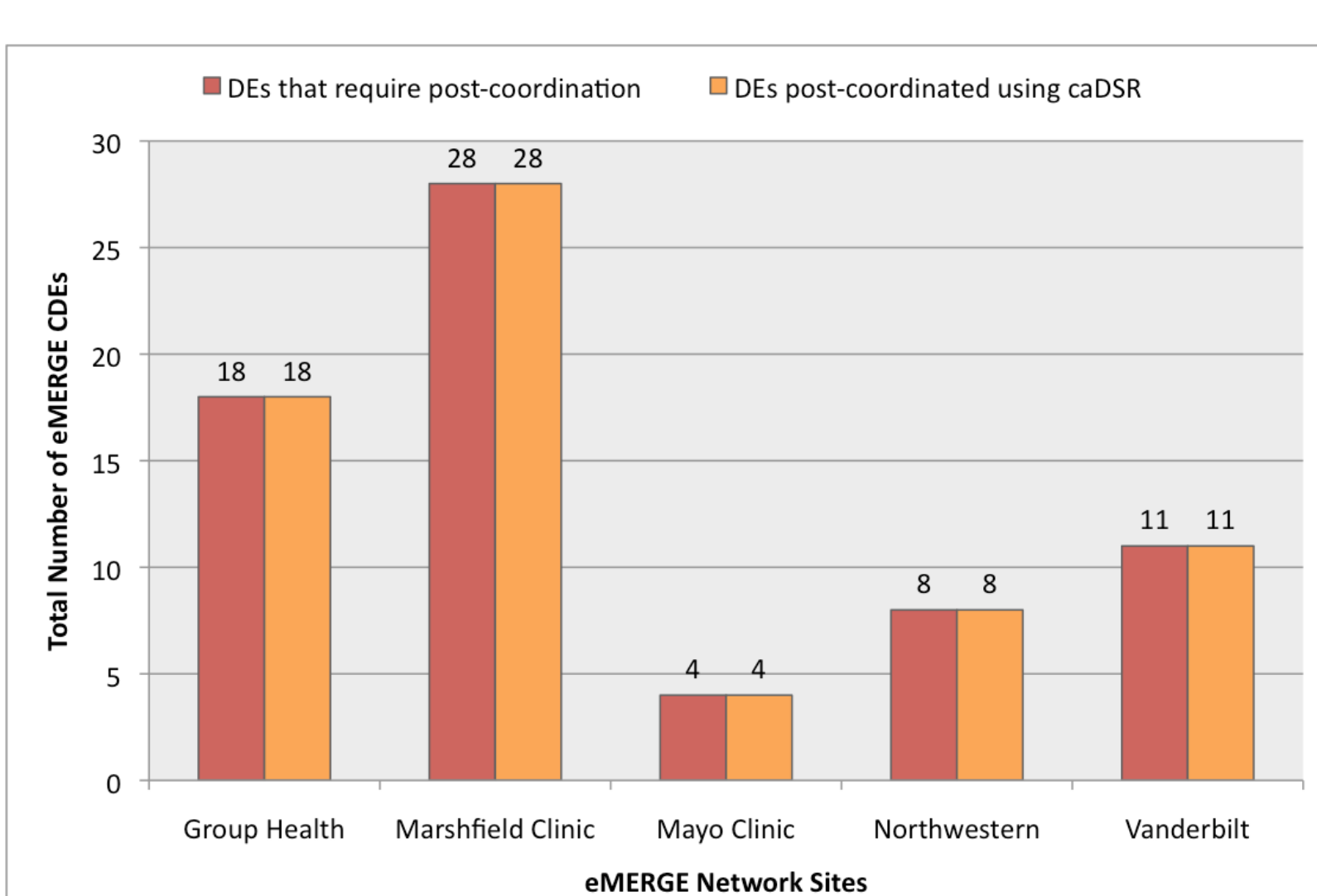


Figure 3b: eMERGE data elements mapped to caDSR (April 20th, 2010 caDSR release)

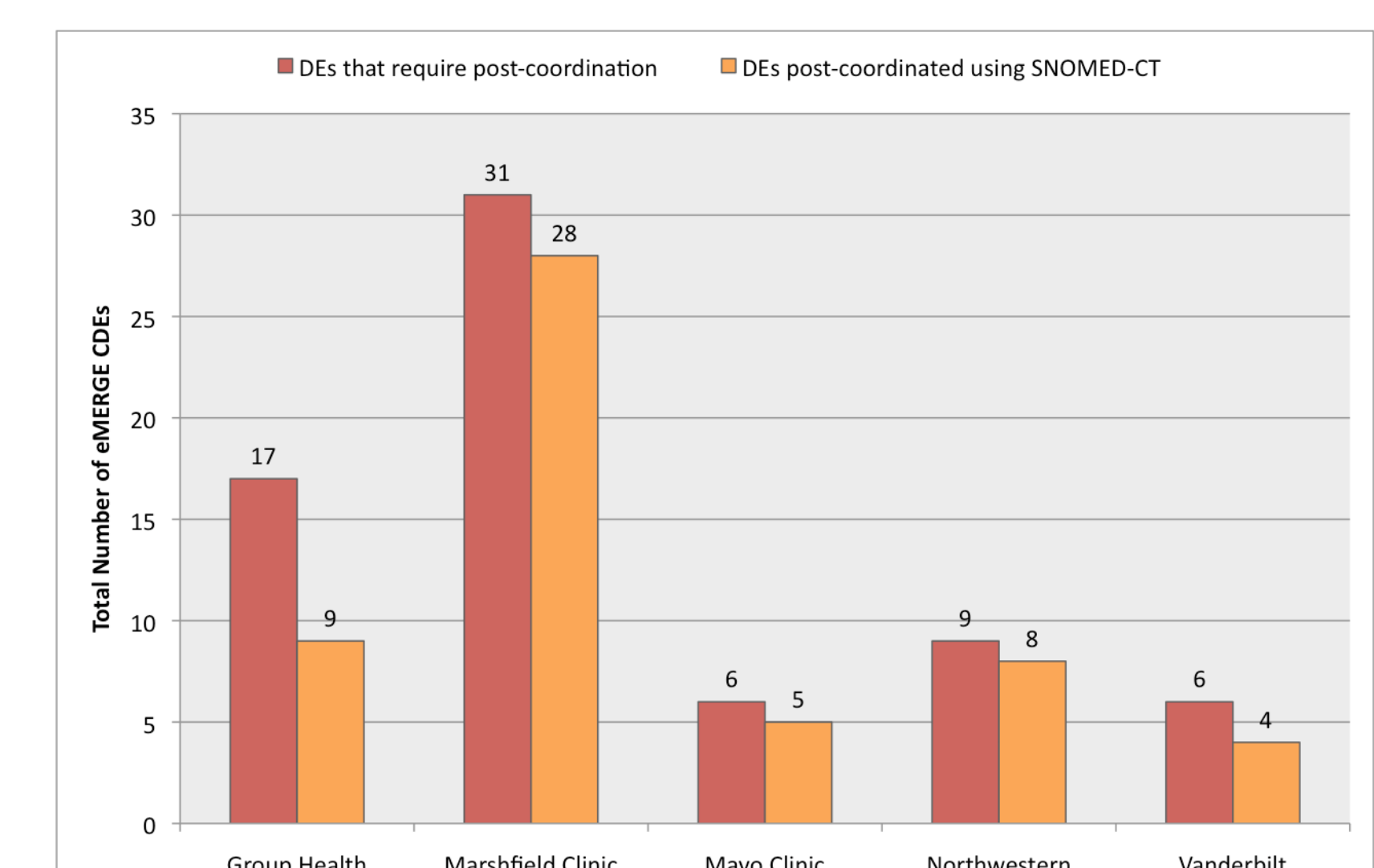
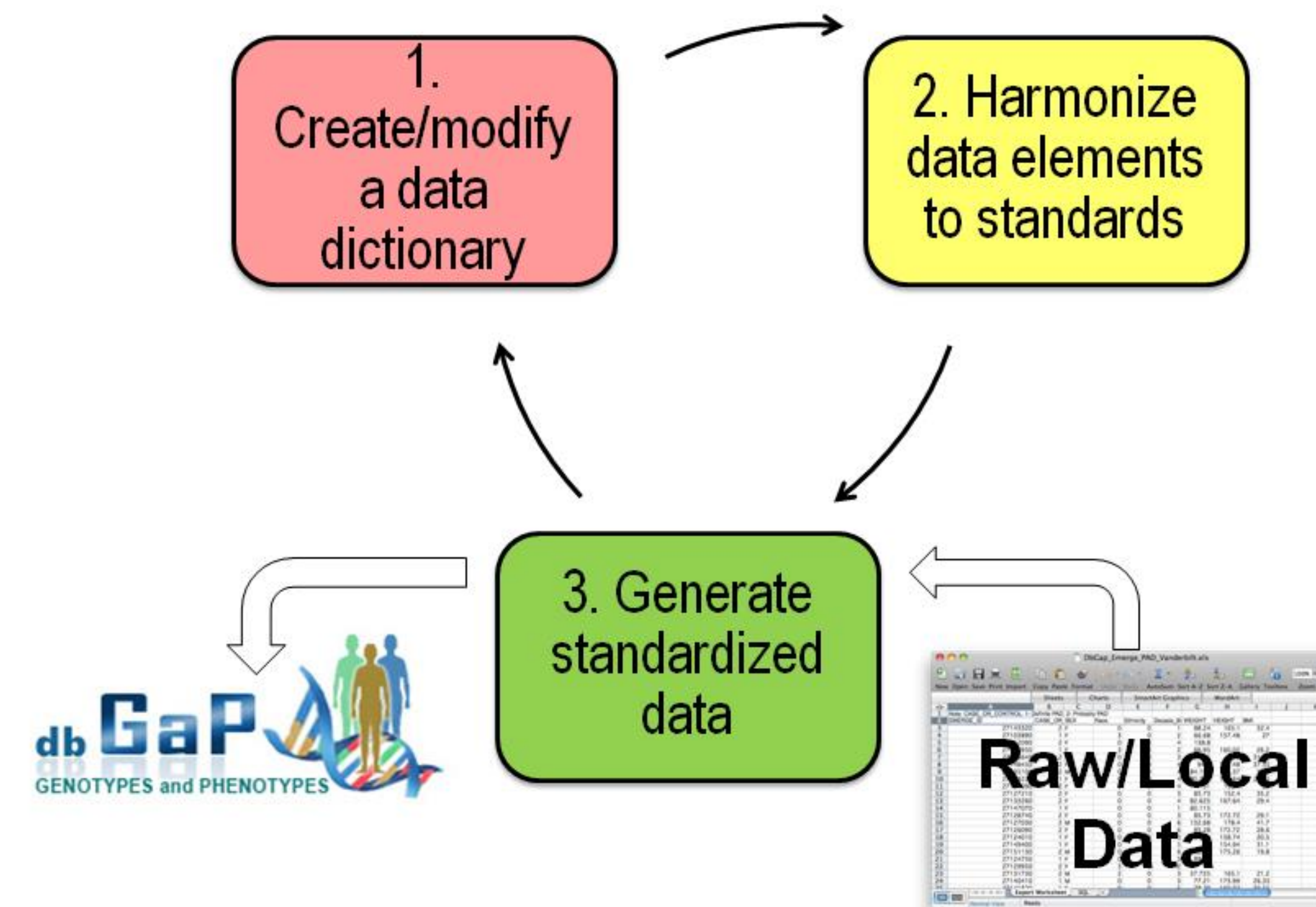


Figure 4b: eMERGE data elements mapped to post-coordinated SNOMED-CT concepts (January 31st; 2010 SNOMED International Release)

eleMAP Data Harmonization Workflow



References

1. caDSR: Cancer Data Standards Registry and Repository. Last accessed: March 6th, 2010.
2. SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms. Last accessed: March 6th, 2010.
3. Noy, N., et al., BioPortal: ontologies and integrated data resources at the click of a mouse.

Future Direction and Next Steps

- Incorporation of additional terminology and metadata resources such as RxNorm and LOINC
- Collaboration with PhenX and incorporation of standardized measures for use in GWAS
- Improved visualization and search capabilities
- Investigate using Clinical Element Models, and ISO 13606 Archetypes
- Release the eleMAP software code, package and database to the open-source community
- Application in projects beyond eMERGE

