

Using SNOMED-CT For Translational Genomics Data Integration

Joel Dudley¹⁻³, David P. Chen¹⁻³, Atul J. Butte¹⁻³, M.D., Ph.D.,
¹Stanford Center for Biomedical Informatics Research, Department of Medicine,
²Department of Pediatrics, Stanford University School of Medicine, Stanford, CA/USA
³Lucile Packard Children's Hospital, Palo Alto, CA/USA
{jdudley, dpchen, abutte}@stanford.edu

As industrial, governmental, and academic agencies place increasing emphasis on translational research, biomedical researchers are now faced with entirely new challenges in regards to both biomedical data integration and knowledge discovery. There is now both a strong need and a tremendous opportunity to apply translational bioinformatics to address the fundamental challenges in integrating the vast bodies of -omics and clinical data. Here we report on our preliminary work in utilizing SNOMED-CT as both a tool for translational data discovery, and a major component in a framework for the large-scale integration of gene expression microarray data and clinical laboratory data. Annotations from microarray experiments in NCBI GEO were mapped to SNOMED-CT terms using UMLS, and these mappings were joined to clinical laboratory data using ICD9CM to SNOMED-CT mappings within UMLS. We find that microarray experiments characterizing 211 distinct diseases can be mapped to clinical laboratory data measurements for 13,452 distinct patients. We maintain that this work represents critical first steps in providing a foundation for large-scale translational data integration, and underlines the important role that controlled clinical terminologies, such as SNOMED-CT, can play in addressing such problems.

INTRODUCTION

Our ability to generate high-quality biomolecular data has advanced at considerably faster rate than our ability to investigate the data generated. This imbalance, driven primarily by rapid advances in high-throughput biological data acquisition technologies and plummeting per-experiment costs, has created an entire spectrum of informatics challenges that are, in many instances, as intangible and complex as the fundamental biological questions that these technologies were designed to address. As a consequence, our ability to formulate and investigate important biological and medical questions is currently limited by our ability to manage and integrate the profusion of biomedical data.

Problems in data integration are moving towards the forefront of biomedical research, driven foremost by the sheer diversity of measurement technologies now

available, and the tremendous volumes of such measurements finding their way into the public domain. The situation is further complicated by the fact that the majority of the public biomolecular data is annotated using unstructured free-text, making it difficult to discern the various biological and medical contexts of the data in an automated fashion. In previous work we demonstrated the feasibility of using controlled terminologies and straightforward text-mining techniques to elucidate clinical, environmental, and phenotypic contexts from free-text annotations associated with public microarray data^{1,2}. The establishment of experimental context is critical to linking genes to environment, phenotype, and ultimately medicine.

While most major types of biomolecular data can be found in the public domain, it is traditionally difficult for researchers to gain access to clinical data. This is unfortunate as the data generated on a daily basis by hospitals and clinicians is perhaps the richest source of phenotypic biomarker data currently available. Fortunately modern Electronic Health Record (EHR) systems such as the Stanford Translational Research Integrated Database Environment (STRIDE)³ and the University of Virginia Health System Clinical Data Repository (CDR)⁴ grant institutional researchers access to large volumes of de-identified, quantitative clinical data in digital form. In recent work, we demonstrated the utility in applying bioinformatics methods to quantitative clinical data to draw new inferences about disease severity⁵, and elucidate novel biomarkers⁶.

Genome Wide Association studies have revealed that for many complex diseases, the pathogenesis of the disease may be facilitated by relatively minor changes across a large number of genes interacting through as of yet poorly understood mechanisms⁷. These findings have therefore highlighted the importance of linking biomolecular data with phenotypic quantifications in order to uncover the full complexity of disease etiology. Recent work in integrating these two data types has offered new insights into disease etiology and pathology with direct clinical implications. Segal and colleagues correlated imaging traits from computed tomography (CT) images of liver cancers with gene expression

data to reconstruct global expression signatures in cancer tumors that are linked to diagnosis, prognosis and treatment⁸. A number of studies have demonstrated the utility of patient microarrays in identifying gene expression patterns linked to disease diagnosis⁹, subtypes^{10, 11}, outcome¹², and treatment^{13, 14}. As significant as the aforementioned findings are, their underlying methods are limited by the fact that, in all instances, they require that the biomolecular and clinical data be derived from the same patient. Given the current high costs and logistical complexities involved in acquiring patient data in a clinical setting, it would be prohibitively expensive to scale the same approaches to address the broad spectrum of human disease. Furthermore, such an approach implicitly eschews the great wealth of public biomolecular data readily available.

A major problem in integrating clinical and biomolecular data derived from disparate sources is to identify attributes by which they can be appropriately joined. This task is complicated by the fact that the majority of biomolecular data is annotated around the concepts of genes and gene products, whereas clinical data is centered on the concept of a patient. We find one concept shared among both clinical data and vast amounts of biomolecular data, and that is the concept of a *disease*. Therefore it is possible to integrate anonymous biomolecular data characterizing an aspect of a particular disease state with quantitative clinical data derived from patients being treated for the same disease.

Central to this approach is the need for a comprehensive controlled disease terminology through which the biomedical and clinical data is joined in a systematic fashion. In general, we would want this disease terminology to maximize three primary criteria: coverage, defined by the number of unique disease terms defined; expressiveness, which is the richness of relationships between disease terms; and resolution, which is the level of detail offered by the terminology structure. A deficiency in any of these could negatively impact the amount and diversity of data that could be integrated, and potentially limit the types of analyses that can be performed on the data downstream. There are a number of well-established disease terminologies in active use that satisfy the above criteria to varying degrees. Chief among these are the International Classification of Diseases (ICD), Medical Subject Headings (MeSH), and the Systemized Nomenclature of Medicine-Clinical Term (SNOMED-CT). Each of these is suited for data integration, yet each of them present particular pros and cons.

The ICD terminology, evolved from a lineage that spans more than 100 years, is the most widely utilized disease terminology, with widespread adoption among a large number of major healthcare providers, the U.S. Federal Government, as well as the World Health Organization. Consequently, the majority of clinical data is codified using ICD codes. Unfortunately the ICD is poorly suited for data integration as the approximately 14,000 unique terms codified by ICD is quite small compared to other terminologies. Furthermore, the ICD is more a compendium of diagnosis and procedure codes, as it lacks any significant hierarchical or relational structure.

MeSH, which is used primarily for the purpose of indexing publications, is only slightly larger than ICD in terms of size with more than 22,000 unique terms. However, the design of MeSH is much more structured and diverse compared to ICD. MeSH terms are arranged into a hierarchy of 14 distinct top-level categories that organize terms by Anatomy, Disease, Chemicals and Drugs, and Geography among other things. MeSH also contains a set of qualifier terms that can be used to narrow the specificity of a descriptor term (e.g. "Measles/epidemiology"). While MeSH possesses many of the attributes desirable for translational data integration, its attributes modest in comparison to those of SNOMED-CT.

SNOMED-CT was born from a medical terminology lineage that traces back more than 75 years, and is currently in use by pathologists worldwide to perform precise classifications of human disease^{15, 16}. With more than 340,000 unique biomedical concepts organized into 19 relational hierarchies linked by more than 1.3 million relationships, it is by far the most expansive and expressive disease terminology in existence. The sheer number of concepts coupled with the rich relational architecture in SNOMED-CT offers attributes superior to other disease terminologies. For example, SNOMED-CT establishes that a *clear cell carcinoma of the kidney* is both a *malignant tumor of the kidney* and a *malignant tumor of the retroperitoneum*. The ICD version 9 (ICD-9) simply asserts that a *malignant neoplasm of the kidney* is a *malignant neoplasm of the genitourinary organs*, which is a much coarser designation. Therefore assert that SNOMED-CT is currently the best-suited terminology for integrating biomolecular and clinical data by disease.

In this study we investigate the feasibility of using SNOMED-CT to integrate gene expression data from a public microarray repository with de-identified clinical laboratory data obtained from a hospital EHR system by disease. We propose that SNOMED-CT is

well suited for this approach as it is the largest disease vocabulary currently available. We evaluate the effectiveness of this approach based on the extent of data successfully joined.

METHODS

A high level representation of the data integration approach is detailed in figure 1. The microarray experiment data was obtained from the NCBI GEO FTP site (downloaded 11/27/2007), which was parsed into a relational structure and stored in a MySQL database. The de-identified clinical laboratory data was obtained from the Lucile Packard Children’s hospital via STRIDE as delimited text files. UMLS release 2007 AA was used as the vocabulary source. The integration steps were performed as follows.

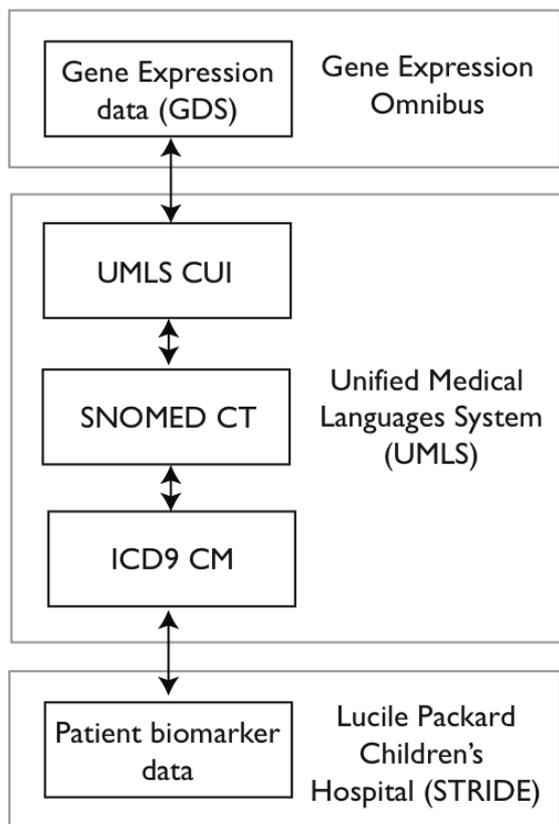


Figure 1 – Schematic representation of the approached used to join gene expression data with clinical laboratory data. Annotations from GDS are first mapped to UMLS CUIs that map to at least one SNOMED CT term, and the ICD9 CM codes from the patient records are mapped to SNOMED CT terms using the relational architecture of UMLS.

Mapping microarray experiments to diseases

Clinically relevant microarray data was identified using a previously described method¹⁷. In brief, we queried the NCBI Gene Expression Omnibus (GEO)¹⁸ to obtain all GEO DataSet experiments with associated PubMed identifiers. For each PubMed identifier we obtained the associated MeSH headings using NCBI eUtils. Each of the MeSH headings was mapped to a UMLS CUI using the MRCONSO table. Using the MRSTY table, we obtained the semantic type identifier (TUI) for the mapped CUIs, and if any MeSH term is found to have a semantic type among Injury or Poisoning (T037), Pathologic Function (T046), Disease or Syndrome (T047), Mental or Behavioral Dysfunction (T048), Experimental Model of Disease (T050), or Neoplastic Process (T191) then the associated experiment is determined to be disease-associated and therefore clinically relevant. This resulted in the positive identification of 737 disease-associated experiments.

The disease-associated experiments are investigated by a second previously described text-mining technique that examines GEO DataSet (GDS) subset annotations to identify when a disease state is being compared to a normal control state². GDS are higher-level representations of microarray experiment in which samples are organized into biologically informative collections known as subsets. The subsets are representative of the experimental axis under examination (figure 2). An attempt is made to map the free-text annotations associated with the GDS subsets to SNOMED-CT disease terms using UMLS concepts. These mappings are subsequently manually reviewed for accuracy, where erroneous codifications are corrected if found.

4 assigned subsets			
Samples		Type	Description
<input checked="" type="checkbox"/> (6)	<input checked="" type="checkbox"/>	disease state	type 2 diabetes
<input checked="" type="checkbox"/> (6)		disease state	non-diabetic
<input checked="" type="checkbox"/> (6)	<input checked="" type="checkbox"/>	age	8 week
<input checked="" type="checkbox"/> (6)		age	16 week

Figure 2 – Example of microarray data subsets defined by GEO GDS experiments.

Mapping patient laboratory data to diseases

Clinical laboratory data for pediatric patients from the Lucile Packard Children’s Hospital was obtained digitally from the STRIDE system. All of the laboratory measurements were received pre-encoded with ICD-9 codes. These ICD-9 codes were mapped to SNOMED-CT codes by first querying UMLS to

find the CUI identifier associated with the ICD-9 code. We then took advantage of the inter-terminology mappings provided by the UMLS (MRMAP) table to translate the ICD-9 codes into SNOMED-CT concepts using associated CUIs.

Joining the microarray and patient lab data by disease

The GDS subsets with mappings to SNOMED-CT disease CUIs were joined with the clinical laboratory data using the UMLS CUIs derived from mapping the ICD-9 codes to SNOMED-CT terms using the UMLS MRMAP table. Of the 238 unique disease concepts mapped to the microarray data, 90% were mapped to quantitative clinical laboratory data for at least one patient.

RESULTS

Using automated methods, were able to identify 737 GDS microarray experiments in NCBI GEO related to human disease. The GDS subsets were investigated for terms related to UMLS concepts that were linked to a SNOMED-CT disease term, resulting in the identification of 238 unique human disease concepts. In total, 29,451 microarray samples were codified with SNOMED-CT disease identifiers. Note however that method was restricted to include only those GDS for which a disease and normal control subset could be identified. This restriction ensures that a disease vs. normal vector of change can be extracted from the data to establish a baseline disease expression signature for downstream analysis.

Disease	SNOMED Terms	ICD9CM Terms	Ind
Allergic asthma	1	1	2240
Asthma	1	1	2240
Allergic asthma NEC	1	1	2240
Esophageal Reflux	1	1	1895
H. pylori infection	1	2	1322
Colitis	1	1	1299
Primary Hypertension	1	1	1017
Hypertension	1	1	1017
Obesity	2	1	1010
Type 1 diabetes	1	1	843

Table 1 – Top ten data mappings ordered by the number of patient lab records matched.

We retrieved quantitative clinical laboratory data representing diagnostic biomarkers for 49,414 patients across 9,997 distinct diagnosis codes. These codes mapped to 20,049 distinct UMLS CUIs. It is interesting to note that in mapping ICD to UMLS we find that twice as many UMLS concepts as ICD-9 terms are found. This likely resulted from the fact that ICD-9 is generally a more high-level terminology, and therefore terms related to rare genetic disorders, for example, may only be represented by one ICD-9 code, whereas UMLS may allow for more fine-grained attribution of specific rare genetic disorders.

In joining the ICD-9 disease codes from the clinical laboratory data to the microarray data using SNOMED-CT disease codes, we find that 211 of the unique disease concepts annotating the microarray data can be mapped to clinical laboratory data. In total, clinical laboratory data for 13,452 patients was mapped to SNOMED-CT disease codes that were used to annotate the microarray GDS experiments. Table 1 shows the top diseases by the number of patients mapped.

Disease	SNOMED Terms	ICD9CM Terms	Ind
Follicular lymphoma	4	3	136
Hamman-Rich syndrome	4	2	18
Mycobacterial infection	3	2	26
Mixed hyperlipidemia	3	2	90
Hepatoma	3	2	67
Fetal alcohol syndrome	3	1	10
Diabetic nephropathy	3	2	30
Megakaryocytic leukemia	2	2	125
Acute monocytic leukemia	2	1	7
Status epilepticus	2	1	84

Table 2 – Top ten data mappings sorted by the number of SNOMED-CT terms matched.

As evident from the data listed in table 1, there are cases in which distinct SNOMED-CT terms will map to the same ICD-9 term. To explore the ambiguities of mapping terms between the SNOMED-CT and ICD-9 using CUIs, we investigated the overall pattern of the mapping cardinalities. Table 2 shows cases in which a single UMLS CUI maps to multiple

SNOMED-CT terms. This could indicate that there is some degree of ambiguity in the SNOMED-CT to ICD-9 UMLS mappings, and perhaps a dampening of SNOMED-CT term resolution when using UMLS concepts.

To better understand the influence of UMLS CUI definitions with regards to source identifier consolidation, we calculated summary statistics for several terminologies with UMLS and restricted the results to CUIs representing a disease. The summary statistics are listed in table 3.

Source	Total disease concepts	Identifiers per concept
SNOMED-CT	74,611	1.4
ICD-9-CM	12,631	1.1
NCI	12,257	1.0
MeSH	6,613	1.0

Table 3 – Summary statistics for select disease terminologies sorted by total number of disease concepts (CUI).

DISCUSSION

The profusion of large public data repositories of genome-scale measures, coupled with the pressing imperative to translate such data into medicine, has precipitated the need to develop informatics tools and techniques for integrating disparate forms of biomolecular and clinical data. The purpose of this investigation was to explore the feasibility of using SNOMED-CT for such integrative efforts. We assessed the feasibility of SNOMED-CT as a translational joining factor by using it to integrate anonymous gene expression data from a public microarray repository with de-identified clinical laboratory data by disease.

We find that SNOMED-CT is effective as a disease terminology for integrating these two types of biomolecular and clinical data. The cases in which microarray data could not be mapped to clinical laboratory data largely reflect the fact that only pediatric data was used. The unmapped terms contain diseases such as *Parkinson's disease*, *macular degeneration*, *Alzheimer's disease* and other diseases not generally found in children. Other failed mappings represent relatively rare disorders, such as *Yersiniosis* and *Luteoma*. Better mappings might be obtained by leveraging the relational structure of UMLS to map terms that are parent or child relationships to the disease terms.

The many-to-many and many-to-one SNOMED-CT to ICD-9 mappings using UMLS CUIs do present an interesting problem. These could lead to ambiguities

in the mappings such that a highly specific disease variant is mapped to a more generalized disease category. This could have a negative impact on the downstream utilization of the integrated data. The data in table 3 suggests that large source vocabularies like SNOMED-CT have been constrained and compressed by the smaller vocabularies within UMLS to the degree that original source vocabulary resolution is lost. This may suggest an alternative strategy in which the biomolecular samples are labeled only with SNOMED-CT identifiers and the translation between SNOMED-CT and ICD-9 is performed outside of UMLS CUI constraints.

There are several caveats in the interpretation of the results. First off, the data sets were not generalized in that the clinical laboratory data only represented pediatric patients and the microarray experiments were limited to those in which a disease and a normal control distinction was evident. Furthermore, this study offered only a focus on SNOMED-CT and did not apply the same techniques to the alternative disease terminologies mentioned to offer any quantitative comparison. Although the investigation revealed that SNOMED-CT was capable of joining the two data types, it offers no statistical characterization of the joining to assess its overall quality and reliability. Of course we also acknowledge that the text mining aspects of this approach are prone to errors, such as miscodings of the data.

The results demonstrate that current and future translational data integration endeavors can leverage existing clinical terminologies, such as SNOMED-CT, to integrate clinical and biomolecular data types and shift valuable efforts to downstream discovery. Furthermore, this study provides support for the continued development and use of SNOMED-CT for translational data integration, and brings to light the importance inter-terminology mappings resources such as UMLS. As demonstrated by our own work, and the work of others, the straightforward act of integrating data from the molecular and clinical worlds can have profound and direct impact on human health.

Although our initial work focused on the integration of microarray data and patient lab data specifically, we are now working to expand the application of the underlying system to integrate additional data types. In order to integrate new forms of biomolecular data into our current framework we must develop improved text-mining methods to map the underlying experimental data to SNOMED-CT identifiers. From the clinical perspective we will continue to integrate new data obtained from the STRIDE system and look to incorporate additional clinical data types as well.

We must also develop methods to test and improve the reliability of the clinical data, as hospital workers will inevitably miscode a small percentage of the data. We must also account for the fact that the application of clinical codes is subject to a number of non-scientific influences, such as hospital billing policies, insurance companies, and pharmaceutical regulations. Any future work in this area should also entail the development of statistical metrics to evaluate the joining terminology, such that a principled decision can be made to identify the most appropriate terminology for a particular integration scenario.

ACKNOWLEDGEMENTS

This work was supported in part by the Lucile Packard Foundation for Children's Health, National Library of Medicine (K22 LM008261), National Institute of General Medical Sciences (R01 GM079719), National Human Genome Research Institute (P50 HG003389), Howard Hughes Medical Institute, and the Pharmaceutical Research and Manufacturers of America Foundation. The authors would also like to thank Alex Skrenchuck for High Performance Computing support.

REFERENCES

1. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nature biotechnology*. 2006 Jan;24(1):55-62.
2. Dudley J, Butte AJ. Enabling Integrative Genomic Analysis of High-Impact Human Diseases Through Text Mining. *Pacific Symposium on Biocomputing*. 2008.
3. STRIDE. [<http://stride.stanford.edu/STRIDE/>]
4. CDR. [<https://cdr.virginia.edu/>]
5. Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ. Clinical Arrays of Laboratory Measures, or "Clinarrays", Built from an Electronic Health Record Enable Disease Subtyping by Severity. *AMIA Annual Symposium Proceedings*. 2007.
6. Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ. Novel Integration of Hospital Electronic Medical Records and Gene Expression Measurements to Identify Genetic Markers of Maturation. *Pacific Symposium on Biocomputing*. 2008.
7. Pickrell J, Clerget-Darpoux F, Bourgain C. Power of genome-wide association studies in the presence of interacting loci. *Genetic epidemiology*. 2007 Nov;31(7):748-62.
8. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nature biotechnology*. 2007 Jun;25(6):675-80.
9. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*. 2001 Dec 18;98(26):15149-54.
10. Pandita A, Zielenska M, Thorner P, Bayani J, Godbout R, Greenberg M, et al. Application of comparative genomic hybridization, spectral karyotyping, and microarray analysis in the identification of subtype-specific patterns of genomic changes in rhabdomyosarcoma. *Neoplasia (New York, NY)*. 1999 Aug;1(3):262-75.
11. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2004 Jan 20;101(3):811-6.
12. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *The New England journal of medicine*. 2007 Jan 4;356(1):11-20.
13. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, et al. Genomic signatures to guide the use of chemotherapeutics. *Nature medicine*. 2006 Nov;12(11):1294-300.
14. Komatsu M, Hiyama K, Tanimoto K, Yunokawa M, Otani K, Ohtaki M, et al. Prediction of individual response to platinum/paclitaxel combination using novel marker genes in ovarian cancers. *Molecular cancer therapeutics*. 2006 Mar;5(3):767-75.
15. SNOMED Intl. [<http://www.snomed.org>]
16. Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc*. 2000 May-Jun;7(3):298-303.
17. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA Annual Symposium proceedings / AMIA Symposium*. 2006:106-10.
18. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic acids research*. 2005 Jan 1;33(Database issue):D562-6.