Novel Integrative Bioinformatics Approaches to Biomedical Ontology Practice for Translational Informatics

by

Sirarat Sarntivijai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2012

Doctoral Committee:

Professor Brian D. Athey, Chair
Professor Howard Markel
Professor Gilbert S. Omenn
Associate Professor Yongqun He
Assistant Professor Kai Zheng

Dedicated to Rosalind E. Franklin.

Acknowledgements

To my mother who has always been my role model of the woman in science, the lady of the family, and the working mother. To my father who has always been the man in science with kind understanding and a great support to everything that the person afro mentioned does, and my life-long encouragement to become the person that I am today. I have been molded, shaped, and prepared to enter the advanced scientific world with the gracious help from my parents and there are no words that could measure how grateful I am for them.

I wish to thank my thesis advisors and committee for their words of wisdom and advices on my dissertation project. Dr. Brian Athey and Dr. Yongqun He serve as my co-advisors. The word "*loog-sit*" which means student in Thai would best describe what I became in Dr. Athey's BIOINF 526 Fall 2005 class. As *loog* means an offspring, and *sit* puts the offspring in the context of apprentice position, I have since then secretly referred to myself as Dr. Athey's adopted student. Dr. He came along as a much-thanked-for big brother that helped guide me through the time of confusion. I then knew that I also had a family at the University of Michigan Bioinformatics that would help me grow and lend me a hand should I stumble. I am deeply grateful for the valuable insights that I have received from my committee; Dr. Gilbert Omenn, Dr. Howard Markel, and Dr. Kai Zheng on the investigation of the field I was a novice at. Their contribution of kind and thorough remarks on the influenza vaccine study will always be recognized. I also wish to thank Janet Peake who would have taken the place of my number one superwoman, had my mother not taken that already.

There are a few people that I will not forgive myself if I fail to mention and give them a proper acknowledgement. Dr. David States who asked me “Sira, what do you know about ontology?” almost six years ago in his office was the first person who set me off as an ontologist (and the honest-to-god answer I gave him was “I do not know what ontology is, but I will find out”). Dr. Matthias Kretzler and his lab members who were great fun to be around, but they were also the great source of knowledge at the same time taught me much about Diabetic Nephropathy which I did use the case study in my preliminary examination. And that work was shaped into this ontology integration dissertation. All Bioinformatics staff and faculty that I dare not list all the names in fear that I might accidentally leave some names out due to the many years that I have been here, and all collaborators I have worked with are also the recipients of my million thanks.

Among the many friends that have helped me maintain my sanity that I am much obliged to accredit are; Yingluck Thongpenyai, my roommate of early years in Michigan who went back to Thailand a few years ago but continued giving me the moral support that I much needed at times; my brother, Patchares Sarntivijai, who refused to listen to my non-sense stories that should not have bothered me. His refusal helped me realize I was wasting my precious time thinking about the non-sense. And my two cats who have always been my very late night/very early morning entertainment to keep me sane and remind me that I was not alone.

Most importantly, I thank everyone and everything that has taken part in my academic experience up until this day. It has not been a smooth ride on the path paved with roses' petals, but I have learnt of what becomes of life as much as I have learnt of what the academia could offer me. Life is a work in progress and I thank this doctoral experience that reminds me of that. I learn something new everyday. The more I see is definitely the less I know.

Preface

"I belong to a new underclass, no longer determined by social status or the colour of your skin. We now have discrimination down to a science." --- Vincent Anton Freeman (GATTACA, 1997)

The movie came out right around the time I was entering college, also around the beginning of the golden era of the Human Genome Project. I was just a computer science major freshman who had always been fascinated by the X-files. With that kind of creativity and imagination in my personality, of course, GATTACA terrified me horribly to every strand of my hair…and having both parents in medical research science who introduced *Dolly** to the dinner table conversation not long before that did not help at all.

In my freshman year, I had also learnt about 'freedom' in Global Issue class. "*Among the basic freedoms to which men aspire that their lives might be full and uncramped, freedom from fear stands out as both a means and an end*" (Aung San Suukyi, 1990).

So I decided to face my fear to be free.

I joined Bioinformatics after I completed my degree in computer science as it was the next thing on my bucket list to cross off. Because we fear the unknown, I had figured that once the unknown became known, I would be free from my fear. Not understanding what this genome thing was all about made me scared of what harm the new technology could bring to me. Two years invested in a Master degree in Bioinformatics gave me the assurance that it was not all that scary (…yet, with the hope that ethics and humanity remained intact for a long while). But then it would be better if I could make a better use of my life and lend myself as a helper, not just explaining and soothing other people who are where I was fourteen years ago, but being part of this new genomic science and inventing new application that will be beneficial to others as well.

So I conquered my fear, I did not just overcome it.

In this thesis, as a result of me conquering my fear in the course of doctoral program, fundamental concepts of reality are investigated. The term *ontology* has its root in philosophy of the study of the nature of being, existence, or reality. How do we represent *reality*, especially in the biomedical science domain? How can this reality help driving basic science to clinical research, *i.e.* achieving translational informatics? How can we use statistical bioinformatics to distinguish between what is true and what is false? Answering these questions lays the foundation for biomedical ontologies to develop greater applications.

As reality is a concept that must remain true regardless to circumstances it is applied under. When speaking of reality (i.e. ontology as reality representation), all parties participating in such discussion should hold the consensus of what that reality describes. Collaboration of methods to accomplish advanced clinical science as in *translational informatics* is a key element in this thesis to demonstrate how multiple techniques can agree on one concept of reality and work together to achieve a greater good, and thus, the birth of integrative bioinformatics method for translational informatics.

**Dolly* (July 5th, 1996 – Feb 14th, 2003) was a first female domestic sheep and the first mammal to be cloned from an adult somatic cell - http://en.wikipedia.org/wiki/Dolly_%28sheep%29

Table of Contents

Dedication... ii

Acknowledgements iii

Preface vi

List of Figures xiv

List of Tables xvi

Abstract xviii

Chapter 1 Introduction and Overview 1

1.1 Assessment of Current Biomedical Ontologies 1

1.1.1 The Use of Biomedical Ontology 2

1.1.1.1 Knowledge Management with Biomedical Ontology 2

1.1.1.2 Data Integration of Biomedical Ontology 6

1.1.1.3 Decision Support in Biomedical Ontology 7

1.2 The Controversy in the Use of Ontologies 9

1.3 Bridging Experimental Informatics to Clinical Informatics 11

1.3.1 P4 Medicine and translational informatics 11

1.3.2 Creating translational informatics 13

1.4 Previous Work on Ontology Mappings 15

1.4.1 The approach to ontology mapping versus ontology integration 15

1.4.2 Ontology integration: tools and applications 16

1.5 Scope of work in this study 17

1.5.1 Motivation 22

1.5.2 Creation 23

1.5.2.1 The Cell Line Knowledgebase/ The Cell Line Ontology 23
1.5.2.2 The Ontology of Adverse Events 23
1.5.3 Application 24
1.5.3.1 Describing cell cultures in experiment 24
1.5.3.2 Clustering of adverse events by OAE structure 25
1.6 Summary and Conclusion 26
1.7 Publications 27
1.7.1 Published 27
1.7.2 In preparation 27
Chapter 2 The Cell Line Knowledgebase (CLKB) 30
2.1 The importance of cell line in biomedical research 30
2.1.1 Motivation of CLKB development 30
2.1.2 Ambiguity introduced by *de facto* naming creation 31
2.1.3 Cell line contamination 32
2.2 Methods 33
2.2.1 Cell line cataloguing in public repositories 33
2.2.2 Parsing individual cell line's attribute values 34
2.3 Results 36
2.3.1 The Cell Line Knowledgebase (CLKB) 36
2.3.2 Data structure 37
2.4 Discussion 40
2.4.1 Cell Line Ontology 40
2.4.2 Cell Line Nomenclature 40
2.4.3 Discovery of uncatalogued cell line names and synonyms 49
2.5 Conclusion 51
Chapter 3 The Cell Line Ontology (CLO) 53

3.1 Call for the Cell Line Ontology: in the wake of biomedical ontology 53

3.2 Methods: developing CLO 54

3.2.1 Cell line data source 54

3.2.2 Importing external concepts – reusing existing ontologies 55

3.2.3 Defining and annotating CLO-specific ontology terms 56

3.2.4 CLO editing and access 56

3.3 Results 57

3.3.1 CLO-top level structure and statistics 57

3.3.2 CLO design pattern 58

3.3.3 Use case: describing cell lines with CLO – example of Jurkat 62

3.3.4 Use case: application of CLO in bioassay data analysis 63

3.4 Discussion 66

Chapter 4 The Ontology of Adverse Events (OAE) 69

4.1 Overview 69

4.2 Introduction 70

4.3 Methods 73

4.3.1 Adverse event data extraction 73

4.3.2 AE report signal detection with Proportional Reporting Ratios (PRRs) 74

4.3.3 Chi-square test to identify statistically significant AEs 74

4.3.4 Proportional Reporting Ratio score for rankings of vaccine-specific significant AEs 75

4.3.5 Comparison of concept reorganization based on semantic similarity of the Ontology of Adverse Events (OAE) 78

4.4 Results 78

4.4.1 Overall study design 78

4.4.2 Overall results: extracting differential AE profiles from VAERS TIV and LAIV data 80

4.4.3 Distinctive underlying biological activities were associated with the two groups of influenza vaccines 89

4.4.4 Severe AEs are highly enriched in killed-inactivated influenza vaccine group 91

4.4.5 Post-immunization Guillain-Barré Syndrome (GBS) in TIV recipients, occurred at a higher rate per number of reports than in LAIV 93

4.5 Discussion 104

Chapter 5 Future Directions and Conclusions 115

5.1 Summary of previous sections 115

5.2 Applications to improved health informatics 118

5.3 Bundling CLKB/CLO as standardized cell line nomenclature source 118

5.4 Maintaining OAE and CLO 119

5.5 The future of OAE: OAE as a hypothesis generator 120

5.5.1 Adverse event data extraction and statistical analysis of AEs 121

5.5.2 Natural language and ontology processing with UMLS 121

5.5.3 Gene-concept network reconstruction with ConceptGen and MiMI 122

5.5.4 Preliminary results 125

5.5.4.1 Statistical analysis of post-influenza-vaccination identified enriched severe adverse events in TIV, but not in LAIV 125

5.5.4.2 Natural language processing of PubMed abstracts extracted gene-AE pairs of TIV-specific significant AEs 126

5.5.4.3 Gene concept analysis revealed different enriched concepts with differential VAEs 129

5.5.4.4 Gene network analysis predicted candidate genes underlying related molecular activities of severe adverse events in TIV 133

5.5.5 OAE as a potential hypothesis generator 140

5.6 Conclusions: Ontology services to bridge to translational research 141

References 142

List of Figures

Figure 1.1 Venn diagram of selected scopes in ontology-driven biomedical questions to aid translational informatics. 18

Figure 1.2 Outline diagram of the overall scope of work in the integrative biomedical ontology study. 21

Figure 1.3 The concept diagram of the integrative biomedical ontology approach to translational informatics. 29

Figure 2.1 UML Diagram describing the entities and relations of the Cell Line Knowledgebase 38

Figure 2.2 Screen capture of the Protégé interface for CLKB 39

Figure 3.1 The top level CLO hierarchical structure of ontology terms. 58

Figure 3.2 Basic design pattern for representing cell lines in CLO. 59

Figure 3.3 Modeling Jurkat and its derivative Jurkat Clone J.Cam1.6 using CLO. 63

Figure 3.4 Application of CLO in bioassay data integration and analysis. 65

Figure 4.1 Signal curves to determine the data cut-off for TIV and LAIV analysis. 77

Figure 4.2 Workflow of integrative AE bioinformatics analysis. 79

Figure 4.3 Venn diagram summary of the three filtering criteria in each group of vaccines from the pool of 3582 AEs analyzed in TIV and LAIV and the universe of 7520 AEs in the entire VAERS database; Chi-square value of >= 4 – χ^2(+), or < 4 – χ^2(-); PRR >= 2 – PRR(+), or PRR < 2 – PRR(-); and number of reports >= 75 in TIV or >= 8 in LAIV – count(+), or else – count(-). 82

Figure 4.4 Diagram of AE count grouped by related symptoms. 90

Figure 4.5 Reporting rate of GBS cases comparison between TIV and LAIV................ 94

Figure 4.6 Report distribution comparison by year (a) TIV (b) LAIV. 102

Figure 4.6 Report distribution comparison by age range (a) TIV (b) LAIV. 103

Figure 4.7 Two groups of vaccine adverse events classified by OAE........................... 112

Figure 4.8 Two groups of vaccine adverse events classified by SNOMED-CT 113

Figure 4.9 Two groups of vaccine adverse events classified by MedDRA 114

Figure 6.1 Summary of the process diagram. 124

Figure 6.2 Overall gene interaction network of TIV-specific associated genes. 129

Figure 6.3 TIV-induced GBS-related gene activities proposed by ConceptGen/MiMI interactions. 139

List of Tables

Table 1.1 Summary of use cases and scope of biological-driving problems (DBP) 22

Table 2.1 Example of synonymous cell lines. 42

Table 2.2 Top ten HyperCLDB cell line names that appeared in non-biomedical text corpus ranked by number of occurrences. 43

Table 2.3 HyperCLDB cell line names that appeared in NCBI GEO Microarray sample description, ranked by number of occurrences. 43

Table 2.4 Ten most common cell line names in GEO microarray sample description demonstrating trade-off of recall and specificity. 46

Table 2.5 Ten most common cell line names in GEO microarray series description demonstrating trade-off of recall and specificity. 46

Table 2.6 Twenty most common cell line names in PubMed sentences containing 'cell' or 'cells' tokens demonstrating trade-off of recalls when increasing specificity. With a large dataset, increasing specificity leads to optimal recalls. 47

Table 2.7 Examples of cross-contaminated cell lines. 47

Table 3.1 Summary of ontology terms in CLO and source ontologies used in CLO. 55

Table 4.1 TIV-specific adverse events. 37,621 TIV-induced AE cases were reported. (* = serious adverse event, (r) = related to serious adverse event) 83

Table 4.2 LAIV-specific adverse events. 3,707 TIV-induced AE cases were reported. .. 85

Table 4.3 Summary of association of GBS to influenza vaccines in peer-reviewed literature 95

Table 6.1 Summary of TIV-specific AE-geneID association by PubMed sentence co-localization. 127

Table 6.2 Enriched gene concepts (GO function or process) associated with TIV- and LAIV-induced VAEs 131

Table 6.3 Summary of GBS related AE molecular network reconstruction data (summarized from Cytoscape input file) 133

Abstract

Ontologies can be used to annotate, integrate, and interoperate against a large amount of biomedical data in support translational informatics research and practice. This thesis demonstrates these capabilities in two case studies: 1) the Cell Line Ontology (CLO), and 2) the Ontology of Adverse Events (OAE).

Cell lines have been generated and are widely used in biomedical research. To analyze cell line nomenclature and support cell line data integration, we first created a cell line knowledgebase (CLKB) with a well-structured collection of names and descriptive data for cell lines cultured *in vitro*. A CLO was then developed based on the CLKB and other related work by our collaborators. The follow-on community-based CLO development has followed the Open Biological and Biomedical Ontologies (OBO) Foundry principles. Cell lines contained in CLO are associated with terms from other ontologies such as Basic Formal Ontology (BFO), Cell Type Ontology (CL), and NCBI Taxonomy. A common design pattern was developed and applied to model cell lines and their attributes. Currently CLO contains over 36,000 cell line entries obtained from American Type Culture Collection (ATCC), HyperCLDB, Coriell, and by manual curation. The Web Ontology Language (OWL)-based CLO is machine-readable and supports automated reasoning. CLO has been used in various applications. For example, CLO has been applied to support cell line data normalization and biochemical assays. CLO structure also allows information storage for cell line's genetic signature that assists the process of cell line authentication. The usefulness of CLO is also highlighted in the effort to create CLO consortium with collaborations of the European Bioinformatics Institute (http://www.ebi.ac.uk/), The Bioassay Ontology group at the University of Miami (http://bioassayontology.org/), and the Riken Cell Bank at Riken Institute in Japan (http://www.brc.riken.jp/lab/cell/english/).

The Ontology of Adverse Events (OAE) was developed to support adverse event (AE) data standardization and systematic analysis. In this thesis, the OAE was used to analyze clinical AE reports from vaccinations with trivalent inactivated influenza vaccine (TIV) from 1990-2011 and trivalent live attenuated influenza vaccine (LAIV) from 2003-2011. From the CDC Vaccine Adverse Event Reporting System (VAERS), 37,621 AE reports were associated with four TIVs (Afluria, Fluarix, Fluvirin, and Fluzone) and 3,707 AE reports for the only LAIV (FluMist) were extracted. Our comprehensive statistical method includes background noise filtering, the Proportional Reporting Ratio (PRR) analysis, and Chi-square significance test. In total, 48 TIV-enriched and 68 LAIV-enriched AEs were identified as statistically significant. These enriched AEs were then classified with reference to OAE structure. TIV-enriched AEs were found to be clustered in neurological and muscular processing. In contrast, LAIV was found to be associated with AEs in the areas of inflammatory response and respiratory system disorders. Higher reporting rate of severe AEs such as Guillain-Barre Syndrome (GBS), and paralysis were associated with TIV. Cases of post-LAIV GBS were manually examined to conclude the lower risk of these severe AEs in LAIV recipients. These results indicate that TIV and LAIV induce different SAEs responses in human patients.

To identify potential genes and gene interaction networks that are critical in regulating the TIV or LAIV associates AEs, a natural language processing (NLP) method was used to search all PubMed abstracts and identify those genes associated with the enriched AE terms. In total, 130 genes have been found to interact with TIV-associated AEs, and 223 genes are associated with LAIV vaccination. The results of the gene concept enrichments are consistent with term classification by OAE. New potential hypotheses have been generated to elucidate the genetic mechanism of the GBS formation.

This thesis presents novel ontology-based approaches to solve biomedical problems raised from basic research (i.e., cell line) to clinical data analysis (i.e., vaccine AEs). An ontology-based method that integrates both basic and clinical data representation and analyses will be critical for advanced translational informatics research.

Chapter 1

Introduction and Overview

1. Introduction

1.1 Assessment of Current Biomedical Ontologies

When the National Center for Biomedical Ontology (NCBO) was founded in 2005, majority of biomedical ontologies were still created *de Facto* by laboratory personnel at work with the primary purpose to generate structured controlled vocabulary for annotation in knowledge representation. Ontologies constructed prior to this era were mainly based on the needs to systematically catalogue of biological findings, such as the Gene Ontology (first known discussion at the International Conference on Intelligent Systems for Molecular Biology 1998 Montreal, http://www.geneontology.org/gene.ontology.discussion.shtml). Very small portion of biomedical ontologies in use at that time were constructed from computational aspect of the biology. Ontologies can facilitate large-scale computations in biomedical research that allows decision support and knowledge discovery. However, with the primary usage of biomedical ontologies in quality assurance and quality control of data annotation, detailed description of ontological elements that was specific to a single user (creator of an ontology) had prevented biomedical ontologies to reach its full potentials of computations. This has been the controversy of biomedical ontology usage in biomedical research. As ontologies should be reused as well as promote interoperability, the lack of consensual format of information stored in individual ontology has obstructed this process. Ontology mapping by importing terms from other reusable ontologies is an approach to the question of knowledge transfer and discovery. This approach supports both reusability and interoperability among ontologies to accommodate and integrate different data types (and sometimes, different domains as well). Although the concept of ontology mapping can be traced back to almost three decades ago[1], it has not been well achieved due to the stated reason. There have been many attempts to overcome this shortcoming. Recent work includes the alignment of anatomical ontologies[2], and multi-level biomedical ontology mapping[3]. The initiatives at the National Centers for Biomedical Computing (NCBC) led by NCBO and the National Center for Integrative Biomedical Informatics (NCIBI) also concentrate on the importance of this issue[4].

1.1.1 The Use of Biomedical Ontology

The major roles of biomedical ontologies in research can be divided into three arenas; knowledge management, data integration, and decision support[5]. This statement however does not imply equal utilization of ontologies among the three areas. Discussion of the insights of this incident is described as follows.

1.1.1.1 Knowledge Management with Biomedical Ontology

An ontology provides a list of entity names in one domain along with description of relations among them. Therefore, it comes naturally that one major role of ontologies is being the source of vocabulary of a particular domain. However, when there are multiple ontologies describing entities in the same domain, but with different perspectives (e.g. genes can be speculated for their functions, chromosomal locations, or even interactions with other molecules), inconsistency of entity naming and description may occur. Terminology and definition provided in an ontology may be inaccurate (examples of this incident are discussed throughout this study). However imperfect biomedical ontologies medical may be at the current situation, the use of ontologies as a source of vocabulary remains of utmost importance.

Roles of biomedical ontologies in data annotation can be differentiated into three groups; indexing and annotating, accessing, and referencing. Assignment of controlled vocabulary to entities facilitates the creation of organized catalogue. Medical Subject Headings [6] have been used in PubMed/MEDLINE indexing for many years. MeSH is a well-defined controlled vocabulary set, but it lacks a full structural definition of relations to be a complete ontology. MeSH hierarchical structure and the terminology of may define an 'ontology' in many cases. However, with the exponential growth of knowledge in the biomedical field both horizontally and vertically, MeSH has fallen behind far from being comprehensive to include all the detailed description of domain knowledge. One of many examples of this incident is, there are less than 40 cell lines listed in MeSH, most of these cell lines are long-established cell cultures, while there are at least 9,000 cell lines (and still counting) catalogued in public repositories[7]. Assigning MeSH descriptor for indexing, though manually for the most part, has been attempted for automated processing via many tools[6,8]. Advanced automated indexing is becoming necessity but it is still a challenging issue because MeSH reference set of controlled vocabulary cannot keep up with the growing depth of information. In addition, a study by Ozgur et al.[9] demonstrated the use of the Vaccine Ontology in querying PubMed database to retrieve a much more comprehensive search results.

Unified Medical Language System (UMLS) Metathesuarus along with other controlled vocabularies (e.g. International Classification of Diseases (ICD), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)) have been used for the purpose of clinical indexing (coding). The UMLS Metathesaurus contains the mapping of entities from different vocabulary sources in the same domain. Even through the UMLS Metathesaurus may appear to be enriched with information across biomedical domain, a close examination has revealed that the mappings are incomplete or inaccurate [10].

The functional description of experimental data in biology (‘annotation’) is also another major utilization of biomedical ontologies which is mostly accomplished through manual curation. Again, semi-automatic methods have been developed for annotation[11,12,13,14,15]. The rapid growth of knowledge pool calls for a fully-automated annotating process. ‘Term recognition’ is a required technique to achieve such process (it is a key process in Natural Language Processing (NLP) that is much needed for automated processing in ontology utilization). Proficient term recognition is still difficult at the moment as the reference ontologies are not well established for the purpose (see 1.2 The Controversy).

Ensuring the accurate information retrieval (accessing) is the next crucial step after indexing and annotating. User-defined query may or may not result in a search term that exactly matches the index in the stored catalogue. Accessing the information should take advantage of the synonym listing as it is part of the features in most ontologies. UMLS Metathesaurus holds an extended collection of synonyms through its mapping and has been used extensively in the information retrieval through its high-level categorization and co-occurrence information features. This includes accessing both the MEDLINE biomedical literature[16], the online artifacts (medical textbooks, images, and computational problems)[17,18,19] Many publicly available databases also exploit the biomedical ontology structure in addressing and accessing the queried information. For example, Gene Ontology has been used to access many microarray experiment databases[20,21,22,23], using SNOMED CT in finding specialized physicians[24]. Other than using the term tokens from ontologies in accessing information, concept features (in terms of structure) can also be used for topic detection in medical texts[25].

Indexing and annotating, and accessing biomedical ontologies lead to the third utilization of these ontologies, reference and mapping. More ontologies have become available and involved in biomedical research each day. Having an ample supply of biomedical ontologies of the same or different domains taken from different perspectives does not guarantee a solution to knowledge discovery. It is always beneficial that user can select one or a few ontologies from the pool of several dozen available biomedical ontologies. However, integration of resources annotated by different ontologies can be difficult to attain. Referencing and mapping of these entities to identify equivalent concepts across ontologies should be acquired and placed for knowledge discovery use. High-level synonym mapping and categorization has been carried out in UMLS Metathesaurus. However, the different scopes and large granularity of biomedical domain are the key factor of the challenge of direct mapping through synonyms: and thus, it is not appropriate to claim UMLS as a complete collection of mappings across multiple ontologies[2,26]. The Foundational Model of Anatomy has been significantly used as a reference ontology for anatomical alignment across anatomy ontologies[27]. But FMA might have the sufficient coverage of knowledge as there are a few other anatomy ontologies such as Cell Type Ontology[28] and BRENDA Tissue Ontology[29] that provide information from other important viewpoints that should also be considered in the anatomical mappings.

1.1.1.2 Data Integration of Biomedical Ontology

One key difference between the two data containments, database and ontology, is that ontology is equipped with the structural element that promotes semantic interoperability and information exchange. Therefore, biomedical ontologies often take the role of *'standards'*. For example, standardization of patients' data for exchange format across multiple electronic health record (EHR) systems[30], or the use of biomedical ontologies in the application of *SAGE* clinical guideline model[31]. We can describe the role of biomedical ontology in data integration as information exchange and semantic interoperability, and information integration.

Other than the use of individual ontologies in information exchange strategy (e.g. RxNorm, UMLS, SNOWMED CT, LOINC, and HL7)[32,33,34,35], there are also semantic interoperability projects in biomedical field in place today, most of which rely on this aspect of ontologies. In cancer research, the Common Ontologic Representation Environment (caCORE) infrastructure[36] was developed to support the interoperability of biomedical information system utilizing ontologies (such as NCI thesaurus) as key elements of the system. The Biomedical Research Integrated Domain Group has also developed *BRIDG* model claimed to support practical application of data interchange for clinical research[37].

Data exchange and semantic interoperability is a very important role of biomedical ontologies. Data and information integration itself is also a crucial aspect in data storage. As clinical and biomedical research grows in its complexity, types of data that are output from the research also become very complex and sophisticated. Information storage and management is a necessity. Data manipulation and management of various complex data types calls for an efficient data integration. Warehousing is a method that takes advantage of ontologies and transforms different information to a common format or vocabulary[5]. This approach has been exemplified in the integration of model organism databases to functional annotation of gene products in Gene Ontology[38]. However, loss of intricate information may occur due to this transformation.

This leads to another approach in data integration to prevent the loss of information, mediation[5]. Mediation-based method uses a high-level ontology as a global schema to create mappings to local schema of lower-level ontologies. For example, UMLS is used as a global schema in *ARIANE* to manage information accessing to various medical databases[39]. One drawback of this approach is the global schema, although global, may not be universal and thus missing necessary links to part of the local schema.

1.1.1.3 Decision Support in Biomedical Ontology

Biomedical ontologies are crucial for decision-support systems. Data selection and aggregation query specific pieces of information needed to make a decision. For example, the International Classification of Disease (ICD) is persistently used to select a group of patients in relations to high-level disease description, or SNOMED CT is used to query clinical data warehouses[5,40,41]. Once groups of candidate information are identified, filtering information (aggregation) for further investigation is also performed through ontologies (e.g. Gene Ontology in gene clustering[42,43].

Role of biomedical ontologies in providing a standard vocabulary and integrating resources aids decision support development (e.g. resolving drug names to standard codes and mapping these codes to a knowledgebase[5]). Furthermore, a more mature feature of ontology as a source of computable domain knowledge can be utilized in the course of decision support. Ontologies developed from advanced computational perspective are often equipped with structures that allow efficient reasoning. The Foundational Model of Anatomy (FMA) is used as a source anatomical knowledge for reasoning of penetrating injuries[44].

Advanced Natural Language Processing (NLP) is another application that utilizes ontology vocabulary and structure for term recognition in extracting specific information such as relations, summarizations, and literature-based discovery[27]. Cell Line Ontology was used to discover cell line names in published literatures that had not been catalogued in public repositories, and identify the issues of common cell line names and cross contamination[7]. Knowledge discovery in the case of Cell Line Ontology is one of many examples of how biomedical ontologies help change the phase of clinical and experimental research from hypothesis-driven to data-driven.

1.2 The Controversy in the Use of Ontologies

Data organization in biomedical field evolved over time, from a structured *'binomial system'* of Carl Linneaus[45] to Charles Darwin's simple yet detailed record keeping of evidences of *'natural selection'*[46] to a more comprehensive data structure of relational database. Today, we are seeing more and more of advanced relational database with knowledgebase that can reason, *ontology*.

Dated back in early 17th century, the word *'metaphysica'* (later known as *'ontology'*) was first used by Aristotle's students to describe "the science of being *qua* being"[47]. In philosophical sense, ontology therefore is the study of existence, or reality. When applied to information science scope of the 19th century, ontology has taken the role of being a representation of a set of concepts that describe reality with the relationships between those concepts in one domain. In theory, ontology is a knowledge representation that provides a shared conceptualization, and hence promotes interoperability, reusability, and homogeneity. Ontology is also created to fit the designer's need/purpose and therefore can be defined in any way taken from any perspective. There are no rules of how the creator should define his/her ontology.

But when it comes to a biomedical domain that contain heterogeneous information, ontologies as *'standard framework'* are facing a difficult time. The nature of ontology creation that is defined to fit the creator's need has introduced obstacles in finding the right standard framework (which one, among the many, is the *'standard'*?). In the initiative led by the NCIBI and NCBO of classifying the biomedical ontologies in use, it was agreed that the more expressive and general an ontology is (for universal coverage as a standard), the less computable it will be (ontology to aid interoperability, reusability, and knowledge discovery)[4]. Automated data integration as a necessity that results from the unbound growth of biomedical information is suffering from this disparity of ontology creation definition. In most cases, the creator must make a decision whether to go with expressiveness or computability. As most-popular use of ontology is annotation, majority of biomedical ontologies are designed to accommodate general expression rather than computability needed for automate data integration. This statement is reflected in a format design decision. While the W3C Web Ontology Language (OWL) format[48] is equipped with the structure that supports automated reasoning and computability, many biomedical ontologies in use these days were originally designed in an Open Biological Ontology [49] [49] format[50] for its simple structure that purposely serves the annotation needed at the time of creation but lacks the advanced computational ability.

There are constructive responses to this controversy of biomedical ontology structure and creation. For example, the conversion of OBO to OWL format for the better computation has been implemented[51,52]. However, unless the converted OBO to OWL format is extensively manually curated or executed with advanced computational method, the resulting OWL file yields not much further information or additional computability. The use of UMLS metathesaurus as a semantic mapping provider may answer the question of this controversial issue. But UMLS contains linkage mapping across hundreds of terminologies and ontologies that versioning of some of these terminologies and ontologies that are still dynamic can cause inconsistencies, or the mapping may be incomplete[51]. This is evidence that a novel approach that combines multiple method to map ontologies should be considered. With this being said, along with the issue of ontology creator's specificity requirement versus generalization for interoperability, the trend in ontology mapping has been moving toward the approach of community consensus. That is, to find and establish some upper ontology that defines the skeletal scope and relations that can further be reused and expanded to fit the creator's need of specificity. The OBO foundry[53] as a biomedical ontology community has played a central role in persuading the bio-ontologists to work together in constructing a guideline to help integrate and reuse available ontologies with customized addition to create a new ontology that fits the creator's need. This approach of ontology integration will be described in later chapters of the creation of the Cell Line Ontology and the Adverse Event Ontology.

1.3 Bridging Experimental Informatics to Clinical Informatics

Other than the data integration within experimental sub-sections of biomedical domain as described previously (e.g. cells-cell lines-genes), data interchange across sections of biomedical informatics pool is also crucial (e.g. health records-phenome-genome). It is a stressed importance that elements in this experimental informatics must be able to move across the boundary to be used practically and efficiently in clinical informatics. Moving from experimental informatics to clinical informatics and vice versa introduces the new approach to medicine resulting in *translational informatics.*

1.3.1 P4 Medicine and translational informatics

The new trend in health care has demonstrated the concept of *'P4 Medicine'*[54] that consists of the four P's: Predictive, Personalized, Preemptive, and Participatory. This concept introduces the movement from doctor-centric health service to interaction-based service that patients take part as participants in the service. *'Predictive'* and *'Personalized'* in the four P's are a result of the modern biomedical research, while *'Personalized'* and *'Preemptive'* have led the patients to *Participate*. The relations among the four P's cannot be made possible without data interchange and integration.

Data integration within the experimental research domain is already a challenging issue in itself. Bridging experimental research to clinical/bed-side research can be even more challenging. Making these connections opens a new page of challenges in data integration and information management. Results from high-throughput experimental methods require certain statistical analyses and corrections to ensure the un-biased findings. Clinical trials produce a large set of data to be used as an evaluation framework that also requires statistical analyses to produce useful findings. It is not always easy to compare the two sides of statistical analyses when it comes to pooling the data from the two sides (experimental and clinical) together for a successful integration. Moore's law[55] indicates that data are doubling in their size every two years and this results in an exponential rate of growth. In biomedical research, this incident implicates the complexity in knowledge of diseases arisen from the growth of data that also transforms biomedical research to a multidisciplinary field. Domain experts from different arenas need to engage in a communication for a successful collaborative work. Methods used in combining and integrating these data must also be scalable as no one can predict the end of a full-grown biomedical data pool capacity.

1.3.2 Creating translational informatics

The interconnections between experimental informatics and clinical informatics have introduced us to the study of *'translational informatics'*. Integration is the key to translational informatics as described in the previous section. Integration of large sets of data necessitates automated processing of these interrelating data, which makes annotation standard a stipulation. Biomedical ontologies appear to be a promising solution to this question.[3] has described the trend and issues of knowledge integration and infrastructure needed for ontology mapping approach to achieve translational informatics. Effective data repositories of experimental data must be implemented and equipped with a well-established standard protocol for data sharing (e.g. the Gene Expression Omnibus (GEO)[56], Stanford Microarray Database (SMD)[57], and ArrayExpress[58,59]). The standard protocol is guided by the Minimum Information About a Microarray Experiment (MIAME). However, even with the guideline such as MIAME, not all the data available in these repositories are equivalently represented as the MIAME framework requires the *minimal* information and most biomedical systems work independently of each other, and therefore there are no common elements or structures[60,61]. Automated systems to combine these heterogeneous data need computable form of knowledge and can be acquired through knowledge sources such as ontologies that contain various knowledgebases[3]. Example shown in this case, as will be describe later, is the collaboration in the work of Cell Line Ontology that is utilized as a component of knowledgebase in the European Bioinformatics Institute's attempt to map GEO to ArrayExpress. Biomedical ontologies are also heterogeneous in their nature that a consortium or a federation as a medium ground to define the quality assurance/quality control of biomedical ontologies needs to step up and take this role. The NCIBI and NCBO under the NCBC NIH Roadmap initiatives have been active and leading this federal role.

Patient's records and reports are also an important source of information once mined carefully. The challenge of using patient's information as an input for bioinformatics analyses will always remain in the area of the data noise management and processing. Patient's records will always be statistically noisy as the data entry procedure both by the health care providers and untrained reporters are sociological science process, not solely computational scientific process. Factors of human behavior always hold accountable in patient's information reporting. However noisy and statistical unstable patient's record information may be, patient's records are still desirable as a great source of information. This comes back to the question that the real world is not perfect, how can we work around this imperfect world? The answer proposed in this study is, we can work with imperfect but useful data by using the combination of statistical bioinformatics and biomedical ontology approach to clean the data, process the data to identify significant signal, and to *translate* these data into a useful knowledge. The work to achieve such process includes 1) transforming the Electronic Health Record (EHR) data entry procedure to a more integratable format with the aid of biomedical ontologies (discussed by not included in the scope of this study), 2) developing a workflow to retrieve patient's data, and statistically identify true signals. This workflow is described in detailed in our case study of the comparative analysis of adverse events induced by killed Influenza vaccines, and live Influenza vaccine.

1.4 Previous Work on Ontology Mappings

Ontology structure can be viewed as a computational graph that consists of nodes (entities/classes) and edges (relationships). Graph matching theory can be dated back to early 1980's[62]. Despite the long years of graph matching study, ontology has only recently begun to show its impact in biomedical world. Kalfoglou et al.[63] has described the insights of issues and techniques in ontology mapping from the computational perspective. Here, they described ontology mapping as *"the state of the art"* with the discussion of role of ontology as a *'mediator', 'translator', 'framework', 'survey',* or *'technique'*. Up until today, the roles of ontologies and what is expected of ontologies remain inconclusive, although we have been observing the trend of ontologies as being the key *integrator,* as will be the theme of this study.

1.4.1 The approach to ontology mapping versus ontology integration

Other than the complexity caused by the bottom-up design of ontology creation (i.e. diverse work on ontologies originated from individual organizations of the same domain), ontology mapping is also very computational challenging as it is a graph matching problem that has been proven to be NP-Complete[64]. Statistical approach to ontology mapping is considered one of the solutions to find mapping via heuristic algorithm. Sub-graph matching and identification in graph-pattern matching by pruning operation to find a homomorphic image in a target graph has been proven efficient in this question[65]. To date, the-state-of-art ontology mapping faces even further questions that have not been answered. Definitions vary from community to community; terminologies of processing techniques are not consensual. The attempt to un-complicate these ontology matters is demonstrated in the invention of the *Relation Ontology* [4] to describe the relations of relations in biomedical ontologies[66]. This would require that the new-released ontology conform with some community-influenced standard as such. The development of Basic Formal Ontology (BFO)[67] along with RO has introduced the biomedical community to a consensus-based approach for ontology mapping. This is the approach that we take in this study in demonstrating that a community-endorsed set of upper-level skeletal ontologies can help gearing towards *ontology integration*, rather than taking the approach of complicated NP-hard problem with ontology mapping as said.

1.4.2 Ontology integration: tools and applications

PROMPT tool under the scope of *Protégé* development at Stanford/NCBO[68,69] is one of the most popular ontology mapping/merging tools in use if one were to take the approach of computational graph-matching mapping. PROMPT has been proven very powerful in version management, and making *semi*-automated suggestions to the ontology comparison at work. However popular it is, PROMPT is not well equipped to answer the question of mapping ontological structural differences that are often seen when dealing with multiple ontologies from different sources. Additional scripting and implementation required to get PROMPT to work well under these circumstances can be equivalent to developing yet another application at all, and this is only accomplished on a case-by-case basis (meaning another scripting and implementation process would be required should the condition of the starting question change).

Why do we need to map ontologies? This can be simply answered by *'integration'*, data integration for knowledge discovery. While the basic comparison of ontologies in different versions can be easily answered by PROMPT tool, a more sophisticated approach is much needed for knowledge discovery in translational science. Extension of Gene Ontology [14] by combining GO with external vocabularies other than the three original universal concept domains in GO yields a larger and more specific vocabulary that includes organism-specific anatomical terms and proliferation of terms and relationships, and provides a possible method for evaluation of hypothetical concepts[70]. Relationships detected from the integration of GO, ChEBI[71], Cell Type Ontology[28], and BRENDA Tissue Ontology[29] provide the insights on relationships between these ontologies when examined with careful evaluation of matching strategies[72]. Point-to-point mapping of concepts among various anatomical ontologies in Zhang's work[2] is a proof of concept that fully-automated large-scale anatomical ontology alignment may provide a comprehensive framework for machine-learning-based computations in biomedical and translational research. This thesis work adapts the said work by Noy et al. and Zhang et al. to create an ontology integration technique from a slightly different dimension (rule-based, natural language processing, and ontology-structure conformance matching) to create a framework that aids hypothesis evaluation in the scope of translational informatics as discussed throughout this paper.

1.5 Scope of work in this study

Ontology mapping and ontology integration is one of the key solutions to answer the question of bridging basic research to translational research. However, it is apparent that the challenge in ontology mapping as stated previously requires a creative approach to overcome the issue. This study demonstrates utilization of biomedical ontology using a novel combinatorial bioinformatics methods in order to answer questions in translational informatics domain. The practice can be logically viewed in three layers; from the perspective of motivation, which leads to the creation of new ontologies, to the application for the better representation in a machine-readable format. The thesis applies both the existing and newly-created ontologies to the translational informatics domain (figure 1.1).

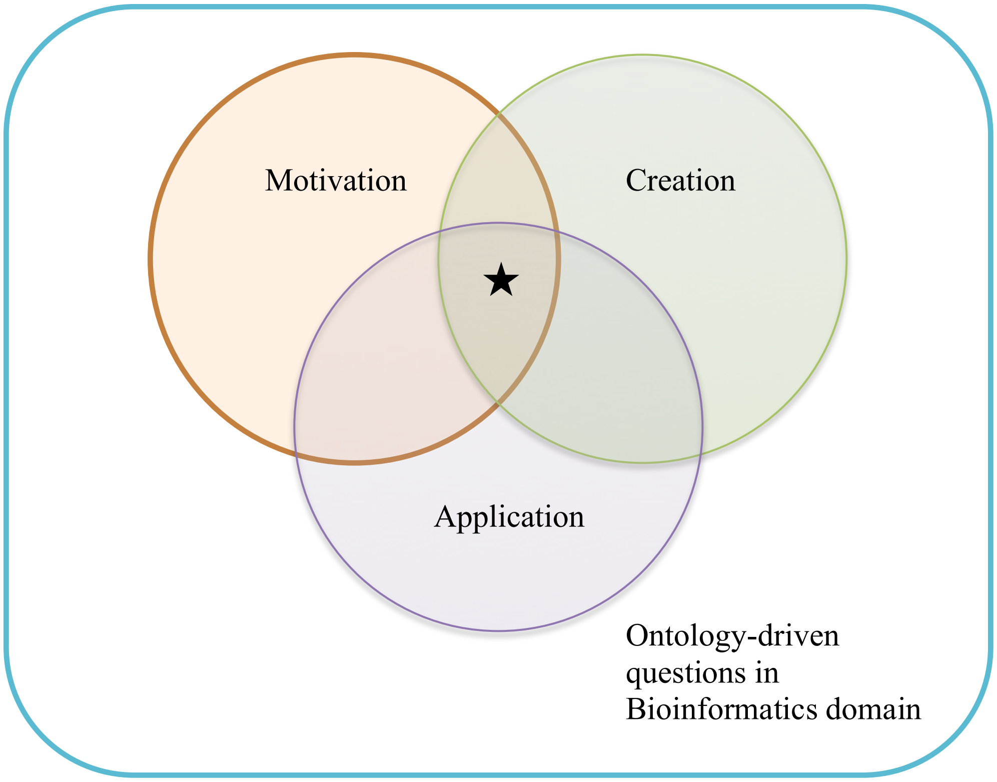


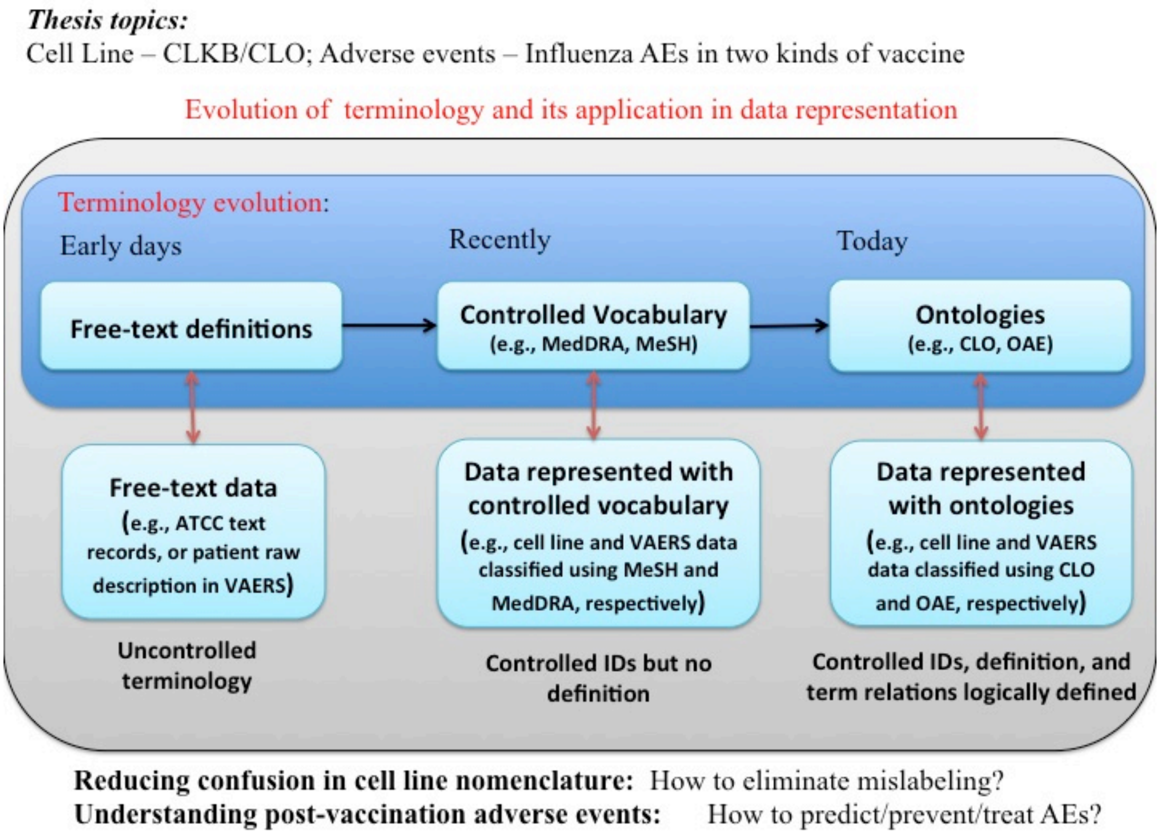
Figure 1.1 Venn diagram of selected scopes in ontology-driven biomedical questions to aid translational informatics.
Among the ubiquitous techniques in the now-advancing bioinformatics methodologies, there exist the three highlighted crucial aspects that are identified as the key components to achieve biomedical translational informatics. Challenges in large data integration management and knowledge transfer lead to the *Motivation* to find a medium tool that best accommodates solution to the challenges. This also prompts to building biomedical ontologies that have not previously existed to help establish the infrastructure to the solution (*Creation*). The *Application* aspect demonstrates the *how-to's* in utilizing

available resources to answer driving biological questions. Designing the framework that links all these three aspects together results in a novel bioinformatics approach that supports translational informatics (★).

This thesis explores two areas of biomedical research that ontologies can assist moving towards translational research. Figure 1.2 outlines the evolution of data in which free-text information evolves to controlled vocabulary and eventually becomes an ontology. A well-designed ontology can then be used as a tool to answer addressed clinical questions from components of basic research. The case study of cell cultures with the creation of CLKB and CLO aims to answer the question of how to reduce confusion in cell line nomenclature caused by mislabeling from either data-entry error or cross-contamination. Cell line data were originally PDF catalogue and HTML file entries before they were processed for a controlled vocabulary used as a dictionary in CLKB. Development of CLO was the bridge to answer the question of how to monitor for quality assurance and quality control of cell line indexing process by data standardization and knowledgebase construction to minimize cell line mislabeling.

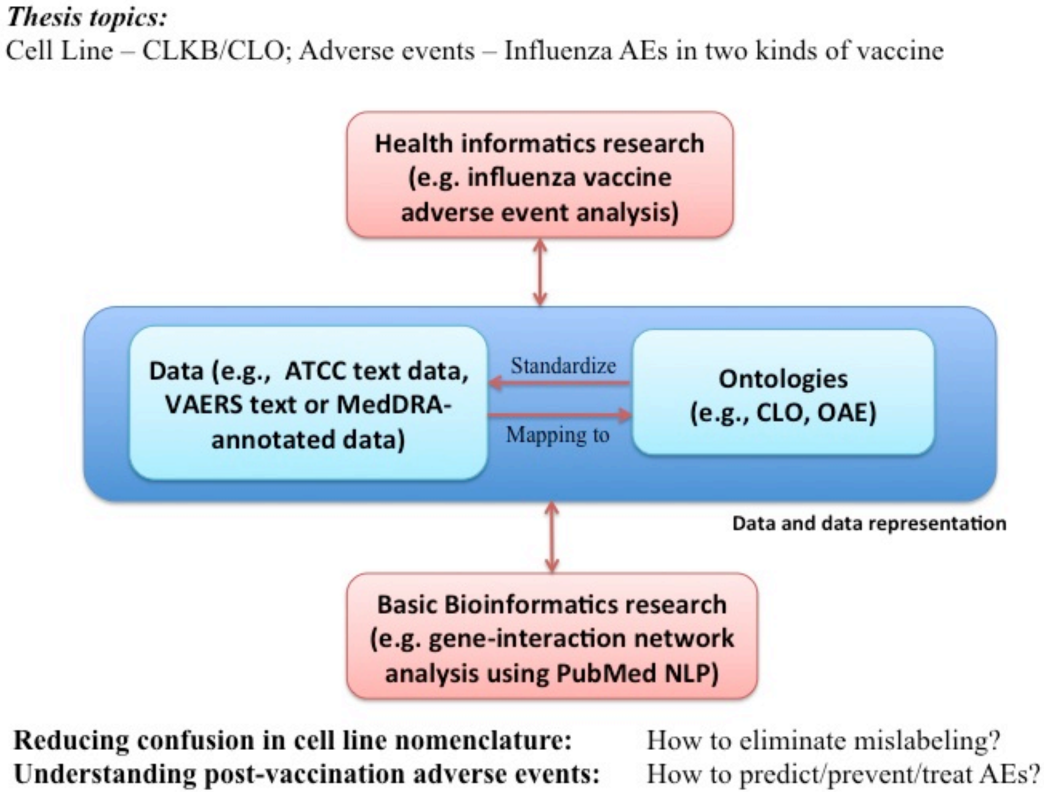
The case study of OAE in analyzing post-vaccination adverse events is described in this thesis to showcase how ontologies can be applied at different levels of biomedical research. Free-text patient records of influenza vaccine recipients from VAERS were retrieved to analyze the differentiate profiles of two types of influenza vaccines (TIV versus LAIV). AE terms that were examined in the two groups of vaccines were reorganized to OAE with reference to MedDRA dictionary (a controlled vocabulary used in reporting AE to VAERS). Gene-network analysis was the basic research component that was used to identify which molecular components were foundation for hypothesis construction in answering the question of how to predict, prevent, or treat AEs.

Detailed discussion of how *motivation, creation,* and *application* concepts were applied in this thesis is described in the following sections.



Thesis topics:
Cell Line – CLKB/CLO; Adverse events – Influenza AEs in two kinds of vaccine
Evolution of terminology and its application in data representation
Terminology evolution:
Early days
Recently
Today
Free-text definitions
Controlled Vocabulary (e.g., MedDRA, MeSH)
Ontologies (e.g., CLO, OAE)
Free-text data (e.g., ATCC text records, or patient raw description in VAERS)
Data represented with controlled vocabulary (e.g., cell line and VAERS data classified using MeSH and MedDRA, respectively)
Data represented with ontologies (e.g., cell line and VAERS data classified using CLO and OAE, respectively)
Uncontrolled terminology
Controlled IDs but no definition
Controlled IDs, definition, and term relations logically defined
Reducing confusion in cell line nomenclature: How to eliminate mislabeling?
Understanding post-vaccination adverse events: How to predict/prevent/treat AEs?

(a)


Thesis topics:
Cell Line – CLKB/CLO; Adverse events – Influenza AEs in two kinds of vaccine
Health informatics research (e.g. influenza vaccine adverse event analysis)
Data (e.g., ATCC text data, VAERS text or MedDRA-annotated data)
Standardize
Mapping to
Ontologies (e.g., CLO, OAE)
Data and data representation
Basic Bioinformatics research (e.g. gene-interaction network analysis using PubMed NLP)
Reducing confusion in cell line nomenclature: How to eliminate mislabeling?
Understanding post-vaccination adverse events: How to predict/prevent/treat AEs?

(b)

Figure 1.2 Outline diagram of the overall scope of work in the integrative biomedical ontology study.
The advantage of integrative biomedical ontology approach has the center role in bypassing the challenges of multi-format data. Information recorded in various formats (e.g. textual records, patient files, or even portable document format (PDF)) can be processed to a standardized form that can further utilize the (semi-)automatic processing and reasoning with biomedical ontologies. This is useful in the situation where the data pool is growing exponentially especially in the clinical setting. This study explores two domains of question; cell line analysis, and vaccine adverse events (VAEs). Figure 1.2 (a) describes the evolution of terminology and its application in data representation that is used as the guideline for method design in this thesis. While figure 1.2 (b) conceptualizes how biomedical ontology data representation is applied to accommodate the practice of health informatics research in combination with basic bioinformatics research.

Table 1.1 Summary of use cases and scope of biological-driving problems (DBP)

Domain of DBP	Ontology (-ies) and Database(s) used	Application
Data management	CLKB	Systematic cell line entry cataloguing
Data integration	CLKB, CLO	Resolving and annotating cell line derived from another cell line
Knowledge transfer between experimental/clinical sources	CLO, EBI Corriell Cell Lines, BAO	Mapping correspondent cell lines among multiple databases
Data management	VO-derived OAE	Identifying adverse events triggered by vaccines
Data integration	Stand-alone OAE	Recognizing adverse events triggered by medical intervention
Knowledge discovery from mined clinical data	OAE, MedDRA, MP, SNOMED-CT	Identifying differential biological profiling of different groups of influenza vaccines
Knowledge transfer and discovery from mined clinical data	OAE, MiMI	Expanding results from previous discovery to further investigate molecular activities of severe adverse events triggered by influenza vaccines

1.5.1 Motivation

The background given in the previous sessions has emphasized on the emergent need to develop tools and framework that assist the new medicine and account for the large amount of clinical and experimental data that are growing exponentially each year. We propose that this framework development can be accomplished by carefully designing and implementing computational architecture with the mechanism of biomedical ontology. With available domain expertise, we have created the Cell Line Ontology (CLO) and Ontology of Adverse Events (OAE) to showcase the myriad possibilities that can be further developed on the basic unit of single ontologies.

1.5.2 Creation

Among the abundance of biomedical ontologies that are publicly available, not all ontologies are created equal for general purposes as previously discussed in the controversy of ontology session. To create an ontology that is specific to one's use, while conforming with an upper ontology as discussed with the skeleton of the OBO Foundry guideline, we have demonstrated such creation in CLO and OAE as followed.

1.5.2.1 The Cell Line Knowledgebase/ The Cell Line Ontology

The question of how to describe a cell line in a computer-readable format came up at the time of the Cell Type Ontology [31] creation where they were defining cell types and tissues based on anatomical aspect of the cell. CL aims to provide the definition of cells regardless of species. However, when they encountered the issue of *experimentally modified cells*, they eventually came to an agreement to leave cell lines out of their scope of work. We then decided to take the leader role in developing a comprehensive catalogue for cell line entries as there were many challenges with regards to cell culture that can be resolved if these cell line entries were properly recorded into a machine-readable format. This attempt resulted in the Cell Line Knowledgebase (CLKB) and Cell Line Ontology (CLO), which will be discussed in the following chapter.

1.5.2.2 The Ontology of Adverse Events

The next step in proposing that biomedical ontologies can help gear research toward translational informatics is to demonstrate that it can be used in the real world. We explicitly show that, even though information obtained from reports in clinical setting is statistically noisy and difficult to work with, with a thoroughly-planned work flow and the aid of biomedical ontologies, bioinformatics methods can be used to process this noisy information efficiently. The Ontology of Adverse Events (OAE) was created as the extension to the existing Vaccine Ontology (VO). OAE utilizes information retrieved from a public surveillance system the FDA/CDC Vaccine Adverse Event Report System (VAERS), and therefore we were faced with a challenge of working the reality into the ideal setting of bioinformatics. Creating ontologies is one challenge, but creating an ontology so that it can be applied to sensible use is even a bigger question. The application of such ontologies will also be discussed as another focus of this study.

1.5.3 Application

Ontology integration has become one of the major themes of biomedical ontology research in recent years because the concept can be applied to answer the controversial argument of reusing existing ontologies in combination of extending the content to convert to a semi-novel ontology that can accommodate a domain specific question. Here we provide scenario of use cases that lend themselves as a good example in how to effectively use biomedical ontology in translational informatics application

1.5.3.1 Describing cell cultures in experiment

In the work of CLO, we have provided evidence that in creating CLO-specific concepts (cell lines and related concepts) while using information imported from other ontologies, we have succeeded in a collaborated project that can further cover cell line content description in a wide universal range including 1) Array Express – Gene Expression Omnibus record mapping project at the European Bioinformatics Institute, 2) Bioassay Ontology at the University of Miami, 3) Experimentally modified cell lines as an extended branch of the Cell Type Ontology (Jackson Laboratory), 4) Reagent Ontology as part of the Eagle-i Consortium (https://www.eagle-i.org/), and 5) Riken Cell Bank at Riken Institute, Japan. The collaboration has beautifully depicted the framework of using biomedical ontology application in knowledge transfer and information sharing.

1.5.3.2 Clustering of adverse events by OAE structure

Moving onto a more clinical aspect of translational informatics, we have applied OAE to the actual records of vaccine adverse event reports submitted to the Vaccine Adverse Event Report System (VAERS) which is a national vaccine safety surveillance program sponsored by the Centers for Disease Control (CDC) and the Food and Drug Administration (FDA) (http://vaers.hhs.gov/). Main advantage of VAERS lies in the coverage of report that is at national level and therefore has the power to detect time-sensitive signal. However, VAERS also presents the classic challenge in the take-all-mine-all approach that the data obtained are very statistically noisy and thus introducing the challenge in predicting any association of an adverse event to a particular vaccine. In this particular case, we have selected our candidate pool of information for investigation based on two types of Influenza vaccines; live-attenuated versus killed-inactivated. We then processed the selected pool with a combination of bioinformatics methods to associate an unsupervised cluster of adverse event with a particular type of Influenza vaccine. This has led us to predicting a model of gene network that triggers biological activities at molecular level. The overall schema shown in this study can be easily adapted to answer other clinical questions in different domains. The summary and case studies of the idea concepts are outlined in figure 1.1, 1.2 and table 1.1.

1.6 Summary and Conclusion

In this thesis, chapter 2 will describe the construction of the Cell Line Knowledgebase followed by the details of the Cell Line Ontology construction in Chapter 3. Chapter 4 demonstrates the needs of a novel bioinformatics workflow to analyze adverse event clinical data with the Ontology of Adverse Events. The bioinformatics framework utilizing biomedical ontologies showcased in this thesis has opened up new possibilities to transition from the basic research to translational health informatics. The concepts and methodologies described throughout have set the *infrastructure of idea* with the provided case examples of how integrative bioinformatics approach is becoming a new driving trend to complex knowledge discovery. Although this is only the beginning of the modern medicine era, the key idea of integrative methodology can be further developed to accommodate higher level of discovery. For example, as modeled in figure 1.3, both OAE and CLO will play a role of mediator in linking between clinical phenotype to molecular genotype. Ontologies modeled in faded dotted boxes are examples of other ontologies not included in the scope of this thesis that can be integrated to further expand this study to a greater value. In the course of collaboration with multiple institutes, we have come across the initiation of the Kinome/Tyrosine Kinase Inhibitor Ontology that aims to build a robust information holder/structure for kinases with an expedient hierarchical structure. In the design of this kinome ontology, a well thought-of plan would include the information of the anatomical compartment/tissue/cell that each kinase would be acting in, and an experimental cell line corresponding to that kinase entry should there be an existing research studying that kinase in.

Furthermore, high-resolution biomedical ontologies are being developed and improved as a result of more research to support a better understanding of biological mechanism. Example ontologies used in this model include the Gene Ontology (GO), Sequence Ontology (SO), and Protein Ontology (PRO). Data mining with NLP along with statistical integrative ontology method as demonstrated in this study has laid out an example of workflow of how it can be implemented to map across the phenotypic characteristics and intermediate level to the genotypic analysis. Altogether, all the components outlined in this model provide the infrastructure pattern that makes the big picture of the modern medicine.

1.7 Publications

1.7.1 Published

- The Cell Line Ontology and its use in tagging cell line names in biomedical text, AMIA Annual Symposium Proceedings, 2007 Oct 11: 1103 [73]

- A bioinformatics analysis of the cell line nomenclature, Bioinformatics, 2008 Dec 1; 24(23): 2760-6 [7]

- Cell Line Ontology: Redesigning the Cell Line Knowledgebase to aid Integrative Translational Informatics, ICBO Proceedings, 2011 Jul 28-30 [74]

1.7.2 In preparation

- The Cell Line Ontology Consortium: a multi-continent collaboration, journal to be determined

- Differential adverse events induced by killed-inactivated and live-attenuated Influenza vaccines, submission for the Journal of the American Medical Association

- The Ontology of Adverse Events to assist post-vaccination adverse events' gene-interaction modeling with combinatorial bioinformatics approaches, submission for ICBO 2012 vaccine/drug ontology workshop

- A review article on biomedical ontologies: the past, present, and future, journal to be determined

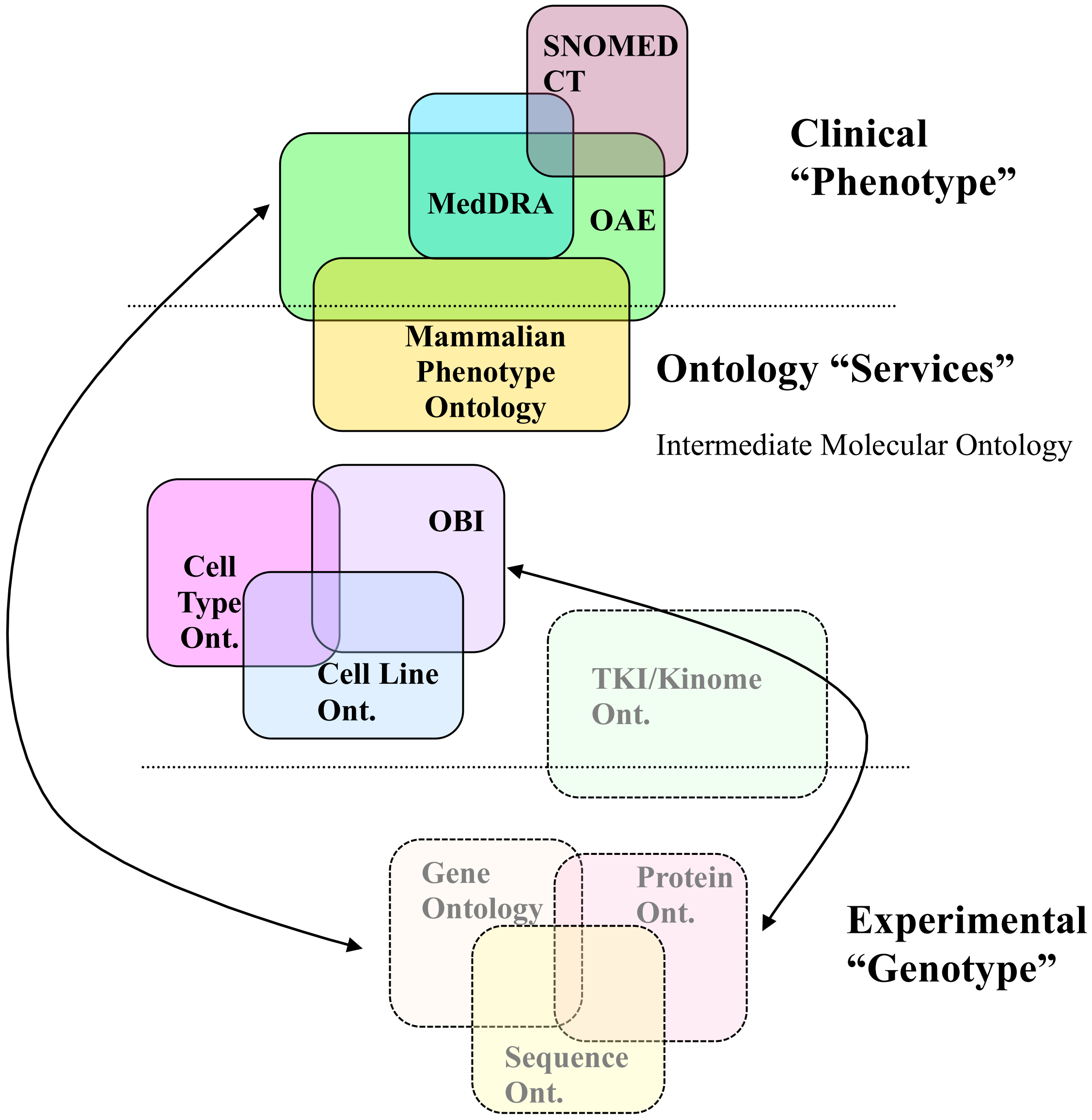


Figure 1.3 The concept diagram of the integrative biomedical ontology approach to translational informatics.
Implementation of linking by importing classes between ontologies expands the knowledge boundary described in a single ontology. Building the bridge to cross from the experimental genotype to the clinical phenotype with this approach is demonstrated with cases of the Cell Line Ontology and Adverse Event Ontology. Overlapping area depicts the shared components (imported classes) among the linked ontologies. Ontologies illustrated in dotted boxes are outside the scope of this thesis (but is discussed in future study).

Chapter 2

The Cell Line Knowledgebase (CLKB)

2.1 The importance of cell line in biomedical research

Cells cultured in vitro are powerful and convenient model systems that are widely used in biomedical research. In the past year alone (2010), some 47,000 papers were published on work using cell lines (Query on: (((cell line[MeSH Terms]) OR cell culture[MeSH Terms]) AND "2010"[Publication Date] : "2010"[Publication Date]) AND "0"[Publication Date] : "3000"[Publication Date]). Often, a great deal of biological information is associated with the particular cell line used for analysis, and knowledge of this context is important to fully understand the implication of a publication. The rapid growth of biotechnology and biomedical research is generating massive collections of free text data that would benefit from improved data organization and management.

2.1.1 Motivation of CLKB development

Cell lines are used extensively in biomedical research, but the nomenclature describing cell lines has not been standardized. The problems are both linguistic and experimental. Many ambiguous cell line names appear in the published literature. Users of the same cell line may refer to it in different ways, and cell lines may mutate or become contaminated without the knowledge of the user. As a first step towards rationalizing this nomenclature, we created a cell line knowledgebase (CLKB) with a well-structured collection of names and descriptive data for cell lines cultured *in vitro*. The objectives of this work are: (i) to assist users in extracting useful information from biomedical text and [75] to highlight the importance of standardizing cell line names in biomedical research. This CLKB contains a broad collection of cell line names compiled from ATCC, Hyper CLDB and MeSH. In addition to names, the knowledgebase specifies relationships between cell lines. We analyze the use of cell line names in biomedical text. Issues include ambiguous names, polymorphisms in the use of names and the fact that some cell line names are also common English words. Linguistic patterns associated with the occurrence of cell line names are analyzed. Applying these patterns to find additional cell line names in the literature identifies only a small number of additional names. Annotation of microarray gene expression studies is used as a test case. The CLKB facilitates data exploration and comparison of different cell lines in support of clinical and experimental research.

2.1.2 Ambiguity introduced by *de facto* naming creation

The names of cell lines grown in culture are frequently generated by the laboratory of origin and have not been subject to systematic organization. As a result, the nomenclature is inconsistent and sometimes ambiguous. For instance, there are numerous examples where the same name is applied to at least two different cell lines and many cases where different names are applied to a single cell line. Improved organization of cell line names would be beneficial to the clinical and experimental research communities. Traditionally, a researcher decides on which cell line he/she will use in addressing the question at hand based on historical knowledge ("what is known to work under the given condition?"). However, in scientific research, the original plan may not work out and different approaches need to be considered. Thus, researchers often need to explore alternative cell lines and to access information about these cell lines in planning their experiments. Repositories of cell lines exist, and names used by these repositories provide a useful starting point for a nomenclature. However, in many cases, the repositories inherit a name from the original developer of a cell line and continue to use it, or they may add catalog numbers or other *de facto* synonyms, further complicating matters. The nature of how cell lines are used in biomedical research leads us to the classic informatics question "How can we transform text data with cell line names into well structured information about these cell lines?" Another issue in information management is quality assurance and quality control, "How do we know if we are looking at an accurate set of data?" Even though there exists an ontology for related information of cell types and other biomedical artifacts[28,76], we are still missing another important component of cell line information; and thus, our attempt to create a CLKB. During the construction of CLKB, we have encountered issues that need to be addressed in the community in order to create structured knowledge about cell lines to answer questions in information-seeking domains. The structure of GENIA ontology also suggests that cell line information from our CLKB can be integrated to create an enhanced network of information[76,77]

2.1.3 Cell line contamination

Cross-contamination is a very serious source of confusion in cell line nomenclature. Cell lines may be contaminated at many steps during propagation and the maintainer of that tissue culture may not be aware of such contamination. Misattribution of cross-contaminated cell lines causes extensive confusion about what a cell line truly is. The widespread contamination with HeLa cells was first recognized by Walter Nelson-Rees using banded marker chromosomes as indicators of intra-species cellular contamination[78]. Cross-contamination of cell cultures has been an on-going issue since[79]. Despite thirty years of effort, cross-contamination remains a problem. A study of contamination in leukemia-lymphoma cell lines has also shown that approximately 15% of these cell lines are not the true representation of what they actually are, or are assumed to be[80,81].

Moreover, even if a cell line is not contaminated, it may not be stable and its character may change from passage to passage. This may cause the birth of a 'new' cell line, sometimes without the knowledge of the users. DNA profiling and cytogenetic analysis are robust methods to identify different cell lines[80,82]. These assays are becoming less expensive, and the community are adopting the protocol as a standard practice for authenticating cell lines as shown in the ATCC Standard Development Organization initiation (ATCC SDO)[83].

2.2 Methods

2.2.1 Cell line cataloguing in public repositories

The CLKB is an initial attempt to normalize the cell line nomenclature. We draw data from Hyper Cell Line Data Base (HyperCLDB) version 4.200201 (http://www.biotect.ist.unige.it/interlab/cldb.html)[84,85,86], and the American Type Culture Collection (ATCC) cell line catalogue (available online at http://www.atcc.org/common/documents/pdf/CellCatalog/CellIndex.pdf as of November

2006), and link names derived from these collections to the National Library of Medicine Medical Subject Headings (NLM MeSH 2007). The focus of this project lies in permanent cell lines, not primary cells. HyperCLDB html files were downloaded and stored in a local storage for data processing. A python script was written to parse out cell line names and their corresponding organism, tissue, pathology, and tumor information, and store the information on our database server. There were 6,609 cell line records including duplicates (cell lines with the same label deposited by multiple laboratories) on HyperCLDB. The collection was then processed to eliminate duplicates keeping only one record for a specific cell line name with reference pointers to the name duplicates. The unique-name version of HyperCLDB contained 5,888 distinct cell lines.

2.2.2 Parsing individual cell line's attribute values

All data were processed in a case-sensitive manner; in the non-standardized cell line annotation across different laboratories, capitalization may signify different representations for different cell lines. Thus we could not normalize capitalization for natural language processing. On the other hand, different capitalization of the same spelling often does not indicate different cell lines. Such variants are treated as synonyms of the cell line name as shown in table 2.1. An advanced algorithm for machine learning would be required in this case. The capitalization inconsistency was manually examined and curated at a later stage, as the machine could not easily distinguish the differences without human intervention (e.g. the case of Hep-2 and HEp-2). ATCC cell lines were converted from PDF format to Microsoft Excel file format. We manipulated the data obtained from ATCC catalogue on Microsoft Excel utilizing VBA macro rather than other programming scripts as some of the ATCC cell line names contained Unicode characters. Implementing text processing using Python or Perl scripting does not usually handle Unicode characters, and converting Unicode encrypting to another workable format adds complexity to calculation (see supplementary note). 3,488 cell lines were extracted from ATCC catalogue.

The attributes for each ATCC cell line were cell line name, ATCC number, species, source/application, morphology, and growth mode. The information stored in association with each ATCC cell line was then processed to a format compatible with the HyperCLDB data set. After text processing, each cell line record has the attributes of CellLineID, Organism, Tissue, Pathology, Growth Mode, Repository Source, ATCC Number, HyperCLDB html, and MeshID. For each attribute used in an individual record, standard nomenclature is applied to normalize textual content for a better organization of the knowledgebase (i.e. we try to use a controlled list of vocabulary to describe definition of terms being used in the knowledgebase; e.g. using NCBI taxonomy to describe organism, or using NCBI MeSH terms where applicable.)

The 8,914 cell line names from both sources were combined in alphanumeric order. Cell lines that appear to share similar names (but different punctuation marks, and/or space) are examined; if they share the same characteristics (same CellLineID, Organism, Tissue, and Pathology), the entries are merged to one primary entry and its cross-reference identifiers are stored in our data table. If a group of names appear to share only the similar names but their characteristics differ, the entries are kept as-is. Table 2.1 provides examples of merged names. Repository source is the name of primary source that the cell line is taken from. If related names for a single cell line exist in both repositories, we use ATCC as the primary repository source since ATCC appears to be dynamically maintained and catalogued by centralized committee. Note that the construction of this CLKB aims to rationalize the nomenclature so that it represents the actual cell lines. At least one of the identifier attributes of the primary repository source (ATCC number, or HyperCLDB html) must be present for each cell line entry. Some cell line names that exist in multiple sources contain the identifier(s) for each source as cross-references. ATCC number is the ATCC catalogue number. HyperCLDB identifiers are the original html file names as taken from the base URL (http://www.biotech.ist.unige.it/cldb/*.html). A few cell lines are also listed in NCBI Medical Subject Headings collection (MeSH). These cell lines also have the corresponding value in their MeSH identifier attributes.

2.3 Results

The structure of this CLKB was constructed using Protégé[87,88]. After manual review, we were left with 8,740 unique-name entries. A Java script was implemented to automate the instance creation in a W3C Web Ontology Language format (OWL-DL file extension).

2.3.1 The Cell Line Knowledgebase (CLKB)

The primary aim of building the CLKB is to facilitate research using cell cultures. Therefore, in addition to the simple underlying ontology structure, CLKB contains information on cell line culture, availability and biological characteristics. We have constructed an online web interface that queries the CLKB. The URL for this knowledgebase is http://clkb.ncibi.org/. This CLKB is a public data warehouse for searching cell line data extracted from ATCC and HyperCLDB as described previously. Constructing this knowledgebase also leads us to the next application of this CLKB, ontology mapping. Often, information can be linked from other sources. For example, the mapping by string literals of attribute 'Tissue' in cell line to the Cell Type Ontology[28]. There are 8,473 cell lines (out of the total 8740 instances in this knowledgebase), each of which has the attribute 'Tissue' that can be mapped to cell type name in Cell Type Ontology using exact-string alignment method. There are 265 cell line instances that do not have tissue information (Tissue = NULL). This leaves us with only 2 cell line instances where their tissue attribute cannot be automatically mapped to a cell type name (SVEC4-10EE2 and SVEC4-10EHR1). Even these two remaining cases could be mapped when extraneous blanks were removed from their tissue attribute values. This preliminary CLKB – Cell Type Ontology mapping clearly demonstrates a promising solution and is a step towards building the 'biomedical knowledge network'.

2.3.2 Data structure

There are two distinct entity types, **CellLine** and **RepositorySource**. A cell line entity contains the attributes as described. A repository entry contains the information of Repository association class, and the URL in which the cell line refers to (html indicator for HyperCLDB instances, and ATCC catalogue PDF for ATCC instances). The CLKB was developed with these processed data as outline in figure 2.1. The majority of cell lines are stand-alone entities, and we include a link to their source of origin.

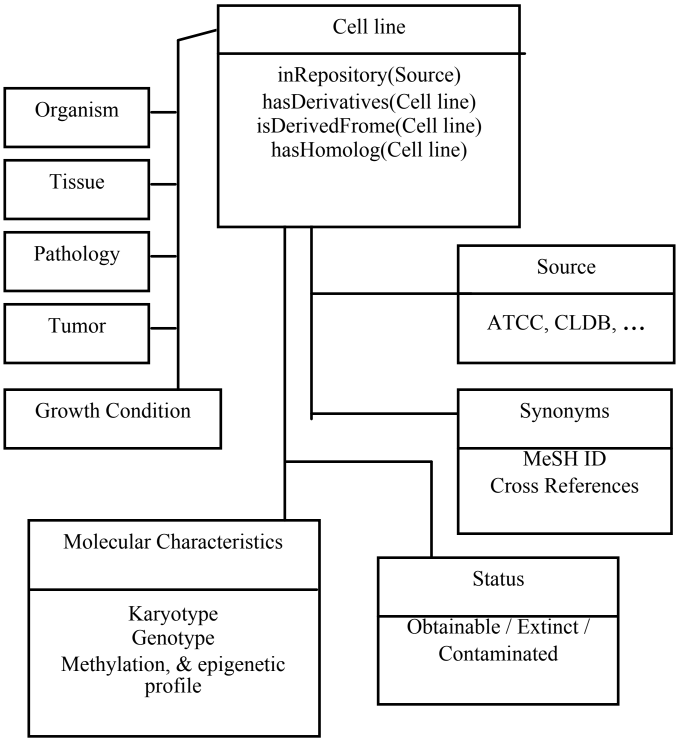
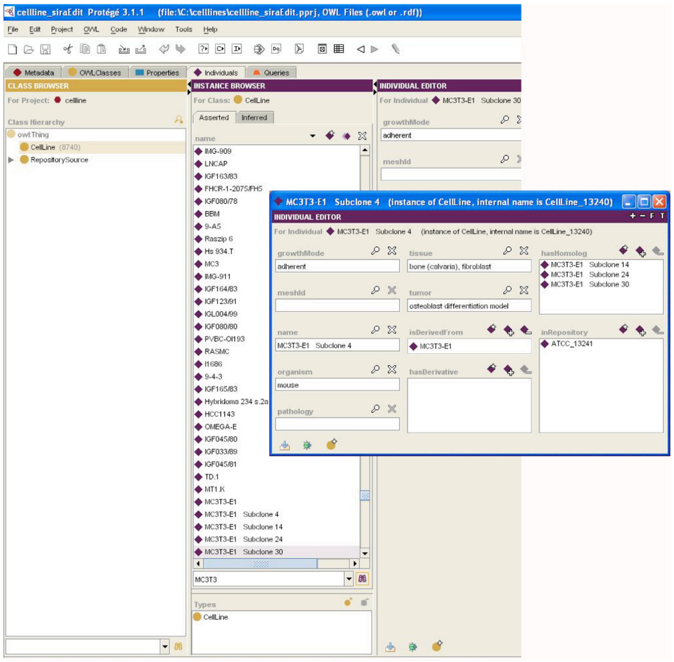


Figure 2.1 UML Diagram describing the entities and relations of the Cell Line Knowledgebase

A number of cell lines are derivatives of some common cell lines. These cell lines contain the **isDerivedFrom** relationship to the parent cell line, and the parent cell line could have one or more **hasDerivatives** relationships as well. Derivatives at the same level are considered **homologs** (for example multiple clones derived from a common parent cell line, thus there must first be a common parent in the dataset). Determination of derivatives was based on explicit information ("subclone") of cell line clones indicated in the cell line name field. Although there may have been some indicators other than this explicit statement for cell line derivatives, we intentionally did not include these rules in our automated script as the common-substring approach could mislead us. We do not want to identify a cell line as a derivative for another cell line when it was not a true derivative (e.g. a substring in a cell line name may be just a manufacturer's abbreviation and a system-generated number). To avoid this foreseen issue, we decided to leave out other clonal implications that were not explicitly stated in the cell line name/label.



cellline_siraEdit Protégé 3.1.1 (file:\C:\celllines\cellline_siraEdit.pprj, OWL Files (.owl or .rdf))

For Project: cellline

Class Hierarchy
- owl:Thing
  - CellLine (8740)
  - RepositorySource

For Class: CellLine

name
- IMG-909
- LNCAP
- IGF163/83
- FHCR-1-2075/FH5
- IGF080/78
- BBM
- 9-A5
- Raszip 6
- Hs 934.T
- MC3
- IMG-911
- IGF164/83
- IGF123/91
- IGL004/99
- IGF080/80
- PVBC-OI193
- RASMC
- I1686
- 9-4-3
- IGF165/83
- Hybridoma 234 s.2a
- HCC1143
- OMEGA-E
- IGF045/80
- IGF033/89
- IGF045/81
- TD.1
- MT1.K
- MC3T3-E1
- MC3T3-E1 Subclone 4
- MC3T3-E1 Subclone 14
- MC3T3-E1 Subclone 24
- MC3T3-E1 Subclone 30

MC3T3

Types: CellLine

For Individual: MC3T3-E1 Subclone 30

growthMode: adherent

meshId:

MC3T3-E1 Subclone 4 (instance of CellLine, internal name is CellLine_13240)

For Individual: MC3T3-E1 Subclone 4 (instance of CellLine, internal name is CellLine_13240)

growthMode: adherent

meshId:

name: MC3T3-E1 Subclone 4

organism: mouse

pathology:

tissue: bone (calvaria), fibroblast

tumor: osteoblast differentiation model

isDerivedFrom: MC3T3-E1

hasDerivative:

hasHomolog: MC3T3-E1 Subclone 14; MC3T3-E1 Subclone 24; MC3T3-E1 Subclone 30

inRepository: ATCC_13241

Figure 2.2 Screen capture of the Protégé interface for CLKB
A user interface of Protégé helps visualize and analyze information of cell lines.

The CLKB can be viewed in most ontology editors that are capable of processing W3C web ontology language file format (.OWL extension). A screenshot from Protégé ontology editor of this knowledgebase is also shown in Figure 2.2. There are likely going to be more novel cell line information available from various sources, we encourage the use of an ontology editor in merging these cell lines into our current version (for example Protégé as we have successfully worked with in this project – http://protege.stanford.edu). The ontology mapping should be relatively easy to accomplish as our CLKB structure was created in such a way that supports an effective and easy-to-understand downstream activities.

2.4 Discussion

2.4.1 Cell Line Ontology

As a first step in developing the knowledgebase, we defined a basic design pattern to later fully be modified and developed to a cell line ontology (Chapter 3). An ontology described and defined in one context may or may not be a consensus to all users. There is no wrong or right ontology as long as it captures what the ontology's users need depending on which perspective the users view the ontology. In this case, we have chosen to simplify the structure of the cell line ontology to be as described in the previous sections as the primary purpose for this type of ontology is to rationalize the cell line nomenclature in experimental and clinical research. However, as we progress through this work, there is a possibility that the cell line ontology's structure can be adapted to assist in standardizing cell line nomenclature and annotation, including the issue of cross-contaminated cell lines that is discussed later in this paper. Classes of cell lines, contaminated cell lines, and homologous cell lines could be drawn for relationships to help solving the issue of cell line nomenclature and annotation.. Since individual cell line names defined in this study as part of the knowledgebase are defined as a class. The structure would also aid other ontology processing at a more-complex level of machine learning with ontology structure.

2.4.2 Cell Line Nomenclature

The compilation of data input for this knowledgebase (American Type Culture – ATCC, and Hyper Cell Line Database – HyperCLDB --- discussed in Method section) reveals a number of cell line names that can potentially be problematic (exemplified in table 2.1).

First, there exist cell lines with identical cell line names, which are, however, distinct cell lines that should be listed separately. For example, the 15C6 cell line is used in more than one distinct ATCC catalogue numbers, CRL-2431 and HB- 326, one referring to a mouse hybridoma line and the other to a rat/mouse line. Second, we have observed inconsistency in the use of cell line annotations between different laboratories when referring to the same cell line. For example, cell lines LLC-MK2, LLCMK2, LLCMK2, all appear to be the same cell line. Other similar cases of NIH/3T3 and NIH-3T3 cells, or BxPC3, BXPC-3, BxPC-3.

Another source of inconsistencies is formatting of text. For example, capitalization, dashes, slashes and other punctuation marks are frequently used inconsistently between different public repositories or even within one repository if the cell line is ob- tained from different depositing laboratories. An NLP approach may be a suitable method to identify these occurrences. Table 2.1 gives some example of one cell line being described with different names. As it appears, many cell line records obtained from HyperCLDB contain different capitalization and punctuation even though they refer back to the same cell line in ATCC. The uncontrolled use of text formatting adds synonyms to the nomenclature and leads to more confusion. Symbols and Greek letters are another source of inconsistency. For example, **Ψ** may be shown as a Greek character in one instance and spelled out as “psi” in another. Unicode presents another issue. When annotating a cell line with Unicode characters (e.g. the use of α, ϒ, ς, or ψ in some cases) different information systems may handle the data differently, in some cases even ignoring such characters altogether. Further, there are visually similar Unicode characters with distinct Unicode encodings. “Ψ” and “ψ” are both renderings of Psi in upper and lower case, but with distinct Unicode encoding. Although they may appear similar to the human eye, an automated script will fail to match them. Fourth, some cell line naes cannot be recognized or distinguished easily when appearing in a different context or even in the same biomedical domain. For example, there exist cell line names that are too common to be easily tagged. Examples include HORSE, OK, WISH, 35, and 81.3. We have identified cell line names that are also common English words by constructing a dictionary of cell line names from HyperCLDB collection and searching against a moderate-size corpus of Wall Street Journal text that is unlikely to include references to cell lines. Table 2.2 shows the top ten HyperCLDB cell line names that occur in a non-biomedical context. Furthermore, we have also investigated the use of cell line names in NCBI GEO microarray description fields. We parsed out 26,109 microarray sample descriptions and tagged the named entities of cell line names based on the HyperCLDB cell line name dictionary that we constructed in the Wall Street Journal corpus experiment as described previously. Some of the occurrences of cell line '35' were false positive as one may have already expected. 'Bm' may as well be an acronym for bone marrow, and not a cell line name at all. Table 2.3 demonstrated the list of tagged cell line names in microarray sample descriptions.

Table 2.1 Example of synonymous cell lines.

CellLineName	**ATCC No.**	**HyperCLDB html**
2HX-2		cl51.html
2Hx-2	HB-8117	
34-5-8 S		cl5138.html
34-5-8S	HB-102	
3T3 L1		cl71.html
3T3-L1		cl72.html

3T3 Swiss Albino 3T3-Swiss albino Swiss-3T3	CCL-92	cl83.html cl4451.html
3T6 3T6 Swiss Albino 3T6-Swiss albino	CCL-96	cl86.html cl89.html
4/4 R.M.-4 4/4 RM-4	CCL-216	cl91.html
72 A1 72A1	HB-168	cl5143.html
7C subscript(2) C subscript(5) C subscript(12) 7C2C5C12	HB-8678	cl5111.html
BALB 3T3 clone A31 BALB/3T3 clone A31	CCL-163	cl386.html

Table 2.2 Top ten HyperCLDB cell line names that appeared in non-biomedical text corpus ranked by number of occurrences.

Cell Line Name	Count
M4	48
35	44
Aa	11
EC	11
P1	11
380	9
BT	8
L	7
CAR	4
OK	4

Table 2.3 HyperCLDB cell line names that appeared in NCBI GEO Microarray sample description, ranked by number of occurrences.

Cell Line Names	Count	False Positives

M9	122	All
F2	120	All
Bm	72	All
IMR-90	41	
MS	38	All
35	35	All
A549	31	
C2C12	30	
FRT	29	All
E2	28	All
NMU	24	All
CCRF-CEM	19	
HCT-8	18	
MDA-MB-231	17	
SC	16	All
HL-60	14	
P4	14	All
Aa	12	All
NIH 3T3	11	
N1E-115	9	
A673	9	
E3	8	All
MCF-7	8	
F1	8	All
D2	8	All
MCF7	8	
697	7	

Note that, only a few of the known cell line names appear in the sample description, and some are more frequently used than the others. This may be due to the fact that many GEO samples contain null description field, or very short phrases. Further investigation reveals that *tissueOrCelllineName* field in GEO sample attributes has been left blank (NULL) in the majority of the data deposited in GEO database. Some GEO samples have information of tissue of origin, and only a small part contain the cell line information.

Furthermore, hardly any values in sampleNameInExperiment from the GEO sample attributes can be a useful pointer to cell line information of the sample used in that experiment. However, some GEO series descriptions contain information that a rule-based NLP implementation may be able to extract useful cell line information from them. A script implementing NLP was written to demonstrate that a simple rule-based NLP approach could help eliminate some common false hits and gain some information of cell line name in text scanning. We tagged cell line names that appeared in 8,091 GEO sample descriptions, 1,059 GEO series descriptions, and 5,187,422 PubMed sentences containing the word ‘cell’ or ‘cells’ (University of Illinois at Urbana Champaign sentence splitter (UIUC, 2004) was used to create this PubMed sentence table).

A comparison of the top ten most common name between tagging any occurrences of cell line names and tagging cell line names with the rules of context where it appeared is given in table 2.4, 2.5 for name tagging in sample descriptions and series description, and table 2.6 for top twenty name tagging in PubMed sentences. The complete tables are given in supplementary materials. The rule-based procedure tagged only cell line names (taken from the existing cell line name dictionary) that preceded ‘cell’, or ‘cells’ tokens, or followed ‘cell line’ token. Also, we used the spaced-tagging strategy (“_% cells|cell_”) to avoid false hits in tagging short cell line names (2 characters long or shorter). An obvious example in this case is a single-letter cell line name like “L cells”; a non-spaced tagging will result in false positives in phrases like “… small cells”, or “… epithelial cells”. The non- spaced tagging (“% cells|cell”) was used for cell line names that are longer than 2 charactors as cell line names can appear at the beginning of a sentence. Another important point regarding rule-based tagging concerns the use of textual qualifiers. There may be other qualifiers for such tagging that researchers use in free text. Further, authors sometimes drop the qualifiers and use only the cell line name token in other sections of a document.

Table 2.4 Ten most common cell line names in GEO microarray sample description demonstrating trade-off of recall and specificity.

Cell line name	# with Rules	# without Rules
L	0	5689
G	0	4213
FR	0	2968
ST	0	2821
U	0	2459
FO	0	2312
NE	0	2133
TE	0	1914
ME	0	1768
MO	0	1243

Table 2.5 Ten most common cell line names in GEO microarray series description demonstrating trade-off of recall and specificity.

Cell line name	# with rules	# without rules
L	0	855
G	0	763
NE	0	504
ST	0	394
U	0	381

FR	0	352
FO	0	331
ME	0	323
TE	0	247
MO	0	226

Table 2.6 Twenty most common cell line names in PubMed sentences containing 'cell' or 'cells' tokens demonstrating trade-off of recalls when increasing specificity. With a large dataset, increasing specificity leads to optimal recalls.

Cell Line Name	# with Rules	# without Rules
L	10401	5187367
U	177	4865827
G	1492	4393413
TE	600	4020437
ST	354	3056536
NE	1352	2815441
EC	3313	2539329
LI	29	2537483
ME	205	2208233
MO	102	1893124
FO	49	1482950
LAT	10	1244902
LT	37	1082924
FR	41	920858
RS	968	895430
YT	491	802212
SC	376	755576
FER	2	751252
TUR	103	652539
HEL	2263	479636

Table 2.7 Examples of cross-contaminated cell lines.

Cell line (Cell type)	Described as	Reference

HeLa (cervical adenocarcinoma)	2563, MAC-21 (lung lymphoma)	Nelson-Rees et al. (1981)
	ADLC-5M2 (lung carsinoma)	MacLeod et al. (1999)
	AO (amnion)	Nelson-Rees & Flandermeyer (1976)
	BCC1/KMC (basal cell carcinoma)	MacLeod et al. (1999)
	BrCa 5 (breast carcinoma)	Nelson-rees et al. (1981)
	CaOV (ovarian carcinoma)	Nelson-Rees & Flandermeyer (1976)
	Chang liver (liver)	Nelson-Rees & Flandermeyer (1976)
	Wong-Kilbourne (conjuntiva)	Nelson-Rees & Flandermeyer (1976)
T-24 (bladder carcinoma)	ECV-304 (normal endothelium)	Dirks et al. (1999)
	GHV (astrocytoma)	MacLeod et al. (1999)
	HAG (adenomatoid goiter)	MacLeod et al. (1999)
	RAMAK-1 (muscle synovium)	MacLeod et al. (1999)

However, we have discovered from this experiment that, in the case of wildly used cell lines, authors seem to conform well with our rules/patterns of “cell” or “cells” token. HeLa cell line was used as an example. Our scripts recalled 22,431 documents that contained “HeLa” (and its spelling variants) tokens. In these documents, there are 43,255 sentences containing the word “HeLa” and its variants. 34,255 sentences out of these 43,255 sentences are sentences where the cell line tokens were tagged in the context of “% hela cell%”. Therefore, 79.19% of sentences tagged with the cell line token were correctly identified as true positives when using rule-based method. Furthermore, when we looked at documents that contained multiple sentences mentioning HeLa and compared to the documents in which only a single sentence was found tagged with HeLa, there are smaller number of documents in the multiple-sentence set (8,558 documents) than the single-sentence set (10,026 documents). Among documents with multiple sentences that co-occur with ‘hela’, looking for an instance that matches the rule “% hela cell%” in one sentence and then assuming that all instances of ‘hela’ in that document were references to an actual cell line found additional 3,196 sentences in 1,215 documents (these 1,215 documents are not a subset of the 8,558 multiple-sentence documents). It should also be noted that while we gain better precision with this rule-based strategy, there is also a trade-off in losing the overall recall. Even though this rule-based NLP approach remains effective in achieving high recall at larger-scale text scanning, further study of finding an optimal adjustment may still be required for a smaller dataset.

2.4.3 Discovery of uncatalogued cell line names and synonyms

As there are often newly-created cell lines in research laboratories, one may wish to utilize a set of existing cell line names for machine learning of biomedical terms using a similar approach to the construction of GENIA ontology (Lee et al., 2004) to recognize novel cell line names that have not been submitted to a repository. This leads to a question that whether or not one can derive a contextual format in which a cell line name may occur. A BayesNet classifier model using Weka (Frank et al., 2004) was introduced to this project in attempt to identify and discover the potentials token of novel cell line names in literature (here, we conducted our experiment with PubMed and NCBI GEO sample description sentences). One generalized observation is that tokens ‘cell’ or ‘cell line’ are often found in co-occurrence with an existing cell line name. The classifier could successfully identify the named entities to be of either cell-type or cell-line classes based on the information from MeSH identifiers under A11 Tissue sub-tree, and the known cell line names in our dictionary. BayesNet classification turned out to be a powerful method to build this classifier. However, over-generalized names and the other natural language processing issues (as described previously) introduce a real obstacle for such discovery because they generate a large number of false matches in text scanning.

We also took another approach to distinguish cell line names from other named entities in attempt to identify novel cell line names. As standard nomenclature, we know that capitalization signifies differentiation between DNA name and protein name. We have also observed the cell line names in parentheses, as many cell line names are acronyms of cell cultures. For example, we have seen the use of **C**hinese **H**amster **O**vary cell culture in many places in literature where the token ‘CHO’ appears in parentheses, as do acronyms of other cell line names. With the observation of short tokens in parentheses, we can narrow down the search space of potential cell line named entities. And with capitalization and numerical character pattern, we could assume that if a token is constructed in the way that it begins with digits that is followed by upper case letter, it may potentially be a novel cell line name. ‘293T’ token comes up in our result as a novel cell line name when using this rule-based NLP named entity tagging (there is not a cell line named ‘293T’ in either ATCC catalogue (the closest name is 293T/17), or in HyperCLDB listing (the closest name is 293)). Out of 46,976 tokens found in the context of “% cell%” ranked by number of occurrences, ‘293T’ comes up at 47th place with 2,809 counts, 2,797 occurrences with ‘293T’ spelling, and 12 occurrences with ‘293t’ spelling. When checking with tokens that appear in parentheses, out of 51,216 tagged tokens ranked by frequency of occurrences, ‘293T’ comes up at 174th place with 2,811 counts, 2,799 with ‘293T’ spelling, and 12 with ‘293t’ spelling. As confirmed by experimental biologists, 293T is a bona fide cell line. Our ability to recognize this as a cell line name based on lexical context suggests a promising direction for further investigation of novel cell line name discovery in free text.

2.5 Conclusion

A recent study of dictionary-based named entity tagging of protein name[75] reveals that, despite the standardized HGNC, issues of ambiguity remain in biomedical applications of automated NLP. Our study shows that there are many other categories of information in the biomedical domain where there is also a need to eliminate ambiguity wherever possible. This includes not only the agreement on a standard nomenclature with unique and distinctive names would greatly facilitate text interpretation, but also the elimination of cross-contamination in cell lines. As a first step toward standardizing the nomenclature for cell lines, we present in the results of this research in supplementary material – cell line token analysis. We also propose that a standard protocol of the minimal set of information including molecular characteristics of cell line should be complied at the development and at repository level as well.

The mapping of cell line names to word tokens in GEO sample descriptions demonstrates that this ontology can be very useful in data quality assurance and control. One may wish to cross check, for example, whether or not the organism labeled for each

GEO sample really matches its corresponding organism in the cell line of that sample that is was grown in. With the issues described here and how much automated NLP applications can potentially accomplish, one cannot stress enough the importance of standardization of biomedical terms. Furthermore, it is to our surprise to find out that every so often, people do not cite where they obtain cell cultures in their research. Without an explicit explanation from the author, it is not practical to assume the source of cell cultures. The current lack of standardized nomenclature in many areas will be an obstacle to the development of effective natural language processing for the interpretation and analysis free-text biomedical information.

The issue of cross-contaminated cell lines remains serious and has been stressed with the NIH notice regarding authentication of cultured cell lines (NIH, 2007).

Chapter 3

The Cell Line Ontology (CLO)

3.1 Call for the Cell Line Ontology: in the wake of biomedical ontology

Cell lines have been widely used in research. Information about cell lines is stored in public repositories and/or indexed catalogues available for open access, and cell lines are commercially available or they are transferred between academic laboratories. Information about cell lines has not been well standardized and machine-readable to date. Each commercial provider generates a catalogue, and academic cell lines are not necessarily included. Integration of data from multiple sources is confounded by: lack of consistent naming conventions for cell lines across providers, contamination of cell lines as they are passaged and transferred between laboratories, and provision of the same cell lines by multiple commercial sources but with different biological attributes.

To address these issues, we previously produced a normalized catalogue of the Cell Line Knowledgebase (Chapter 2 - CLKB; http://clkb.ncibi.org/) as a project in the National Center for Integrative Biomedical Informatics (NCIBI)[7]. Since the release of CLKB, biomedical research has rapidly evolved toward integrative translational bioinformatics. In order to support translational research, conform to OBO foundry standards, and produce a resource that can be used in queries and data integration we have transformed the CLKB into an ontology available in OWL format (http://www.w3.org/TR/owl-guide/). Here we present the design patterns, design methodology, and content of the Cell Line Ontology – CLO.

When the Cell Type Ontology (CL) was first introduced to represent *in vivo* cell types[28], primary and permanent cell lines were included in the ontology and no separate cell line ontology existed. The Cell Type Ontology no longer includes primary or permanent cell lines as the CLO has now become the source ontology for permanent cell lines as agreed by the maintainers of the CL, the Ontology for Biomedical Investigations (OBI), and OBO Foundry. The top-level terminology required for generating a primary cell line is provided by the OBI. The CLO is therefore a collaborative development between the CL, OBI and the CLKB developers at NCIBI and references terms from these and other ontologies in the definitions and modeling of cell lines.

In addition to CLO design and methodology, we also include examples and applications of the CLO in this study.

3.2 Methods: developing CLO

3.2.1 Cell line data source

The CLO uses data from multiple sources, which are described in Table 3.1. The CLO cell line data were first drawn from CLKB entries, which consist of 8740 cell lines stored in ATCC (http://www.atcc.org/) and HyperCLDB (http://bioinformatics. istge.it/cldb/). CLKB will be kept as a backup source but will become obsolete at the release of the new CLO. Additional 27,000 permanent cell lines are obtained from European Bioinformatics

Institute Coriell Catalogue Ontology that models cell lines from the Coriell cell repository (http://ccr.coriell.org/), and cell lines (both primary and permanent) provided by the Bioassay Ontology (BAO; http://bioassayontology.org/) development team. Cell lines that are listed in multiple repositories contain cross-reference pointers to these repositories. Cell line names can be misleading. Similar or synonymous names do not guarantee identical cell lines. Automatic mapping and manual annotation have been combined to ensure correct cell line annotation in CLO.

Table 3.1 Summary of ontology terms in CLO and source ontologies used in CLO.

Ontology	Classes	Object Properties	Datatype Properties	*Total*
CLO (Cell Line Ontology) specific	36879	14	0	36893
Imported full ontologies				
BFO (Basic Formal Ontology)	39	0	0	39
RO (Relation Ontology)	6	25	0	31
IAO (Information Artifact Ontology)	102	14	5	121
Imported terms from other external ontologies				
OBI (Ontology for Biomedical Investigation)	15	6	0	21
CL (Cell Type Ontology)	194	0	0	194
UBERON	622	34	0	656
NCBITaxon (NCBI Taxonomy)	217	0	0	217
Total	38074	93	5	38172

3.2.2 Importing external concepts – reusing existing ontologies

CLO imports the whole Basic Formal Ontology (BFO)[89] as its upper level ontology and the Relation Ontology (RO)[66] as its core relations. The use of these ontologies promotes integration as these resources are used by many biomedical ontologies. We used OntoFox[90] - a technology for merging ontologies to integrate external ontologies such as NCBI_Taxon and Cell Type Ontology into the CLO. All namespaces are preserved for these ontology terms.

3.2.3 Defining and annotating CLO-specific ontology terms

All cell lines and cell line-specific terms are given unified CLO IDs. The cell line data from the Coriell Cell Line ontology (http://bioportal.bioontology.org/ontologies/45331) have been merged to CLO with newly assigned CLO IDs. The BioAssay Ontology [91] has also provided a list of cell lines for inclusion in CLO. In these two cases a namespace is not preserved. When a cell line term is imported from the Coriell Cell Line ontology, we have provided a cross reference to the ontology using the *seeAlso* annotation property. Using the annotation property *comment*, BAO is noted as the source for those cell lines coming from BAO. A cell line design pattern is developed to make generic pattern between CLO cell lines and other ontology terms.

3.2.4 CLO editing and access

The development of CLO follows the OBO Foundry principles[50]. Specifically, we use unique IDs, and provide text definition for each cell line. The Web Ontology Language (OWL) is used as the default CLO format. CLO is edited using Protégé 4 Ontology Editor (http://protege.stanford.edu). The latest CLO is available for public view and download at http://sourceforge.net/projects/clo-ontology/. The latest version of CLO is also available for visualization and download from NCBO BioPortal: http://purl.bioontology.org/ontology/CLO.

3.3 Results

3.3.1 CLO-top level structure and statistics

The key top level classes in CLO are shown in Figure 3.1. To support data integration and automated reasoning, CLO imports many terms from existing ontologies as upper level terms (e.g., *material_entity* from BFO) or terms needed for association (e.g., *cell* in CL). Cell line-specific terms are assigned with CLO IDs (Figure. 3.1). The CLO-specific class *cell line* is the parent class for all specific cell lines in the CLO. The classes *permanent cell line* and *primary cell line* are the major differentia based on culture for cell lines in the CLO at present. The majority of cell lines in the CLO are permanent cell lines. Normalized cell line entries are entered as asserted CLO classes under these two subclasses. A cell line can be cultured or modified, and supplied or managed by a cell line repository (e.g., ATCC) (Figure 3.1). The detailed relations among these terms are described in our cell line design pattern (Figure 3.2). Currently CLO contains 8797 cell line-specific terms with unique CLO identifiers. In total, CLO contains 38172 terms (Table3. 1). The Coriell cell line records were integrated and assigned CLO-specific identifiers.

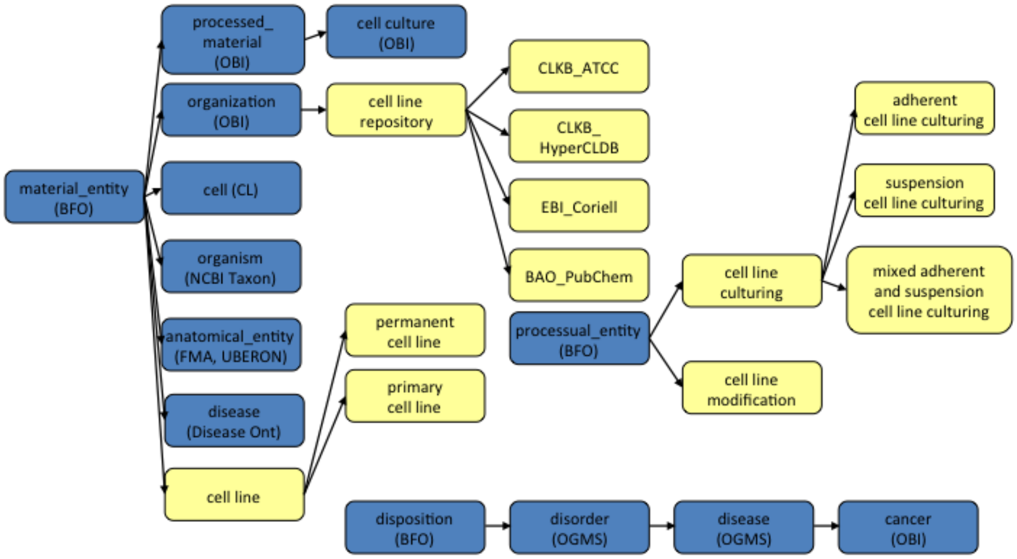


Figure 3.1 The top level CLO hierarchical structure of ontology terms. The terms in light blue boxes are imported from existing ontologies. The terms shown in light yellow boxes are terms with CLO unique IDs.

3.3.2 CLO design pattern

The CLO design pattern supports representation of anatomy, cell types, disease and pathology, source information in the form of ownership and derivation where cell lines are related, and technical information such as culture conditions. Figure 3.2 depicts the design pattern developed to model this information retrieved from data sources. Briefly, a cell line is originally derived from a cell type that is part of an anatomical part (e.g., liver) of a specific organism (e.g., human) having a cancer (e.g., lymphoma) or some other disease. A cell line can be derived from another cell line through a particular cell line modification. A cell line is cultured differently (e.g., *suspension cell line culturing*), which reflects a particular culturing condition or growth mode (e.g., suspension). A cell line is supplied, owned, or managed by a specific organization such as *ATCC* that has a *cell line repository role*. Since relation terms such as *supply*, *own*, or *manage* do not exist in any ontology, we use a similar relation, *mentions* (Figure. 3.2).

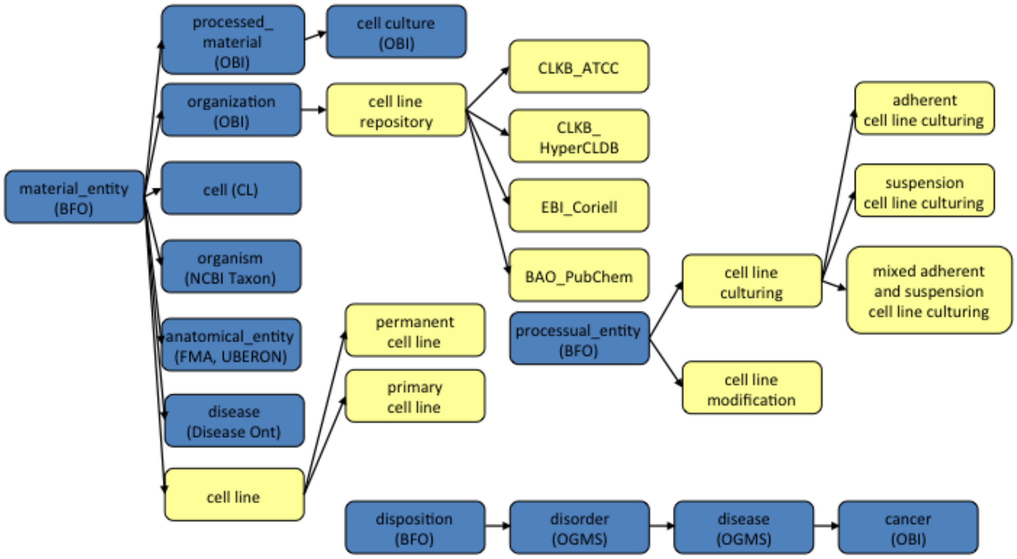


Figure 3.2 Basic design pattern for representing cell lines in CLO.
Components shown in yellow boxes are specific to CLO, while those in blue boxes signify classes imported from other ontologies. Depending on the cell line being described, *suspension cell line culturing* and *ATCC* can be replaced with another cell line culturing process and another cell line repository, accordingly to its associated attributes respectively.

The basic cell line design pattern is followed in our CLO development. In many cases, we have also extended this design pattern by adding more content. For example, we may add a sex (female or male) quality to the organism. The pathology of the cell lines in Coriell represents one of the most important aspects to users of these artifacts and many are used as models for a particular disease. A cell line may derive from an organism that has a specific disease (Figure 3.2). The *is_model_for* relation is used to link a disease to a cell line. This relation has been created as a shortcut relation to represent the association between a cell line and a disease. In the case that a cell line derives from a normal tissue, the information of disease is omitted.

A deep understanding of the cell line design pattern that portrays a true composite architecture of cell lines requires more discussion and explanation on the relationship between CLO, CL and NCBI Taxonomy. More information is provided below.

A cell line is derived from a cell type (Figure 3.2). The CL developers have been working with the CLO to ensure adequate representation of cell types from which cell lines originate. This allows mappings between the CLO and the CL using the *derives_from* relationship. Such integration also promotes error detection. To enhance interoperability with other OBO Foundry ontologies, CL-CLO mapping associates cell-types with anatomical structures using the species-neutral UBERON ontology. Thus, mappings between CLO and CL allow for associations from cell lines to anatomical structures. Sometimes a cell line cannot be mapped directly to CL as the cell line may contain multiple cell types, which can be a case of anatomical part + cell type (e.g., HCC cell line is annotated as having tissue type '*mammary gland, epithelial'*), or cell type + pathological description (e.g., AtT-20 cell line is annotated as having tissue type *pituitary tumor, small, rounded*), or multiple cell types (e.g., p53NiS1 cell line has annotated tissue type *fibrous histocytoma, fibroblast*). In this case, this cell line is related to all associated cell types using the same *derives_from* relation. According to the original repository, a cell line may derive from a cell type named by its associated anatomical part (e.g., *liver cell, peripheral blood*). These anatomy associated cell type terms have been added to CL as new CL terms to support this design. In total, 194 CL terms and 656 UBERON terms are imported to CLO (Table 3.1) and CLO development has expanded the CL.

The NCBI Taxonomy is the source ontology for CLO to import organism information associated with individual cell lines (Table 3.1). A cell line may be listed as a hybrid from multiple organisms and therefore organism and not species is modeled. In this situation, the cell line will be linked to multiple organisms. One exception of this mapping occurs when a cell line is recorded as being part of mouse/rat hybrid as there exists a class named *Mus musculus x Rattus norvegicus* as a special class of the taxonomy. Investigation of the NCBI Taxonomy also reveals that a few classes relating to those of a cell line have a place holder within NCBI Taxon, such as *mouse/rat hybrid cell lines being* classified under parent term *unclassified Muridae*. However, since the primary purpose of importing NCBI Taxon terms to CLO is to use the information to define organism classes, and not to redefine cell lines, we do not import these terms to CLO. A few organism values that could not be mapped to NCBI_Taxon appeared to be the result of typographical errors or spelling variants. For example, there are a few cell line entries with annotated organism ‘Agrothis segetum’, which is believed to be a spelling variation of ‘Agrotis segetum’ (NCBITaxon: 47767). We do not omit or modify these original values, keeping ‘Agrothis segetum’ as obtained from the source (e.g., ATCC), and putting a remark in the cell line class’ comment with the information pointing to NCBITaxon: 47767.

3.3.3 Use case: describing cell lines with CLO – example of Jurkat

We have modeled the Jurkat Clone E6-1 cell line (ATCC # TIB-152) and its derived cell line J.Cam1.6 (ATCC #CRL-2063) as a demonstration of our cell line design pattern usage (Figure 3.3). Jurkat Clone E6-1 is a clone of an immortalized line (Jurkat) of T lymphocyte cells that was established in the late 1970s from the peripheral blood of a 14 year old boy with T cell leukemia[49]. The J.CaM1.6 cell line is a derivative mutant of Jurkat E6-1 by treatment with ethylmethanesulfonate J.CaM1.6 cells are deficient in Lck kinase activity and miss exon 7 in their lck mRNA.

It is noted that J.CaM1.6 cell line is not a child term of Jurkat Clone E6-1 cell line in CLO. Cell lines derived from one base cell line (e.g., J.CaM1.6 cell line deriving from Jurkat Clone E6-1) is by definition not an *is_a* relation to the base cell line but rather a *derives_from* relation. Based on this *derives_from* relation, we generate a term *Jurkat derivative cell line*, and J.CaM1.6 cell line can be inferred to be a *Jurkat derivative cell line*.

The sharing of tissue, tumor, and organism can be used to group different cell lines, such as Jurkat and Jurkat Clone E6-1. The original value *'peripheral blood'* obtained from source is mapped to anatomical term *'blood'* that best fits this term mapping as there are no such terms in FMA or UBERON that describe peripheral blood. Furthermore, a cell line deriving from T cell such as Jurkat is potentially problematic as T cells scatter throughout the body. Not all T cells are *part_of* some blood. Jurkat was extracted from an instance of lymphoma that was in blood. But CL's definition of lymphocytes does not restrict all lymphocytes to blood tissue. This information is however described by *'isolation'* (an OBI term used to associate tissue and cell type) inside CL. A cell line's description embedded in CLO is specific to each individual cell line being conceptualized in each class. Reasoning with knowledge obtained from CLO and CL can capture this issue of specificity.

Many CLO specific terms (e.g., *peripheral blood cell line*) have been generated. A reasoner can be used to infer what cell lines belong to such CLO terms. Such terms are needed for many applications. For example, the ArrayExpress staff needs to know all the blood-derived cell lines and cell types for a meta-analysis of gene expression data on blood. Without such defined classes, it is difficult to obtain the results.

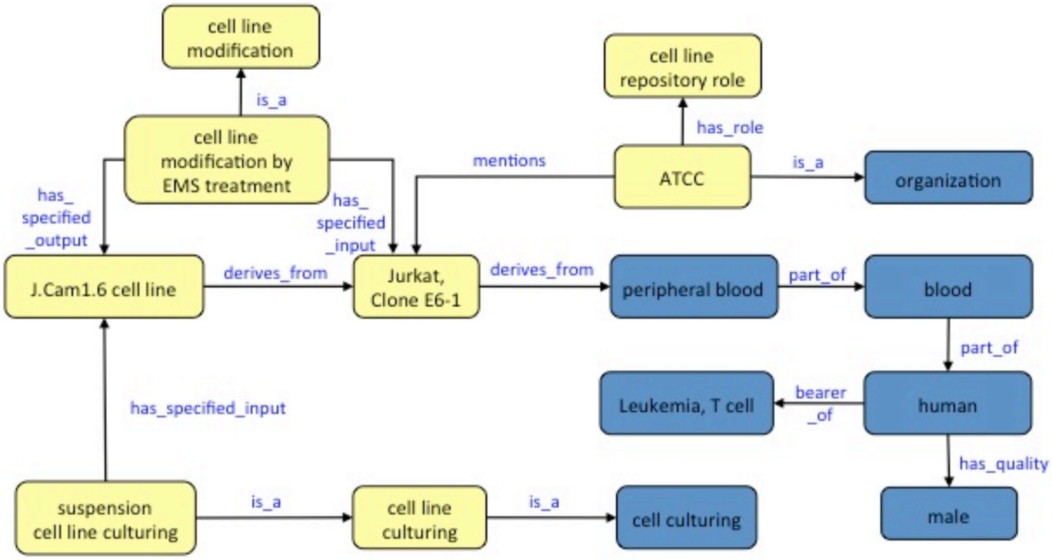


Figure 3.3 Modeling Jurkat and its derivative Jurkat Clone J.Cam1.6 using CLO.

3.3.4 Use case: application of CLO in bioassay data analysis

The Bioassay Ontology (BAO – http://bioassayontology.org/) describes bioassays and results obtained from small molecule perturbations, such as those in the PubChem database [92]. Integrating a formal representation of cell lines will benefit researchers in interpreting and analyzing cell-based screening results. It will also enable linking PubChem assays to other types of information (such as diseases and pathways). Moreover, formally described cell lines can help researchers in the design of novel assays, for example with respect to choosing the best cellular model system, and also in identifying which modified cell lines are available and which ones work best in existing assays.

To describe and annotate cell-based PubChem assays and screening results comprehensively, BAO is being extended through collaborative development of the CLO. By integrating BAO with CLO, those cell lines that are typically used in cellular assays are added into CLO. Based on the demands of BAO bioassay modeling, extended parameters are being added to CLO, including different sources of cell lines (normal/healthy tissue, pathological tissue, or tumor), cell modification methods (plasmid transfection, viral transduction, cell fusion, *etc.*), culture condition (composition of culture medium), morphology (epithelial, lymphoblast, *etc.*), growth properties (adherent or suspension), short tandem repeat (STR) profiling and other properties that are relevant for cellular screening.

As a demonstration of the use of CLO in BAO bioassay modeling, we have modeled the HeLa cell line in the context of a PubChem assay (AID 1611) (Figure 3.4). HeLa is an immortal cell line established from cervical adenocarcinoma of a patient in 1951 [93] and available from the ATCC (catalog # CCl-2). In the PubChem assay, HeLa cells were modified by stable transfection with a heat shock promoter driven-luciferase reporter gene construct. In this assay, the modified HeLa cells were used to screen for compounds that could induce heat shock transcriptional response as a potential therapeutic for Huntington's disease and amyotrophic lateral sclerosis (ALS).

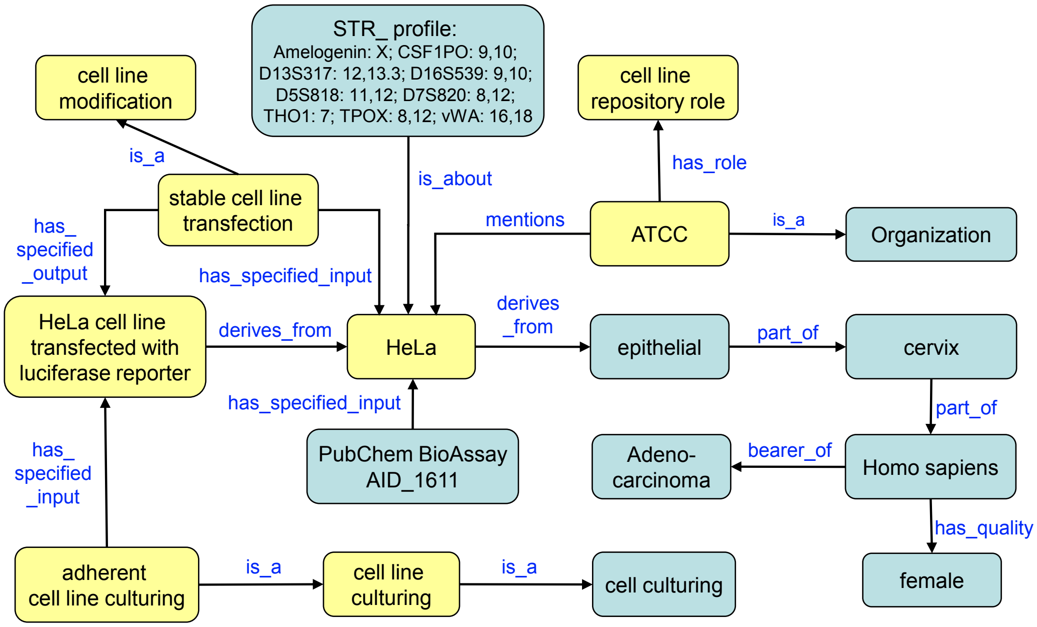


Figure 3.4 Application of CLO in bioassay data integration and analysis.

HeLa as demonstrated in figure 3.4 can be described as a *permanent cell line*, which is 'part_of' cervix and is 'derived_from' *Homo sapiens* that is 'bearer_of' cervical carcinoma. HeLa is an *epithelial cell of cervix* (CL:0002535), whose *growth mode* is *adherent*. Describing the other details of the assay are out of the scope of this paper, as they require concepts from BAO.

Many other CLO applications are being studied. For example, cell line knowledge can be used for microarray data analysis. A separate paper has been submitted to the International Conference on Biomedical Ontology (http://icbo.buffao.edu/) that provides more details of how the Coriell Cell Line Ontology, which has been merged to CLO, is used for ArrayExpress microarray data analysis.

3.4 Discussion

The availability of cellular assays and the ability to sequence DNA and RNA from single cells has promoted the use of cell lines in research and highlighted the role of cell lines in biomedical research. The release of CLO is therefore timely and will support many applications in biomedical informatics. First, CLO can be used as a tool to facilitate the data entry process for public cell line repositories (e.g., ATCC) and the referencing of these by resources such as archival repositories (e.g., ArrayExpress) Cross-referencing with other source ontologies that are imported to CLO will allow a standard controlled vocabulary to be utilized at the data-entry point to avoid typographic errors and aid better annotation, while the depositor can also verify if the cell line being deposited already exists in the ontology, thus eliminating redundant data. Although there are currently no central authorities to assess cell line nomenclature and a cell line name is often assigned by the lab of origin, utilizing the CLO structure to frame the process will help reduce the use of duplicate names for different cell lines. It is our plan to solicit directions of new cell lines to CLO through a community-based agreement. This is also a crucial step to achieve an efficient cell line authentication process.

Furthermore, information stored in CLO can potentially validate other existing cell culture information in various sources. Gene expression data that contain the information of cell line used can be analyzed to observe if there is any data inconsistency when compared back to the information received in CLO based on the same cell line.

Inconsistency in the record's attributes such as organism, tissue, tumor, or genetic mutations based on the cell line's modification may signify the possibility of cross contamination.

Cell line contamination occurs easily. It is reported that 15% of the times cell lines being used are not what they are assumed to be[80] [10]. Contamination also leads to issue of misidentification and mislabelling. To address this issue of contamination and mislabelling and improve cell line authentication, the American Type Cell Culture: Standards Development Organization (ATCC SDO) has proposed to establish a community-supported central authority and to use short tandem repeats (STR) as one method of verification. As a normalized indexed catalogue with ontological structure and semantics, the CLO will play a critical role in standardizing and representing cell lines and properly addressing the issue of cell line contamination. CLO can also be further expanded to link out to this STR verification information of each cell line. CLO is currently being studied for use in the ATCC SDO's authentication process.

Normalized cell line data and additional features in CLO also support applications in translational informatics such as cell line-disease association analysis, annotations of complex organ/tissue in cell cultures, and combined studies of cell culture and bioassay data.

The creation of international BioSamples databases at the EBI (http://www.ebi.ac.uk/biosamples) and NCBI (http://www.ncbi.nlm.nih.gov/ biosample*) provides a strong use case in that storage of non-standardized* data on thousands of cell lines is not useful for high level query purposes and queries such as 'retrieve data on all

ENCODE cell lines' or 'all Drosophila cell lines' will be facilitated by the addition of defined classes to the CLO, and the submission process to archival repositories will be easier if the users are able to query using the CLO ontology to retrieve validated cell line information instead of providing all this information again.

Future work of the CLO development includes the insertion of more cell lines and cell line-associated attributes. Additional CLO applications are under investigation.

Chapter 4

The Ontology of Adverse Events (OAE)

Use of OAE to characterize VAERS data
The comparative analysis of killed influenza vaccines and live influenza vaccine

4.1 Overview

Vaccine adverse events (AEs) are adverse bodily changes occurring after vaccination. Understanding the adverse event (AE) profile is a crucial step to identify serious AEs triggered by influenza vaccines. Two different types of influenza vaccines have been used on the market: trivalent live attenuated influenza vaccine (LAIV) and trivalent (killed) inactivated influenza vaccine (TIV). Different adverse event profiles induced by these two types of influenza vaccines were studied based on the data drawn from the CDC Vaccine Adverse Event Report System (VAERS). Extracted from VAERS are 37,621 AE reports for four TIVs (Afluria, Fluarix, Fluvirin, and Fluzone) and 3,707 AE reports for the only LAIV (FluMist). Our comprehensive statistical method includes Proportional Reporting Ratio (PRR) measure, Chi-square significance test, and base level filtration. In total 48 TIV-enriched and 68 LAIV-enriched AEs were identified (PRR > 2, Chi-square score > 4, and number of cases > 0.2% of total reports). TIV-enriched AEs clustered in neurological and muscular processing. In contrast, LAIV was found to be associated with AEs in the areas of inflammatory response and respiratory system disorders. Severe adverse events such as Guillain-Barre Syndrome and paralysis were at low incidence rate but were found to be higher enriched in TIV. It was then suggested that LAIV had lower chance of inducing these two severe adverse events.

4.2 Introduction

Vaccination is a standard protocol practiced in the public health domain. However, patients may react to the administered vaccine, resulting in adverse changes in health. These changes or side effects are commonly referred to as an adverse event (AE) [94]. Note that AE stands for adverse *event* throughout this thesis because the data used in this study asserted no causality in the AE reports, and therefore, it is not to be confused with adverse *effect*. Some AEs can be serious and fatal. To monitor post-marketing adverse events associated with released vaccines, the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA) have established the Vaccine *Adverse Event* Report System (VAERS) surveillance program[95]. The primary strength of VAERS lies in the national coverage of its reporting network that it can pick up a rare incident of an adverse event in a timely manner. VAERS has been used in many studies that resulted in useful insights of post-vaccination incidents[96,97,98,99,100].

Since VAERS records contain high-noise data, a well-defined method is necessary to analyze VAERS entries. The VAERS reporting protocol is a *passive surveillance* system that accepts data from any reporter, and explicitly does not take in the consideration of causal relationship between the vaccine and adverse event reported to the system. Although each case report is curated with individual adverse events manually assigned to MedDRA codes by VAERS personnel, VAERS data entry process still introduces strong biases caused by reporting efficiency (over- or under-reporting), reporters’ inability to assess causality, and temporal association of the report (e.g. inconclusive symptoms get inserted as post-vaccination AE). Incidence rates and relative risks of specific adverse events cannot be calculated by processing the raw VAERS data without a careful analysis with further combinatorial methods[101]. Current methods of analyzing AE data include the Proportional Reporting Rate Ratio (PRR)[102,103] and Chi-square[104], and Bayesian network approach[105].

There has been proof of plausibility of utilizing monitored post-release drug adverse event data with combinatorial bioinformatics methods in analyzing the occurrence of serious events, for example, the exploitation of the WHO Uppsala center drug safety reports’ pharmacovigilance database using a Bayesian network approach[106],[105] , or the implementation of Proportional Reporting Rate Ratio (PRR) and Chi-square significance test methods in the Medicines Control Agency, and the Drug Safety Research Unit[102],[103],[104]. A recent study of the construction of the Ontology of Adverse Events (previously known as Adverse Event Ontology (AEO)[107]) has addressed the issue of information structure of standardized vocabulary.

VAERS utilizes the Medical Dictionary for Regulatory Activities (MedDRA) system as a coding vocabulary nomenclature. MedDRA has been widely used by physicians and health care researchers in annotating AE information. Therefore, it has played a central role in standardizing vocabulary in the scope of AE reporting. While it was an improvement in the effort of creating an AE nomenclature, MedDRA had several issues in domain completeness and discrepancies with a physician’s AE description that results from MedDRA’s lack of hierarchical structure[108]. To best utilize VAERS content that was transcribed with MedDRA, we introduced the Ontology of Adverse Events (OAE) to the study in an attempt to reorganize MedDRA terms into a logical hierarchical structure based on pathology of the AE symptoms. We tried to keep the original MedDRA term without editing the term as much as possible to avoid any confusion that may arise.

Pandemic influenza has been the center of attention in recent years. It is an illness common but fatal enough that the CDC's Advisory Committee on Immunization Practices (ACIP) recommends that everyone of 6 months of age or older should receive influenza vaccine every year (http://www.cdc.gov/mmwr/preview/mmwrhtml/rr59e0729a1.htm). However, it is possible that vaccine recipients may develop post-vaccination reaction (adverse event), and this possibility should be made aware of when considering influenza vaccine, especially when there were reports of severe adverse events such as Guillain-Barre Syndrome (GBS), and paralysis. There are also a few types of seasonal influenza vaccines available in the market. Assessing adverse events triggered by different influenza vaccines leads to understanding vaccine safety.

In this study, we hypothesized that killed-inactivated and live-attenuated influenza vaccines, two subtypes of trivalent seasonal influenza vaccines, induce different types of adverse events. The rationale behind managing Influenza vaccines into two groups is: 1) They are two different types of vaccines, one is live attenuated, and the other killed inactivated; 2) They both have been widely used after their releases, and have resulted in a significant number of adverse event records reported to VAERS, and 3) Both groups of vaccines aim for protection against Influenza A/B, while having different features of administration methods. TIVs are vaccinated through intramuscular route, and intranasal spray is used for LAIV vaccination. We demonstrated that the workflow of combinatorial bioinformatics methods shown in this study could lead to a novel discovery from VAERS clinic-based adverse event reports. Comparative analysis of adverse event information associated with two groups of Influenza vaccines: Trivalent Killed Inactivated Influenza Vaccine (TIV) and Trivalent Live Attenuated Influenza Vaccine (LAIV), suggests that TIV recipients have an, even though rare, higher chance of developing GBS and other related adverse events such as paralysis and paraesthesia. While LAIV recipients have not shown significant statistical associations with GBS or other related neurological symptoms.

4.3 Methods

4.3.1 Adverse event data extraction

Records of post-vaccination adverse events were queried from the CDC Vaccine Adverse Event Reporting System (VAERS; access date: May 18th, 2011). The query was constructed to retrieve adverse event information of killed inactivated vaccines (TIV group) consisting of *Afluria, Fluarix, Fluvirin,* and *Fluzone*, and live attenuated vaccine (LAIV group) *Flumist*. From the query, 37,621 records were retrieved for TIV, and 3,707 records were retrieved for LAIV. The names and the numbers of AEs for each type of vaccines were summarized. Some AE symptoms are common in both groups, while a significant number of symptoms are unique to each cohort. Symptoms from both AE record tables differ in rankings (number of occurrences) and the nature of adverse events themselves. We hypothesize that performing a comparative analysis of different physiological responses implied by AE symptoms taken by each vaccine would lead us to understand the underlying disease mechanism of the influenza A/B vaccines.

4.3.2 AE report signal detection with Proportional Reporting Ratios (PRRs)

Each group of reports (TIV and LAIV) was analyzed independently with PRR method (as introduced by Evans et al.[102]). PRR calculations to detect signals from the data pool utilize the total number of reports for each vaccine as a denominator to determine the proportion of all reports that fall in the type of interest (which in this case is the individual AE that was retrieved by each group of vaccine). The PRR score of individual AEs in each group is then used as one of the composite criteria to compare for significant AEs in each group.

4.3.3 Chi-square test to identify statistically significant AEs

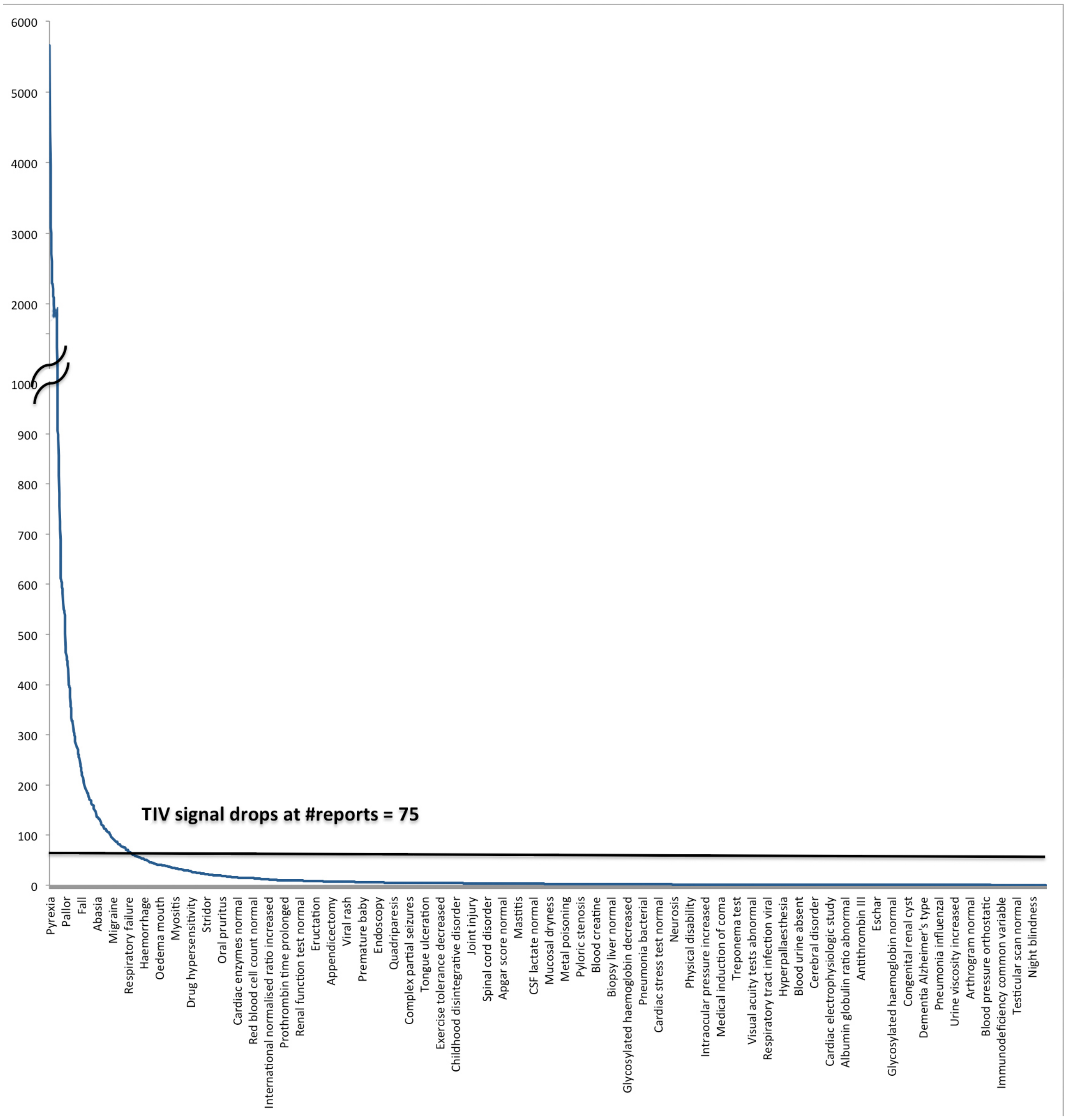
Independently from PRR normalization, the Chi-square significance test was applied to individual AEs on each list of queried data (TIV or LAIV) to determine how likely an AE would occur by random chance by looking at the probability distribution with the following formula.

$$\chi^2 = \Sigma_{i,k} ((X_i - \mu_i) / \sigma_i)^2 \quad (4.1)$$

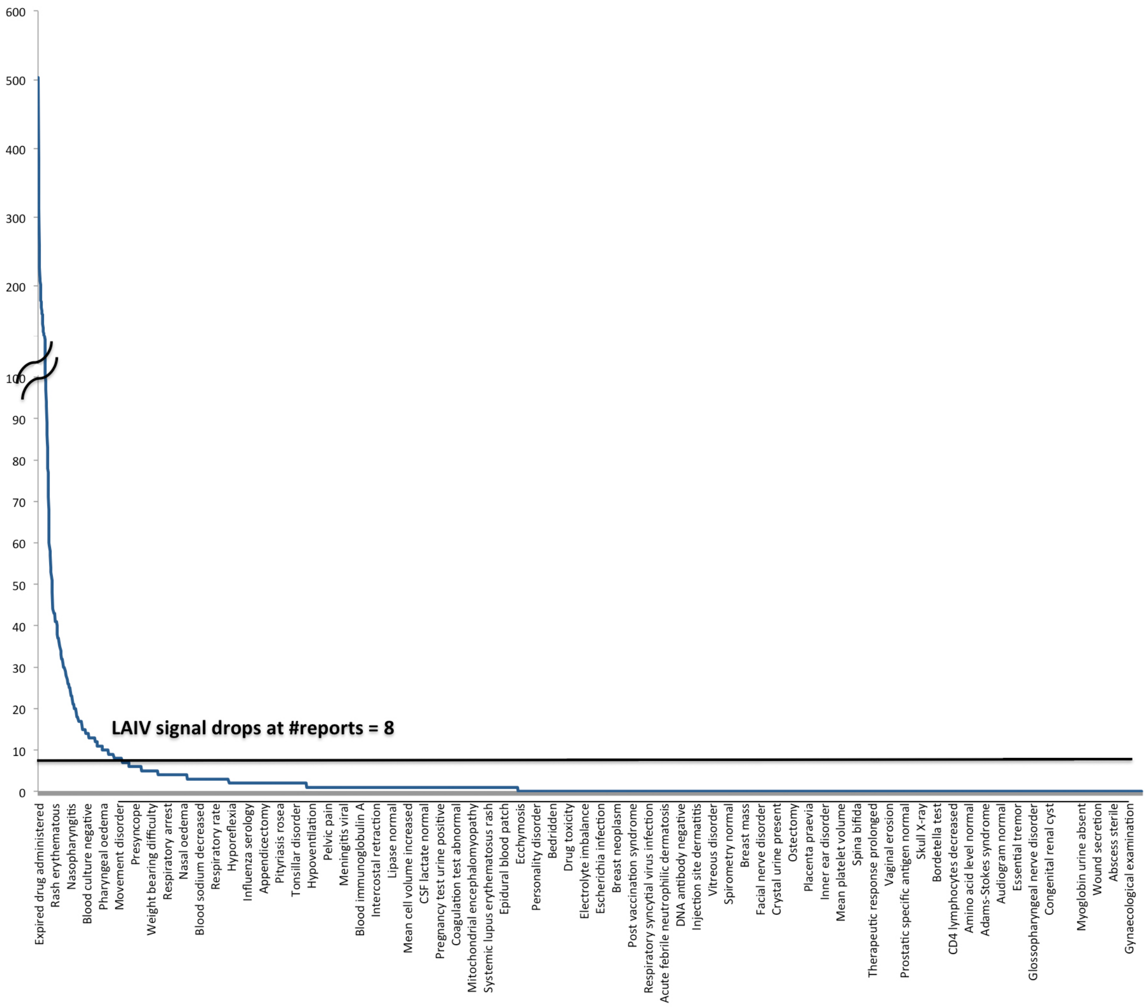
Under the assumption that the dataset was normally distributed, the formula annotated the calculation based on the sum of difference observed value [107] and the expected population mean (μ_i) denominated by the population standard deviation (σ_i). The calculation presented in (4.1) led to the computation for each AE in each group using a 2 X 2 frequency table based on the total number of all reports in each group (37621 TIV cases, 3707 LAIV cases) against the overall VAERS data (616215 cases) to derive the probability distribution for each AE in each group. For each AE term (AE_i), the number of TIV- or LAIV-post-vaccination occurrences of AE_i was used to compute for Chi-square distribution against the entire VAERS dataset. The 2 x 2 contingency table was composed of four disjoint counts for calculated for individual AEs in each group. An AE was called significant when its Chi square score was greater than 4 which implied P-value of approximately 0.05 or smaller.

4.3.4 Proportional Reporting Ratio score for rankings of vaccine-specific significant AEs

In order to avoid statistical error of insufficient information (outliers and by-chance occurrences), The sample size cut-off threshold of the number of reports for both groups was determined to be approximately 0.2 % of the total number of reports of each group. Using this cut-off, the biological implication would mean that at least 2 out of 1000 cases reported the AE of interest. These cut-offs were also supported by the signal curve on the total data. The fitting of this selection was selected by the plotting of report signal cut-off in the number of total reports of each group (figure 4.1). The number of cases for one AE to get called in for TIV group was evaluated to be 75 (number of reports >= 75), while the cut-off for LAIV group was evaluated at 8 (number of reports >= 8). Note that previous studies by Evans et al. recommended the minimal number of three reports in the traditional PRR method. At cut-off number of report = 3, the significant AE lists covered hundreds of AEs due to the nature of vast data pool.



(a) TIV cut-off signal curve



(b) LAIV cut-off signal curve

Figure 4.1 Signal curves to determine the data cut-off for TIV and LAIV analysis. Rendering and visualization of this plot was based on the adjusted display in MS Excel 2011 version 14.1.4. Some AE labels, though existed in dataset, were omitted on the plot.

To determine which AEs were exclusively enriched for TIV or LAIV, we excluded AEs that appeared as common signal in both lists. We also excluded ambiguous AEs such as *no adverse event,* or those of lab test result *normal.* We were then left with 48 TIV-enriched AEs and 68 LAIV-enriched AEs. These are AEs that their corresponding PRR score is at least 2, and Chi-square is greater than 4 (approximately of probability value of 0.05 or smaller).

4.3.5 Comparison of concept reorganization based on semantic similarity of the Ontology of Adverse Events (OAE)

OAE[107] was downloaded from http://sourceforge.net/projects/oae/. OAE was visualized with Protégé 4.0.2. TIV- and LAIV-related AEs were then clustered based on these structures for comparative analyses. Comparative analysis of TIV- and LAIV-specific AEs and their parent terms was extracted from the proposed OAE, visualized, and manually studied and compared between the two cohorts to derive a hypothesis.

4.4 Results

4.4.1 Overall study design

Data obtained by clinical observations are often high in statistical noise, which sometimes leads to temporal association that is wrongly believed to be causal without a sound conclusive argument [109]. In this study, we demonstrated that, by combining multiple statistical and bioinformatics methods, background noise and irrelevant information can be reduced to a minimal level. Scientist can draw a sensible hypothesis from these processed data and test it under experimental laboratory settings.

Figure 4.2 outlined our combinatorial bioinformatics approach. Starting from using Proportional Reporting Ratio (PRR) method as described in methods for normalization of data to distinguish true signal from background. Mining textual data of VAERS records is similar to the literature mining research. Some AEs are more frequently observed in the general population (*i.e.*, background) than others. PRR is implemented to over counter the background. We then calculated a Chi-square value to verify the significance level of each AE term in the specific group of vaccines. The Chi-square analysis and PRR were performed separately in parallel with each other. Those AEs with Chi Square score greater than 4 were kept for further studies. A PRR score is greater than 2 for TIV- or LAIV- enriched AE.

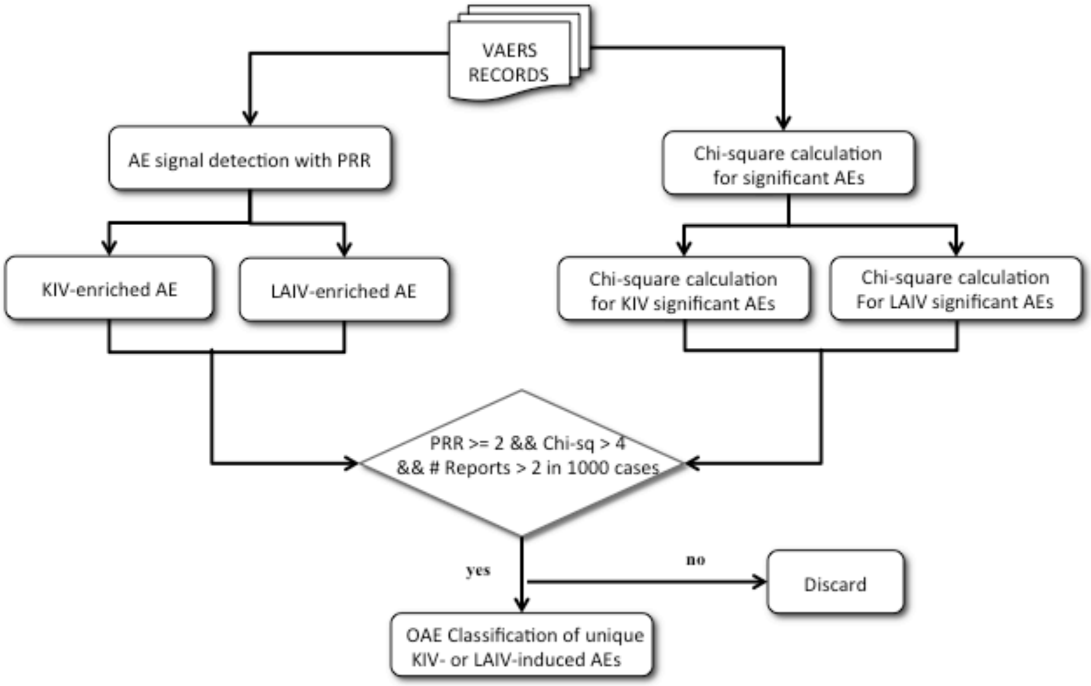

VAERS RECORDS
AE signal detection with PRR
Chi-square calculation for significant AEs
KIV-enriched AE
LAIV-enriched AE
Chi-square calculation for KIV significant AEs
Chi-square calculation For LAIV significant AEs
PRR >= 2 && Chi-sq > 4 && # Reports > 2 in 1000 cases
yes
no
Discard
OAE Classification of unique KIV- or LAIV-induced AEs

Figure 4.2 Workflow of integrative AE bioinformatics analysis.
VAERS records were retrieved based on the query criteria of 4 TIVs (Afluria, Fluarix, Fluvirin, and Fluzone) year 1990-2011 and 1 LAIV (Flumist) year 2003-2011. Parallel analyses of the Proportional Reporting Ratios and Chi-square significant test were performed on individual AEs to identify enriched and significant AEs in each group. Base level filtration of 0.2% of total number of reports was also applied to each AEs. AEs that were identified to have PRR >= 2, Chi-square >= 4, and number of reports >= 0.2% of total reports were then classified based on OAE structure. Classification of AEs filtered out AEs that overlapped between the 2 groups.

To ensure the specificity of the study, the condition also included the number of reports as a parameter to account for the possibility of outliers. The number of reports per AE must be at least 75 for TIV and 8 for LAIV to avoid the statistical error of random occurrences. These numbers were designed to determine the cut-off for manageability of the sets of AEs studied in each cohort by the signal curve drop (figure 4.1). In the established PRR method study, this cut-off number was recommended to be at least 3 occurrences [102,110].

4.4.2 Overall results: extracting differential AE profiles from VAERS TIV and LAIV data

As of May 18th, 2011, there were 7,520 MedDRa AE terms and 616,215 VAERS case records (one record may contain multiple AEs) listed for 75 vaccines reported to VAERS. The two subsets studied (TIV and LAIV) held 3,582 AE terms in total, and the comparison set contained 37,621 TIV reports and 3,707 LAIV reports. Following the overall analysis pipeline (Figure 4.2), TIV- or LAIV- enriched AEs were determined by Chi-square score (>4) , PRR score (>2), and the number of reports (>= 75 for TIV, >= 8 for LAIV). Figure 4.3 provides a Venn diagram showing the results after the three criteria of AE selection were applied in this study. In TIV group, there are 1236 AEs that their chi-square value is greater than 4 (χ^2(+)), and 2346 that did not pass this condition (χ^2(-)); 271 AEs passed the condition of sample size of at least 75 (count(+)), and 3311 AEs that did not (count(-)); and 1083 AEs that contained PRR score of at least 2 (PRR(+)) and 2499 AEs did not pass the PRR criterion.

The numbers as described indicated that even though sample size filtering screened out majority of the low-signal AEs, additional filtering by χ^2 and PRR scoring provided screening measures that could detect true signals of enriched AEs with high significance. Among 271 AEs that passed the sample size screening, there existed 223 AEs that passed the χ^2 test and 128 AEs that passed PRR evaluation. There were 80 AEs that overlapped within the screening of χ^2 and PRR tests, leaving 48 AEs out as the result of 3-criteria elimination.

In LAIV group, 757 AEs passed the χ^2(+) filtering, 2825 remained in (χ^2(-)); 274 AEs contained at least 8 records per AE (count(+)), and 3308 were left in count(-) group; and 898 AEs passed PRR condition (PRR(+)), while 2684 AEs were excluded by PRR (PRR(-)). Note that in LAIV cohort, χ^2 and PRR provided identical screening results after the sample size cut-off. This was coincidental and should not be taken that either screening method did not deliver further or useful filtering. There were 80 TIV AEs and 118 LAIV AEs that passed all three conditions with 31 AEs overlapped between the two lists. After screening out ambiguous or common AE terms, 48 AEs were included in TIV analysis, and 68 AEs were included in LAIV analysis. Tables 4.1 and 4.2 summarize the list of AEs for TIV and LAIV that were used for analytical clustering respectively. The AEs in these two tables were clustered based on an ontological classification method using OAE. These *clusters of adverse events, grouped by biological relevance based on physiological symptoms of individual,* are explained below.

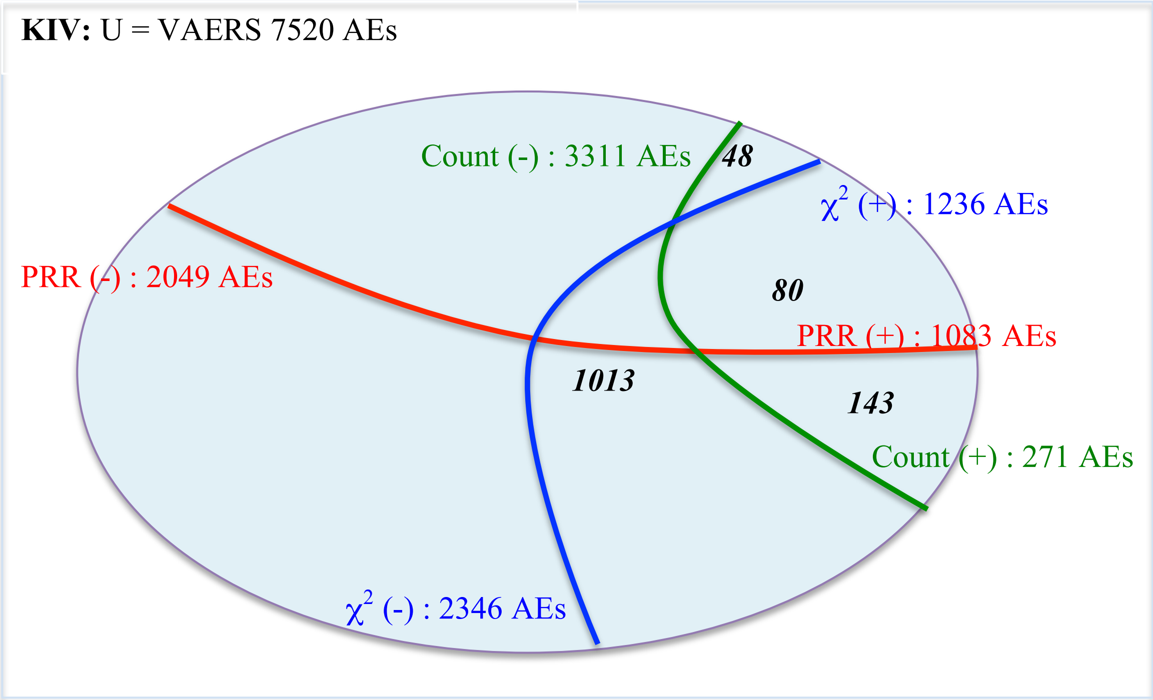
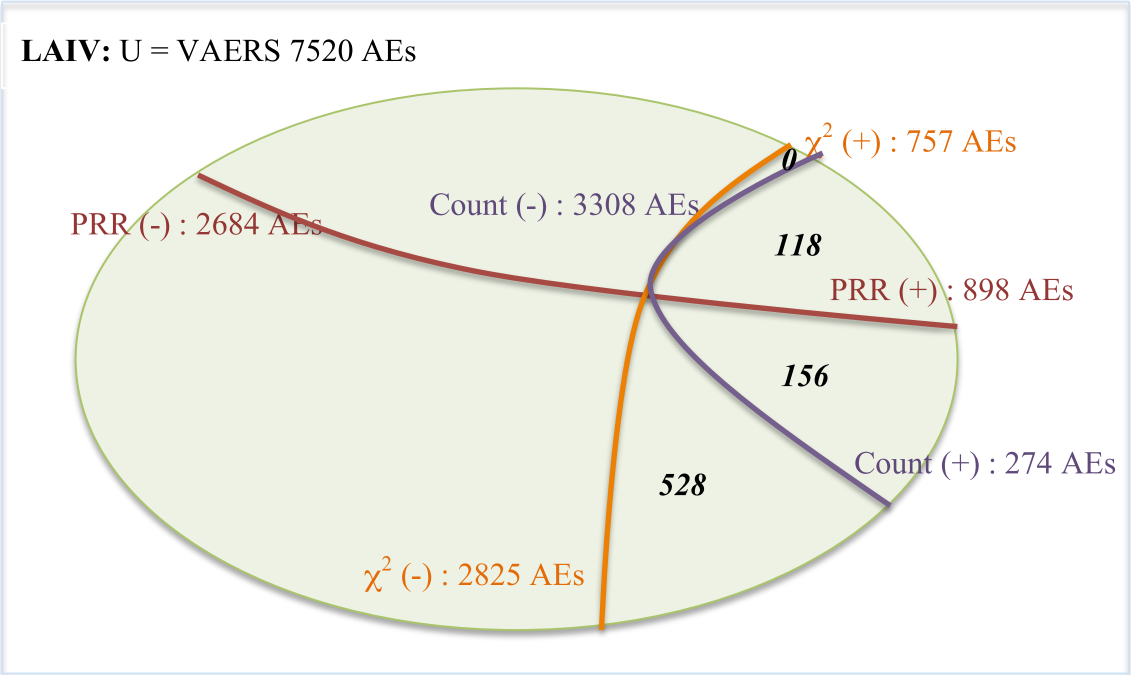


Figure 4.3 Venn diagram summary of the three filtering criteria in each group of



[1]vaccines from the pool of 3582 AEs analyzed in TIV and LAIV and the universe of

7520 AEs in the entire VAERS database; Chi-square value of >= 4 – χ^2(+), or < 4 – χ^2(-); PRR >= 2 – PRR(+), or PRR < 2 – PRR(-); and number of reports >= 75 in TIV or >= 8 in LAIV – count(+), or else – count(-).

Table 4.1 TIV-specific adverse events. 37,621 TIV-induced AE cases were reported. (* = serious adverse event, (r) = related to serious adverse event)

Adverse Event	Count	PRR(TIV)	Chi-sq(TIV)
AE with an outcome of lab test abnormal			
Electromyogram abnormal	107	4.87	248.14
behavior and neurological AE			
Dysarthria	91	2.80	75.22
behavior and neurological AE -> movement disorder AE			
Paralysis *	181	2.22	105.00
Hyporeflexia (r)	77	2.46	56.62
behavior and neurological AE -> sensory capability AE			
Pain	4516	2.12	2475.50
Chills	2286	2.78	2237.97
Pain in extremity	2106	3.40	2944.16
Paraesthesia (r)	1360	2.29	858.59
Hypoaesthesia (r)	1035	2.94	1114.16
Chest pain	725	2.37	492.53
Neck pain	543	2.06	260.28
Throat tightness	449	5.20	1134.22
Musculoskeletal pain (r)	432	4.00	767.24
Palpitations	267	2.70	240.23
Feeling cold	186	2.65	161.93

Skin burning sensation	113	3.04	127.95
Sensation of heaviness	100	2.99	109.56
Shoulder pain	78	3.40	107.20
Neuralgia (r)	77	2.37	52.32
cardiovascular disorder AE			
Heart rate increased	397	2.47	297.95
Hypertension	306	2.28	189.79
Blood pressure increased	216	4.11	398.44
Injection site haematoma	175	4.08	318.99
digestive system AE			
Dysphagia	299	2.23	175.09
Dry mouth	75	2.68	66.49
eye disorder AE			
Eye discharge	135	6.05	406.32
Eye irritation	115	4.62	249.17
hometostasis AE			
Pharyngeal oedema	256	4.30	503.52
Swollen tongue	158	4.56	336.46
Tongue oedema	105	2.51	81.09
Local swelling	88	2.05	40.97
medical intervention (not under 'adverse event')			
Accidental overdose	83	4.59	177.98

muscle adverse event			
Muscular weakness (r)	594	2.89	614.83
musculoskeletal system AE			
Laryngospasm	143	2.83	141.06
nervous system AE			
Guillain-Barre syndrome *	606	4.63	1321.69
Mobility decreased (r)	161	3.40	221.81
nervous system AE -> mobility decreased AE			
Injected limb mobility decreased (r)	561	4.72	1253.55
Joint range of motion decreased (r)	317	4.36	635.69
respiratory system AE			
Dyspnoea	2088	2.18	1180.96
skin adverse event			
Flushing	403	3.00	447.05
Eye pruritus	168	9.06	754.39
Hot flush	109	3.37	147.90

Table 4.2 LAIV-specific adverse events. 3,707 TIV-induced AE cases were reported.

Adverse Event	**Count**	**PRR(LAIV)**	**Chi-sq(LAIV)**
AE with an outcome of lab test abnormal			
Influenza serology positive	32	28.46	715.61

Chest X-ray abnormal	24	4.00	51.89
Blood creatine phosphokinase increased	19	3.96	40.33
Blood glucose increased	18	2.25	11.95
Urine analysis abnormal	15	2.78	16.43
Computerised tomogram abnormal	13	2.23	8.44
Nuclear magnetic resonance imaging brain abnormal	13	2.74	13.77
Electrocardiogram abnormal	11	2.36	8.27
Neutrophil percentage increased	11	3.00	14.13
Urine ketone body present	10	7.11	49.64
Lymphocyte percentage decreased	9	3.13	12.51
behavior and neurological AE			
Headache	383	2.22	257.06
Fatigue	159	2.16	97.17
Abdominal pain upper	54	3.49	92.50
Ear pain	22	3.01	28.48
Migraine	21	2.54	18.85
Abdominal discomfort	15	3.38	24.20
Burning sensation	15	2.25	9.97
Sinus headache	14	18.72	208.53
VIIth nerve paralysis	11	11.08	93.29
Facial paresis	9	5.37	30.53
Ataxia	8	3.50	13.72
cardiovascular disorder AE			
Epistaxis	71	14.71	823.81
Pericarditis	9	3.97	19.16

digestive system AE			
Retching	12	3.27	18.19
Dry throat	10	10.32	78.17
gustatory system AE			
Throat irritation	17	3.05	22.55
errored drug administration			
Expired drug administered	503	90.04	28507.94
Inappropriate schedule of drug administration	169	4.15	390.47
eye disorder AE			
Photophobia	17	2.66	16.90
Eye irritation	10	3.41	16.37
Visual impairment	9	3.05	11.92
homeostasis AE			
Swelling face	44	2.51	38.58
Eyelid oedema	13	2.03	6.49
immune system disorder			
Immunisation reaction	9	2.43	7.30
infection adverse event			
Croup infectious	11	9.58	78.80
injury and procedural complication AE			

Pregnancy test positive	10	3.21	14.60
medical intervention			
Drug exposure during pregnancy	77	3.90	159.92
Accidental exposure	30	20.56	490.83
Vaccination error	13	30.19	306.90
Underdose	11	11.64	98.65
Drug administration error	9	5.81	34.07
muscle disorder AE			
Bronchospasm	8	4.60	21.52
respiratory system AE			
Rhinorrhoea	210	9.47	1493.03
Nasal congestion	177	11.64	1593.58
Sneezing	53	10.78	436.15
Pneumonia	43	2.01	20.90
Sinusitis	43	5.92	167.24
Asthma	35	2.07	18.45
Respiratory tract congestion	32	7.68	175.14
Upper respiratory tract infection	25	2.22	16.05
Nasopharyngitis	25	4.08	55.69
Bronchitis	23	2.81	25.84
Sinus congestion	13	8.49	80.57
Nasal discomfort	12	47.77	422.18
Stridor	10	3.82	19.93
Postnasal drip	10	20.90	166.27

Lobar pneumonia	10	15.77	124.77
skin adverse event			
Pruritus generalised	15	2.94	18.45
Rash pustular	15	2.19	9.29
Henoch-Schonlein purpura	15	8.20	89.03
social behavior AE			
Activities of daily living impaired	23	2.36	17.29
Impaired work ability	8	3.96	16.94

4.4.3 Distinctive underlying biological activities were associated with the two groups of influenza vaccines

In summary, biological systems highlighted in TIV AEs were the behaviour/neurological system, immune system, and muscle/nervous systems. LAIV AEs appeared to cluster heavily in the respiratory system. Behavior/neurological adverse events were triggered by both TIV and LAIV. However, manual examination revealed that TIV-induced behaviour/neurological adverse events clustered around muscular, motor and movement disorders, and LAIV-induced adverse events were mainly pain in the head. Figure 4.4 illustrates the reorganization of TIV- and LAIV-induced AEs in details based on the Ontology of Adverse Events (previously known as the Adverse Event Ontology (AEO)[107]).

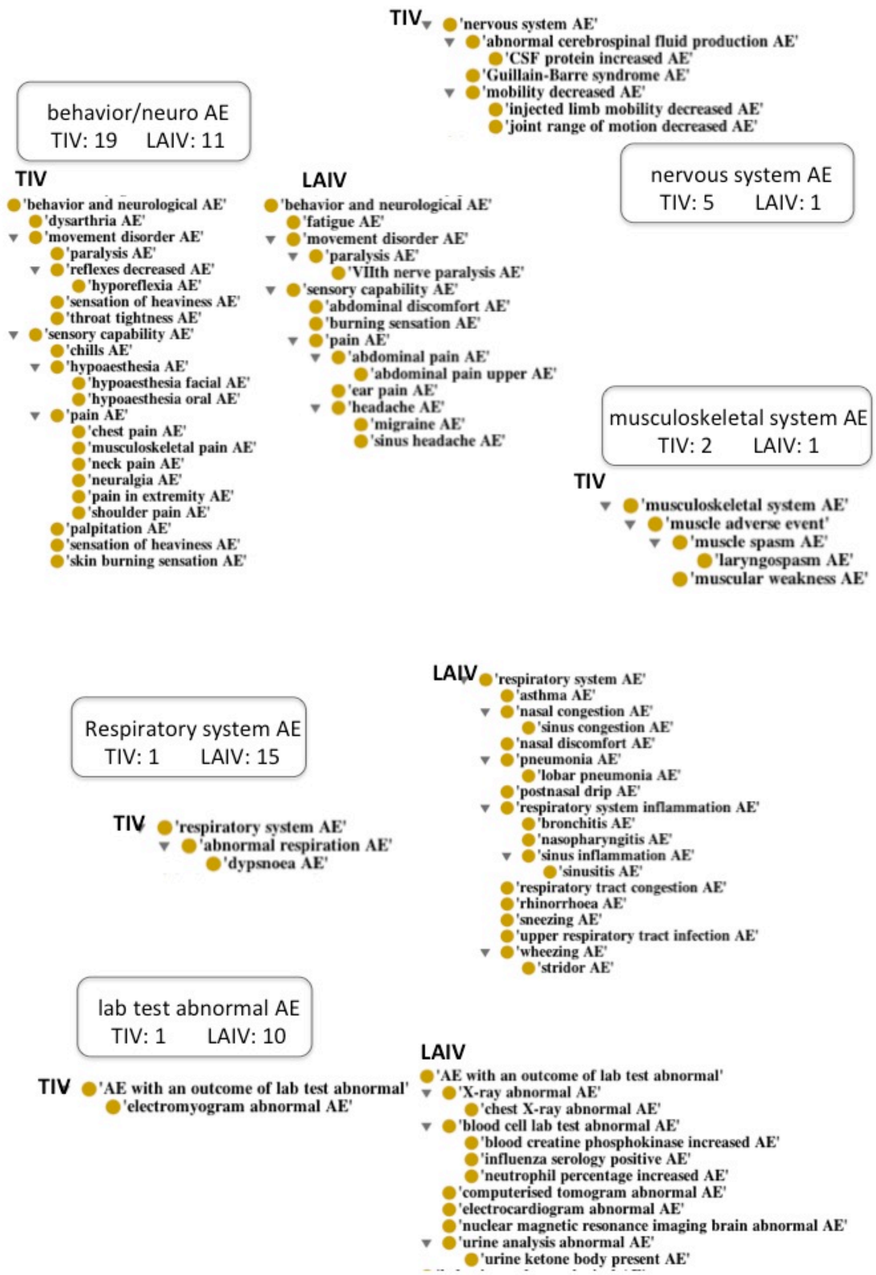


Figure 4.4 Diagram of AE count grouped by related symptoms.
Behavior/neurological system contains the most adverse events distributed in two groups of vaccines (40 adverse events; 25 in TIV, 15 in LAIV) but the clusters are significantly different in processes. TIV's behavior/neurological AEs are much more closely related to those of muscle and movement disorder while LAIV's behavior/neurological AEs cluster around pain in the head. Respiratory system AEs is listed as the most significant cluster in LAIV group with 16 AEs. Full listing can be found in table 1(TIV) and table 2 (LAIV).

For TIV's AEs that were clustered and enriched in movement and muscle disorders, AEs that displayed its symptom in movement and muscular disorders included joint range of motion decreased, mobility decreased, muscular weakness, Guillain-Barré syndrome, paralysis, and hyporeflexia. Another set of TIV-induced adverse events that was observed exclusively in TIV was edema (homeostasis and fluid dysregulation) in various parts of the body. The *Guillain-Barré* syndrome (GBS) was also a TIV-enriched AE. Detailed description of GBS as a TIV-enriched AE is provided in the following section. It should also be noted that no inflammatory responses came up as significant TIV-induced AEs (as opposed to LAIV).

LAIV influenza vaccine triggered other sets of biological activities in processes of the respiratory system (*e.g.* sinus headache, nasal congestion) and the respiratory system disorders that were characterized by inflammation – upper respiratory tract infection, pneumonia, bronchitis, and nasopharyngitis. Fifteen distinct adverse events were reported as over-represented respiratory system disorders. Activities in the hematopoietic system also suggested evidence in responses to stimulus related to inflammation. Furthermore, LAIV-induced adverse events showed a set of activities involved in the gustatory system and gustatory system-related activities.

4.4.4 Severe AEs are highly enriched in killed-inactivated influenza vaccine group

Based on FDA's definition of serious adverse event (http://www.fda.gov/safety/medwatch/howtoreport/ucm053087.htm), SAE is an adverse event that results in serious or fatal health condition such as death, permanent damage, or hospitalization. It should be noted that some TIV-specific AEs were serious adverse events (SAE). These included Guillain-Barre syndrome (Chi-square: 1172.79/P-value: 4.99E-257, PRR score: 4.63), and paralysis (Chi-square: 85.48/P-value: 2.34E-20, PRR score: 2.22). AEs that may be related to these SAEs and that are also enriched in TIV are hypoaesthesia, mobility decreased, joint range of motion decreased, musculoskeletal pain, paraesthesia, and neuralgia.

The association of GBS to influenza vaccination has long been debated in public health discussion. GBS is a serious immune system disease that has often been reported after influenza vaccine immunization[111,112]. GBS is categorized under immune system disorder based on cause of disease, or nervous system disorder based on its biological responses (muscular weakness, and paralysis). It is notable that a significant number of adverse events cases reported as a consequence of administrating influenza TIV vaccines are related to loss of muscle strength in various forms without the development of GBS. GBS often results in a key symptom in movement disorder. There are reports of associations between GBS and TIV Influenza vaccines, but the cause-effect relations remain inconclusive (Table 4.3).

Another evidence for TIV-associated compromised muscular system activities was the significantly ranked abnormal electromyogram result from TIV AE case reports. Electromyogram (EMG) is a test that evaluates electrical activity of muscle. Often, physicians utilize EMG as a method to diagnose GBS and other muscle-related disorders[113,114]. Observation of both abnormal lab test result AE (electromyogram abnormal) and physiological evidence in nervous and muscular disorders pointed toward an TIV-triggered inter-connecting activities in human body that were key symptoms of severe AEs (GBS, and paralysis).

No SAEs were enriched in LAIV AEs listed in Table 4.2. Pain in the head/neck area exists exclusively on the LAIV list, which may be explained by the route and method of LAIV administration (nasal spray). These respiratory system disorder AEs, for example, include pneumonia (Chi-square: 20.9/P-Value: 4.85E-06, PRR score: 2.01), lobar pneumonia (Chi-square: 124.77/P-value: 5.7E-29, PRR score: 15.77), Bronchitis (Chi-square: 25.84/P-value: 3.71E-07, PRR score: 2.81), and upper respiratory tract infection (Chi-square: 16.05/P-value: 6.16E-05, PRR score: 2.22). Operational errors reported as post-vaccination AE were reported with higher significance in LAIV's result.

4.4.5 Post-immunization Guillain-Barré Syndrome (GBS) in TIV recipients, occurred at a higher rate per number of reports than in LAIV

GBS is, although enriched in TIV, an infrequent incidence. Based on our literature review, the incidence rate of GBS in Influenza vaccine recipient is considered rare (approximately 1 in 100,000[112]). When looking at the number of reports of GBS in LAIV, we found this reporting rate may indicate a potentially rarer incidence rate in when compared to those of TIV recipients. Post-vaccination GBS was ranked among the over-represented AEs in all scoring matrices of TIV group (Table 4.1). In contrast, the information retrieved via this study showed a small reporting rate of GBS in the LAIV group that the statistical analysis did not recognize GBS as LAIV-enriched in Table 4.2. However, without consideration of PRR, GBS would pass the criteria of Chi-square and number of reports in both TIV and LAIV. Manual examination and cross referencing of input data (VAERS records) confirmed that the incidence was not manufacturer lot specific. Investigation of number of GBS cases reported per year in each group suggested that LAIV was less likely to induce post-immunization GBS than TIV (figure 4.5).

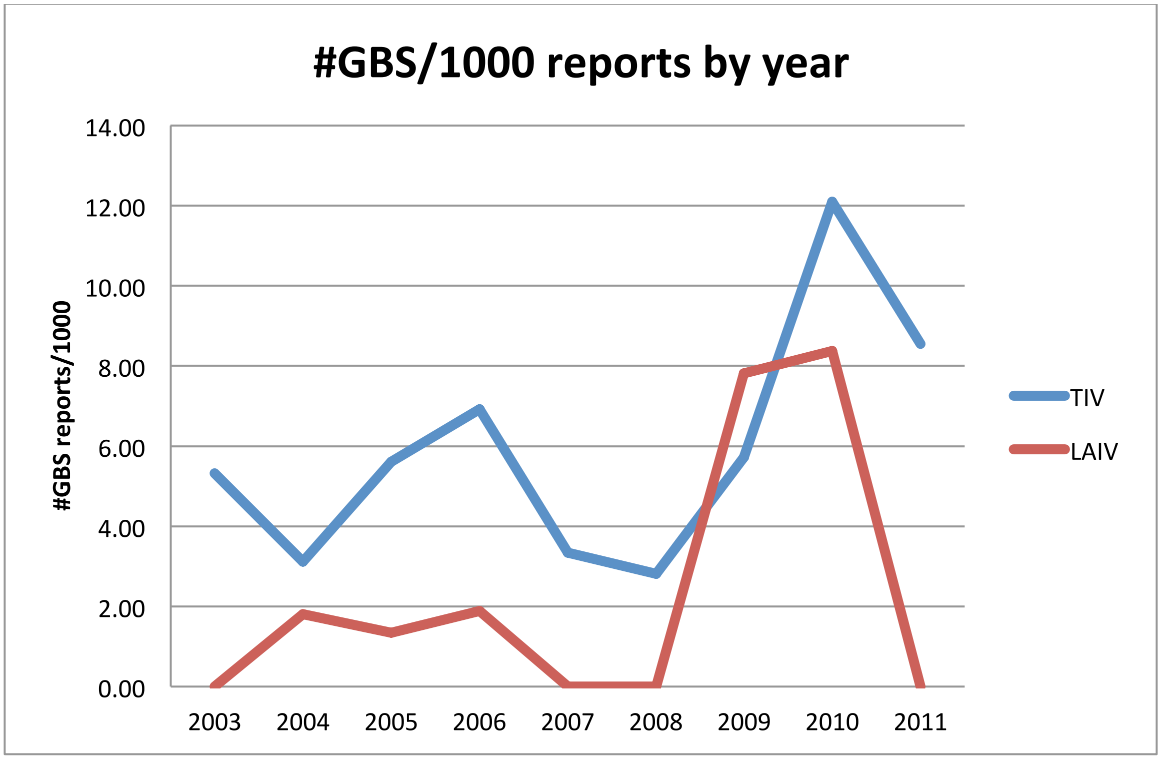


Figure 4.5 Reporting rate of GBS cases comparison between TIV and LAIV
Reporting rate as shown in this figure is based on the number of GBS cases per 1000 reports in VAERS. Results indicates the trend of higher reporting rate in TIV when compared to LAIV.

Table 4.3 Summary of association of GBS to influenza vaccines in peer-reviewed literature

Author	Title	Publication year	Type(s) of Influenza vaccine(s) studied	DB used	Method	Conclusion
Baxter, R.[115]	Recurrent Guillain-Barre Syndrome Following Vaccination	2012	TIV	Kaiser Permanente Northern California	review of medical records of GBS confirmed cases	low risk of recurrent GBS
Lee, S.J.[116]	Neurologic adverse events following influenza A (H1N1) vaccinations in children	2012	H1N1 MIV	N/A	single case study of 14 cases Nov.09-Mar.10	No major Neurologic AEs
Andrews, N.[117]	Guillain-Barre syndrome and H1N1 (2009) pandemic influenza vaccination using an AS03 adjuvanted vaccine in the United Kingdom: Self-controlled case series	2011	H1N1 MIV (2009)	N/A	review of patient records of post-vaccination GBS cases using self-controlled case series method on case identified in hospital episode statistics	no evidence of increased risk of GBS 6 weeks after vaccination
Cheo, Y.J.[118]	Serious adverse events follwing receipt of trivalent inactivated influenza vaccine in Korea 2003-2010	2011	TIV	Korea National Vaccine Injury Compensation Program (2003-2010)	retrospective review of clinical records, case investigation reports, conference materials and billing records.	GBS was the most-common SAEs reported after TIV immunization
Dieleman, J.[119]	Guillain-Barre syndrome and adjuvanted pandemic influenza A (H1H1) 2009 vaccine: multinational case-control study in Europe	2011	H1N1 MIV (2009)	N/A	GBS and Fisher syndrome case-control study (matched on age, sex, index date, and country)	no increased risk of GBS

Lee. G.M. [112]	H1N1 and Seasonal Influenza Vaccine Safety in the Vaccine Safety Datalink Project	2011	H1N1 MIV, LAMV, TIV, LAIV	Vaccine Safety Datalink	Nov'09-Apr'10 weekly signal detection analysis using self-controlled design /or/ current-vs-historical comparison	No association of GBS and other neurologic outcomes. For MIV - signal of Bell's Palsy. Higher reports of GBS in TIV than LAIV.
Sejvar, J.J.[120]	Guillain-Barre Syndrome Following Influenza Vaccination: Causal or Coincidental?	2011	MIV, TIV	N/A	analysis of past studies	question not answered, causality of GBS with regards to influenza vaccination remained inconclusive
Verity, C.[121]	Guillain-Barre syndrome and H1N1 influenza vaccine in UK children	2011	H1N1 MIV (?)	N/A	Sep'09- Aug'10 follow-up clinical questionairs	no association of GBS to influenza vaccination
Williams, S. E. [122]	Causality assessment of serious neurologic adverse events following 2009 H1N1 vaccination	2011	H1N1 MIV (2009)	VAERS	review of SAE reports in Oct'09-Mar'10	inconclusive association assessment, investigation of GBS and other SNAEs causality is difficult, VAERS reporting process can be improved
Burwen, D.R.[123]	Evaluation of Guillain-Barre Syndrome among recipients of influenza vaccine in 2000 and 2001	2010	not specified	Medicare claim data & hospital records	Incidence Rate Ratio	slightly non-significant elevated incidence rate ratio of GBS for all seasons combined

McNeil, M.M.[124]	A cluster of nonspecific adverse events in a military reserve unit following pandemic influenza A (H1N1) 2009 vaccination-Possible stimulate reporting?	2010	H1N1 MIV	VAERS	surey & review of index cases' VAERS reports, hospital records, vaccination status, aiagnostic results and outcome in comparison to VAERS reports of the same screen lot	GBS in the index case not confirmed, possbile stimulated reporting among reporters from the index case's cohort
Vellozzi, C. [125]	Adverse events following infuenza A (H1N1) 2009 monovalent vaccines reported to the Vaccine Adverse Event Reporting System, United States, October 1, 2009-January 31, 2010	2010	H1N1 MIV	VAERS	empirical Bayesian data mining and reporting proportions with clinical review of reports	death, GBS, and anaphylaxis were rare (<2 per million doses from ~10,000 VEARS reports)
Evans, D.[126]	"Prepandemic" Immunization for Novel Influenza Viruses, "Swine Flu" Vaccine, Guillain-Barre Syndrome, and the Detection of Rare Severe Adverse Events	2009	clinical trials of H5N1	N/A	review of past pandemic influenza studies	GBS association with Influenza vaccination remained inconclusive. Possible mechanism of GBS association with influenza vaccine was discussed.
Vellozzi, C.[127]	Safety of trivalent inactivated influenza vaccines in adults: Background for pandemic influenza vaccine safety monitoring	2009	TIV (1990-2005)	VAERS	PRR, review of reports of recurrent events and death	slightly elevated risk of SAEs, GBS - most frequently reported SAE, GBS requires continued monitoring

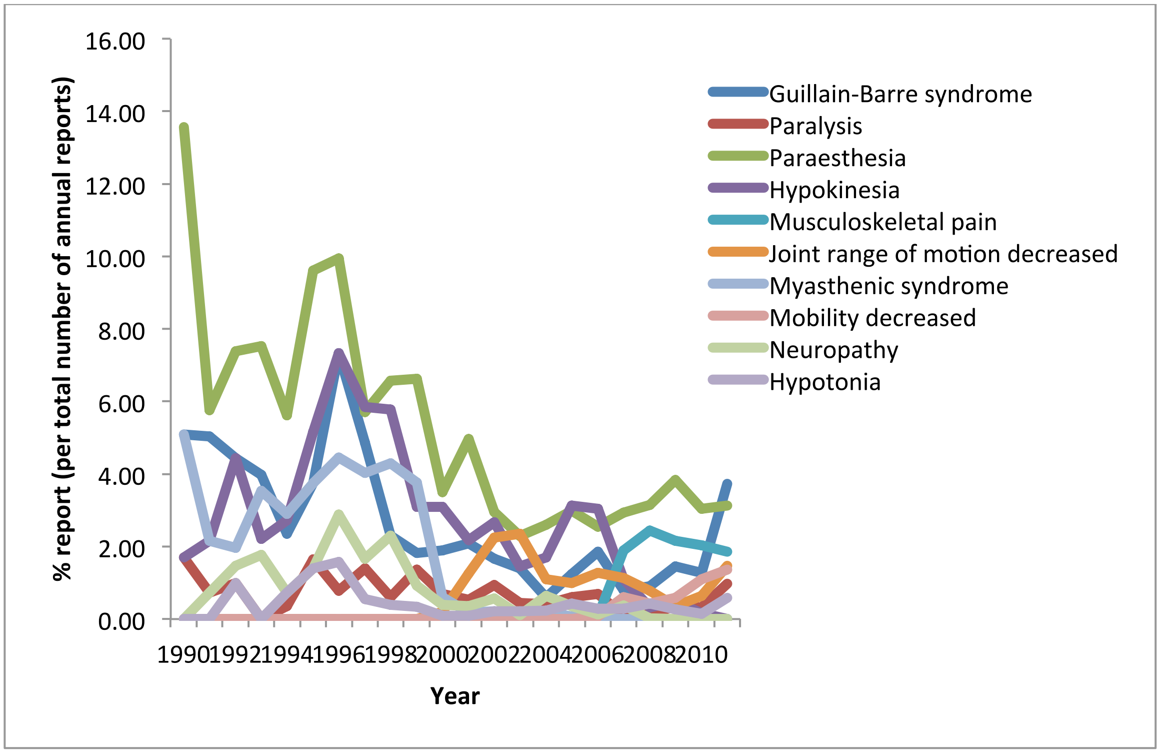
Juurlink, D.N.[128]	Guillain-Barre Syndrome after influenza vaccination in adults: a population-based study	2006	not specified	N/A	self-matched case-series method and time-series analysis	Influenza vaccination is associated with increased risk of hospitalization due to GBS
Izurieta, H.S.[129]	Adverse Events Reported Following Live Cold-Adapted, Intranasal Influenza Vaccine	2005	LAIV	VAERS (2003-2005)	report rate per 100,000 vaccinees	No unexpected serious risks, no GBS, may rarely cause anaphylaxis
Kao, C.[130]	Guillain-Barre syndrome coexisting with pericarditis or nephrotic syndrome after influenza vaccination	2004	not specified	N/A	2 individual case studies	pericarditis or onset nephrotic syndrome may coincidoe with GBS development.
Geier, M.R.[131]	Influenza vaccination and Guillain Barre syndrome	2003	not specified 1991-1999	VAERS	stat. analysis with Corel's Quattro Pro	increased risk of acute GBS and severe GBS in comparison to tetanus-diphtheria control group
Lasky, T.[132]	The Guillain-Barre Syndrome and the 1992-1993 and 1993-1994 influenza vaccines	1998	not specified	VAERS	patient survey and review of hospital discharging records based on VAERS reports	elevated relative risk of GBS 6 weeks after vaccination when combined the two reporting years. No increase in the risk when looking at each year individually.

MIV – Monovalent Inactivated Influenza Vaccine, TIV – Trivalent Inactivated Influenza Vaccine, LAIV – Live Attenuated Influenza Vaccine, VAERS – Vaccince Adverse Event Reporting System

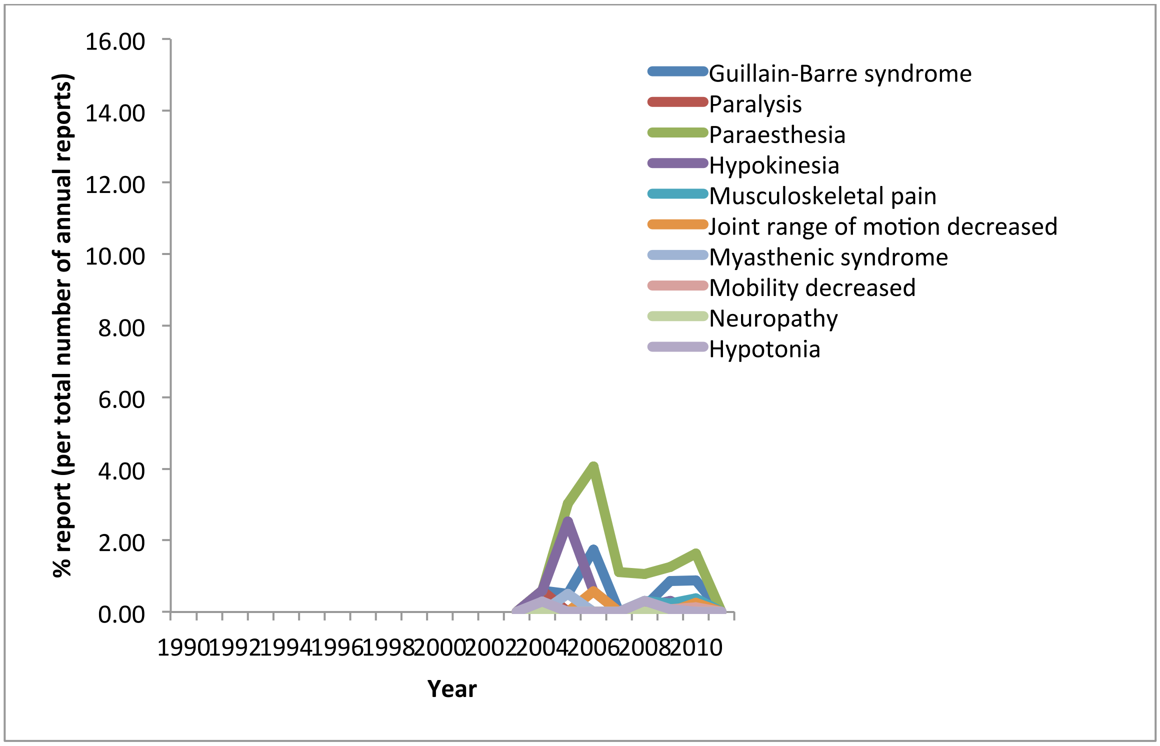
Table 4.4 Summary of statistical analysis testing if GBS and selected related AEs occurs independently of vaccine type.

year	#selected muscle AEs(TIV)	total cases reported in that year (TIV)	#selected muscle AEs(LAIV)	total cases reported in that year (LAIV)	P-Val (selected AEs occur independently of TIV)
2003	150	1790	0	34	7.81E-02
2004	163	2187	9	339	1.10E-03
2005	278	2842	13	198	1.37E-01
2006	244	2430	12	172	1.92E-01
2007	292	3165	2	180	1.84E-04
2008	321	3725	14	660	7.01E-09
2009	498	5231	36	1272	6.46E-15
2010	625	7165	26	797	9.44E-08
2011	134	1021	0	75	8.12E-04

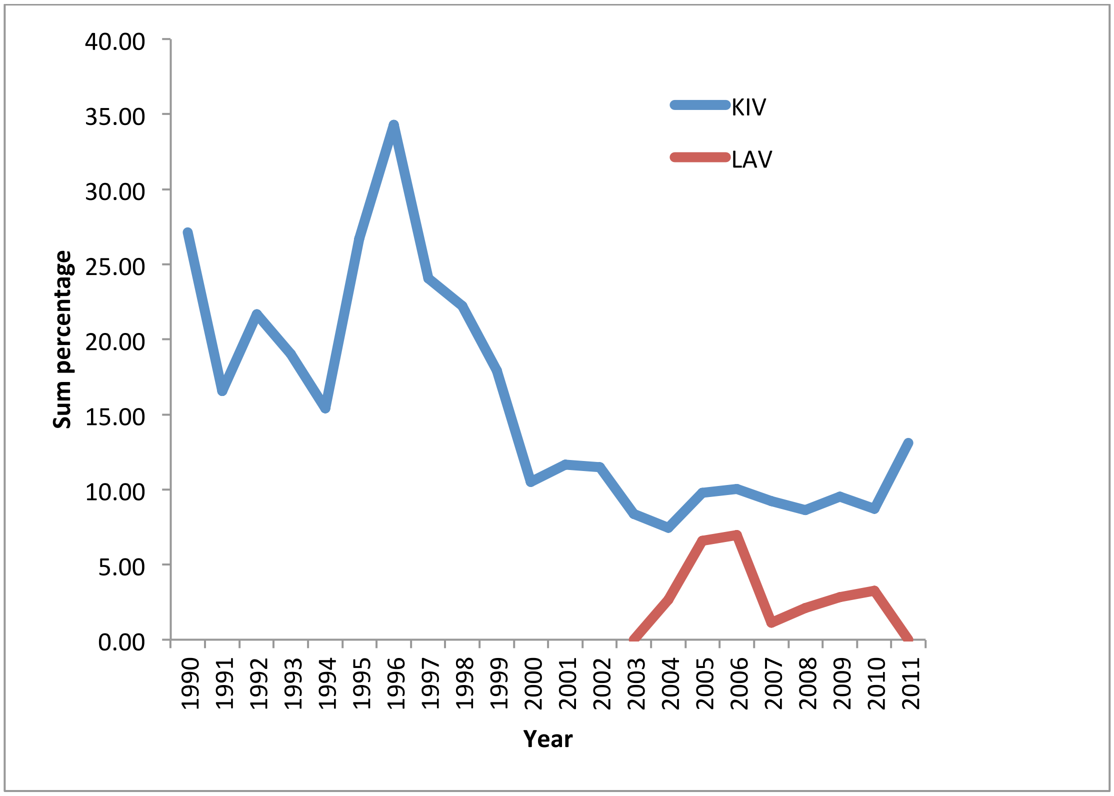
To further investigate the incidence rate of GBS and GBS-related AEs among the patients who reacted to the trivalent seasonal Influenza vaccines (both TIV and LAIV), we first tried to determine whether or not these incidences were specific to the year of reports (i.e. verified that GBS and GBS-related AEs occurrences were not manufacturer's lot specific). Figures 4.6 (A, B) display the percentage ratio of the reported cases of TIV- and LAIV-induced GBS and other related symptoms per year of reports. These AEs include symptoms resulting in movement and muscular disorders; paralysis, paraesthesia, hypokinesia, musculoskeletal pain, joint range of motion decreased, myasthenic syndrome, mobility decreased, neuropathy, and hypotonia. We extended the scope of GBS examination to include other muscular and nervous disorders in attempt to avoid the possibility of overlooking AEs that were closely related to GBS (i.e. symptoms of GBS include muscular weakness, or loss of muscle strength. These cases would have been ignored as non-important when focusing on GBS alone). Figure 4.6(C) shows the combined percentage of these selected AEs as one cluster based on year of reports. The denominator for the individual year calculation is the number of cases reported in that particular year. These three figures (4.6 (A)-(C)) have pointed towards the over-represented incidence rate of GBS and GBS-related AEs in TIV.



(A).


16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
% report (per total number of annual reports)
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Year
Guillain-Barre syndrome
Paralysis
Paraesthesia
Hypokinesia
Musculoskeletal pain
Joint range of motion decreased
Myasthenic syndrome
Mobility decreased
Neuropathy
Hypotonia

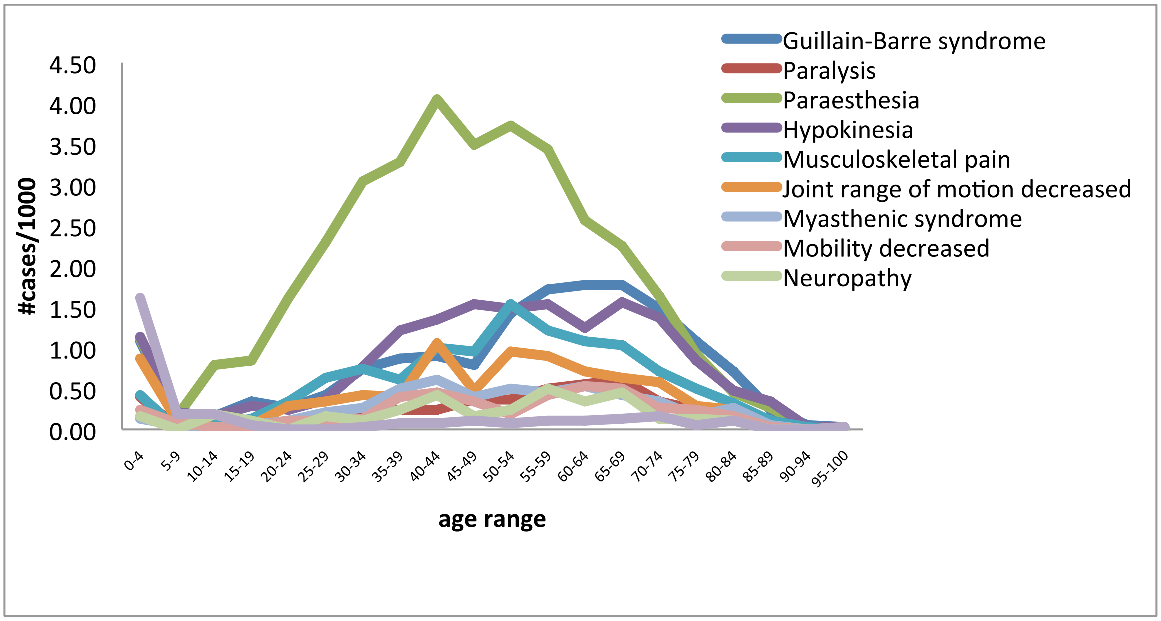
(B).


40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Sum percentage
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year
KIV
LAV

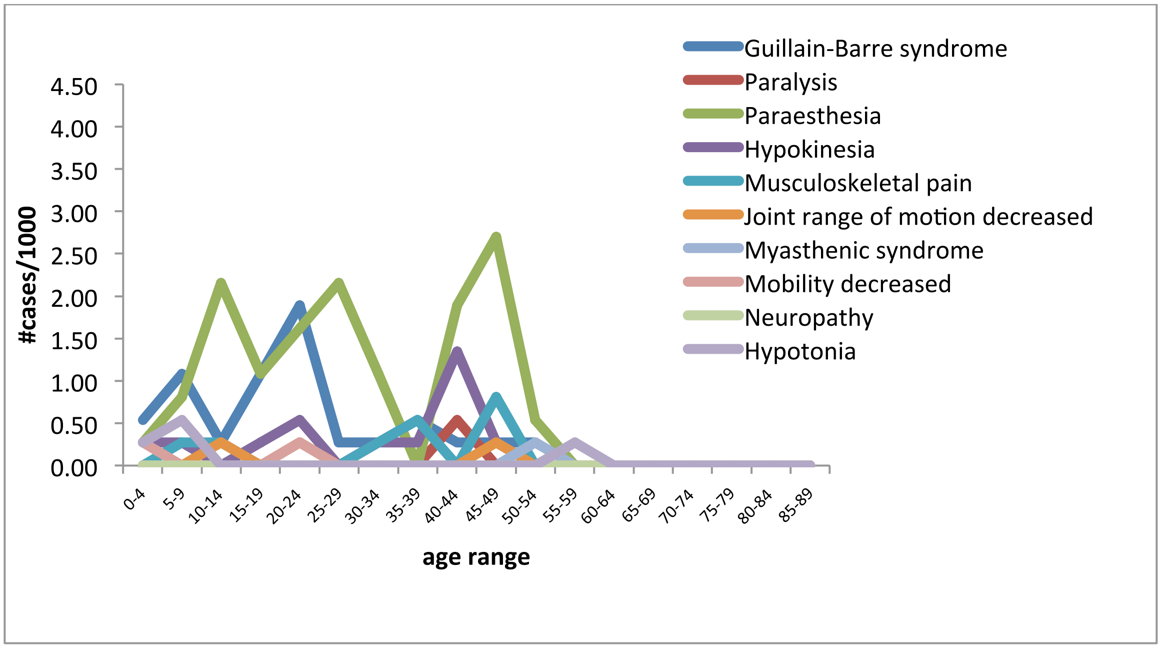
(C).

Figure 4.6 Report distribution comparison by year (a) TIV (b) LAIV.
LAIV was recently released and therefore data available is from 2003 onward. The raw number of occurrences was scaled to percentage by the number of reports in each year. Guillain-Barre Syndrome (GBS) and other GBS-related symptoms synchronized trends in both TIV and LAIV groups need further exploration. (c) depicts the combined percentage number of GBS and its related AEs. Careful examination is recommended to investigate GBS-musculoskeletal pain-muscular weakness correlation in 2008-2009 incidents. Percentage of occurrences by year show that the incidents of GBS and other related symptoms are at a much smaller rate in LAIV when compared to TIV group

Since LAIV surveillance data recently became available in 2003 (*Flumist* was released in late 2003), the comparison of TIV-LAIV by Chi-square significance test against each other could only be calculated from 2003 on. Table 4.4 summarizes TIV's probability value of how much more GBS and GBS-related AEs are more overly represented in comparison to LAIV. The results show that all but three P-values are smaller than 0.05, signifying that GBS and GBS-related AEs are more enriched in TIV. One of those three years with the P-Value bigger than 0.05 is a result of statistical artifact while the raw number of occurrence in LAIV is equal to 0 (year 2003). Figures 4.6(a,b) plot the age-range percentage ratio in which each examined AE occurred based on the total sample size in this study of TIV and LAIV, respectively. The results indicate that all GBS and GBS-related AEs occurred at higher rate in early age group (0-5 years) with another trend of increasing occurrence in middle to later age range (40-75) (Figure 4.6(A)). There was not a suggestion of age pattern in GBS and GBS-related AE occurrences in LAIV group (Figure 4.6(B)).



(a).



(b).

Figure 4.7 Report distribution comparison by age range (a) TIV (b) LAIV. Age distribution of occurrences do not show significant differences in which GBS and other related symptoms occur between the two groups

4.5 Discussion

We hypothesized that the AE differences in the two sets of recipients (TIV vs. LAIV) emerged from different immune-response pathways induced by each type of vaccine, so we demonstrated that the combinatorial bioinformatics approach can overcome the complex challenges in public post-vaccination event record data. The strategy of this study resolves the issue of high-noise data, especially when these data contain high-value hidden knowledge that can be evaluated by robust statistical tests. It is crucial to identify background information, as some AEs are common to many vaccines. Because the number of reports in VAERS database is large (616,215 cases, 75 vaccines), background information is not sensitive to minor change or adjustment such as removing reports from one or two vaccines from the studied sample set. Though the combinatorial workflow (summarized in figure 4.2) was applied specifically to VAERS data, the concept can also be adapted and applied to other questions in the Translational Informatics domain.

Furthermore, the preliminary result in the form of flat list may be informative at an individual AE level. However, it is difficult to examine the flat list to identify the underlying biological systems when the system is composed of multiple interactions among multiple participating AEs. It is challenging to draw any connections between biological processes while the significant individual AE terms scatter across various different biological functions and systems. Examining these AEs based on their score rankings along with reorganizing result by their semantic similarity and functional relevance leads to a better representation of data that can overcome this issue.

Incidents of serious AEs are not always easy to detect in terms of population statistics, as they may require a long period of observation. Therefore, detection and confirmation of such incidents can be inconclusive or take a long time. Examples of time-consuming observations of vaccine post-marketing AEs include Guillain-Barré Syndrome (GBS) after 1976 Swine Flu to 2009 Influenza vaccine campaigns[94,111], anthrax vaccine adverse event studied from 1990 to 2007[133], and 1990-2007 measles vaccine adverse effects studied in the Ivory Coast[134]. Although 1976 incidence of Swine Flu vaccination was detected in real time, debate and discussion of the incidence remained inconclusive.

One interesting finding from this study was the occurrence of GBS in TIV recipients. There have been many controversial results with regard to the post Influenza vaccination incidents as demonstrated in table 4.3. Haber et al. concluded in their studies that the occurrence of GBS is Influenza vaccine recipients was merely temporal association, and the causal association was not implicated with any solid evidence[94]. Furthermore, Haber et al. had previously challenged the study of Souayah et al.[135] that used VAERS dataset similar to our study by pointing out VAERS limitation of omitted information of case follow-up procedure. Haber et al. also stated in their debate that GBS incident as Influenza vaccine adverse event reporting interval should be determined by Influenza season rather than calendar year[136]. After a careful examination of the data, we considered that pooling entire VAERS data with our methodology could overcome the issues of omitted data or reporting intervals. Whether or not the reporting interval was based on the season or calendar year, overall incidence rate was not depending on any one particular year or season. We also took into account of the number of post-Influenza-vaccine GBS confirmed by neurologists in VAERS (1995-2003) as investigated by Haber et al. to be 82% which, when combined with additional data that became available in the later years, was still significant in a large dataset such as those studied in this work. Souayah et al.'s study in 2009 remained firm in their conclusion of Influenza vaccine-associated GBS with significant incidence rate[137]. Evans et al. also associated GBS and rare adverse events with Influenza vaccine by conducting the a comparative study of the novel Influenza (swine flu) prepandemic data in 2009 to 1976 National Influenza Immunization Programme data[111]. Our study found that Souayah's and Evans' GBS association to Influenza vaccines held true only when considering TIV, not LAIV. To the best of our knowledge, this study is the first to systemically compare the differences between TIV- and LAIV-associated VAEs.

Even though the safety of TIVs is generally accepted at population level, our analysis points towards LAIV as an alternative that is less likely for the recipient to develop severe AEs such as GBS or paralysis. However, although the number of cases of SAEs in LAIV were small, and therefore SAEs were not significantly enriched in LAIV, occurrences of SAEs should still be investigated carefully. While GBS and paralysis (as categorized to be severe adverse events) were enriched in TIV, weighted-AE scoring method should be applied in the future direction of this study to properly address the issue of SAEs. All SAEs should automatically rank high in the significance of AE. Lee et al. has conducted a similar investigation on killed and live Influenza vaccines. With different approach to the same question, our results were compatible with Lee's work in terms of enriched GBS in killed Influenza vaccine group as opposed to live Influenza vaccine group. Lee et al. examined both trivalent and monovalent seasonal Influenza vaccines. The rate of GBS in both trivalent and monovalent killed Influenza vaccines in Lee’s work was approximately 1 in 100,000, which was significant and above the background level[112]. Our study of trivalent killed Influenza vaccines revealed a similar story of the incidence rate of approximately 1 in 100,000 as well. Further study to verify the relevance of post-vaccination GBS in TIV group as opposed to LAIV group is needed. Currently, the results suggested enriched TIV-induced GBS and paralysis, while LAIV Influenza vaccine did not show any evidence of correlations to GBS based on vaccine-specific enriched AEs in both groups.

One observation from the compilation of table 4.3 came to our notion that majority of peer-reviewed publication on influenza vaccine-induced GBS in the recent years were the studies of monovalent influenza vaccine. These studies concluded that there were no associations of GBS to monovalent inactivated influenza vaccines. It should also be noted that all peer-reviewed publications on the same topic that concluded other wise were tested on trivalent inactivated influenza vaccines. Further investigation on the subject of monovalent versus trivalent inactivated influenza vaccines as the trigger of post-immunization GBS is required before conclusion can be made.

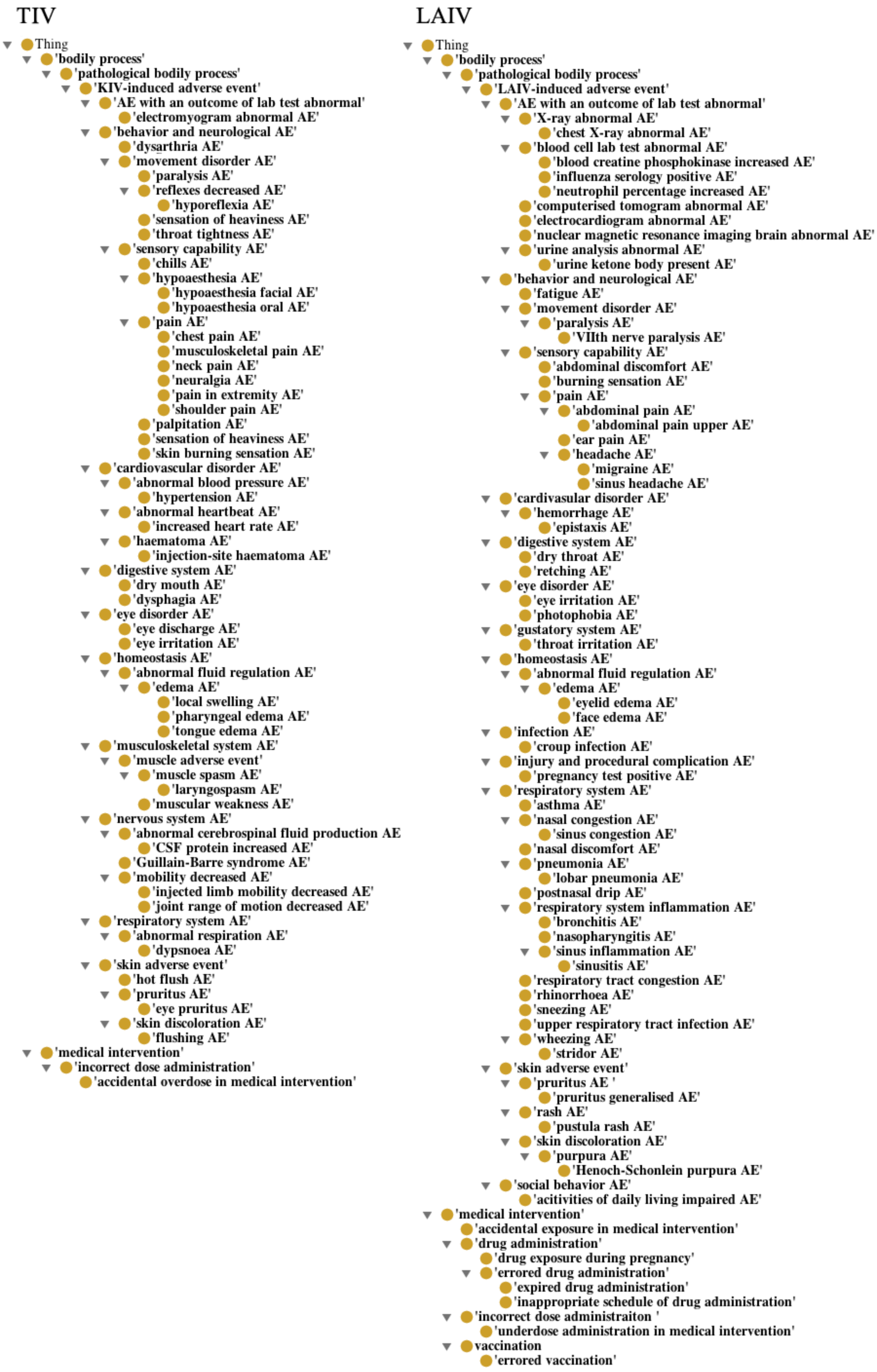
Further analysis of TIV- and LAIV-induced AEs by reorganizing into an ontological structure with reference to other community-accepted ontologies reveals certain challenges that need to be properly addressed. We have clustered AEs of each group of vaccines to COSTART (1995) (the foundation vocabulary that MedDRA was built upon) with embedded hierarchical structure available on BioPortal (http://bioportal.bioontology.org/visualize/40390). We found that COSTART hierarchical structure might not be a suitable term reorganization reference as COSTART/MedDRA structure lacked a specific definition in which the aspect of this hierarchical tree was based on. It was not clear if the hierarchy was defined by biological processes, or anatomy of the body. Hypothetically, because COSTART/MedDRA is a comprehensive dictionary of adverse event descriptors but it was not created for the purpose of computation, the structure organization may not be fully equipped for ontologically machine processing. Many concepts in COSTART/MedDRA listed synonyms that were not true synonyms; e.g. sinus headache was defined to be synonymous to headache, infection upper respiratory was defined to be synonymous to infection, or chest X-ray abnormal was defined as a synonym of lung disorder. Many examples of this kind of synonym error occur throughout the COSTART hierarchy. Another major issue in using COSTART/MedDRA as an ontological reference was that COSTART contained duplicate classes that caused ambiguity in many situations (figure 4.9). For example, ear disorder was a child under a parent class of the same class name ear disorder, hemorrhage was a child of parent class haemorrhage [same word], hypotension was a child of parent class shock syndrome which was, in turn, a sibling class of another concept that also has class identifier of hypotension.

We then explored another clinical ontology of SNOMED Clinical Terms (Version 07/31/2010) to find an alternative for AE term reorganization for the purpose of recognizing AEs based on biological relevance (figure 4.8). We found that while SNOMED CT was very thoroughly defined with the most detailed information of anatomical and physiological description, this ontology still may not be the best alternative for such purpose. The comprehensive organization of terms in SNOMED CT resulted in the structure that did not provide an apparent clustering for term recognition based on biological process, because classes at the individual AE level (leaf nodes) were scattered across the ontology due to the nature of very detailed parental subclasses. To find a plausible solution that accounted for such situation, we had to examine the issue at hand; i.e. MedDRA vocabulary may not be the best nomenclature system for the purpose of AE reporting, but it had been in use for VAERS for many years. Therefore, to be able to mine for discovery within VAERS records, we must find methods to process and interpret VAERS data in an efficient way that discovers that knowledge embedded with it. Also, sometimes, COSTART/MedDRA terms that are reported in VAERS fall into many semantic types in SNOMED CT, namely Body structure, Clinical finding, Procedure, Special concept, or Qualifier value. Such terms are within the same semantic type and there are also sub-structures that may further divide COSTART/MedDRA terms into many separate groups. One example scenario as appeared in this situation is how Edema was described and categorized on the two ontologies. In SNOMED, Edema is a Clinical finding while Edema of pharynx is a child of Disorder characterized by edema, Disorder characterized by edema is a subclass of Disease, while Disease is a Clinical finding. This, in turn, resulted in Disease that was a sibling of class Edema while containing a child of a child of class Edema of pharynx. This separation by different semantic types occured throughout SNOMED CT.

The creation of OAE to complement for issues described above, it is evident in this study that semantic-similarity-based term recognition and reorganization (figure 4.7) can provide insights into underlying processes that may be overlooked or hard to detect without prior-knowledge structuring. By embedding top-level information obtained from other sources to reconstruct branches of related adverse events, we have shown that examining OAE, parent nodes of these AEs can identify biological systems that may be associated with post-vaccination reactions in the two groups of vaccines in this study. This combinatorial bioinformatics approach can also be modified to answer biomedical questions in other domains of interest.

The role of gleaning and cleaning *real-world* clinical data of this approach also introduces a novel hypothesis generator tool to aid translational informatics as the results are supported by statistical evaluation and validation of the findings. The method is also designed to be discovery-driven rather than the traditional research hypothesis-driven approach. Two possible hypotheses that are derived from this post-vaccination adverse event investigation could be: (1) Hypothesis 1: TIV induces the occurrence of GBS that may be explained by the trigger in behavioral & neurological process, and (2) Hypothesis 2: LAIV is more likely to trigger respiratory inflammatory response than TIV due to its mode of administration.

Arising from this finding lies in the interest of personalized medicine of individuals. As a recent study by Liang et al. indicates the occurrence of GBS as below the background rate of severe adverse event induced by Influenza vaccine[91], those observations were made on the whole population of Influenza vaccine recipients including the majority who did not develop any major post-vaccination complication. Our study, in contrast, focuses on the sub-population of those whose case has been submitted to VAERS as having a post-vaccination complication. This difference in the focused population group may lead to the hypothesis as to which molecular or genetic variation of the person can cause the occurrences of Influenza-vaccine-induced severe AEs.



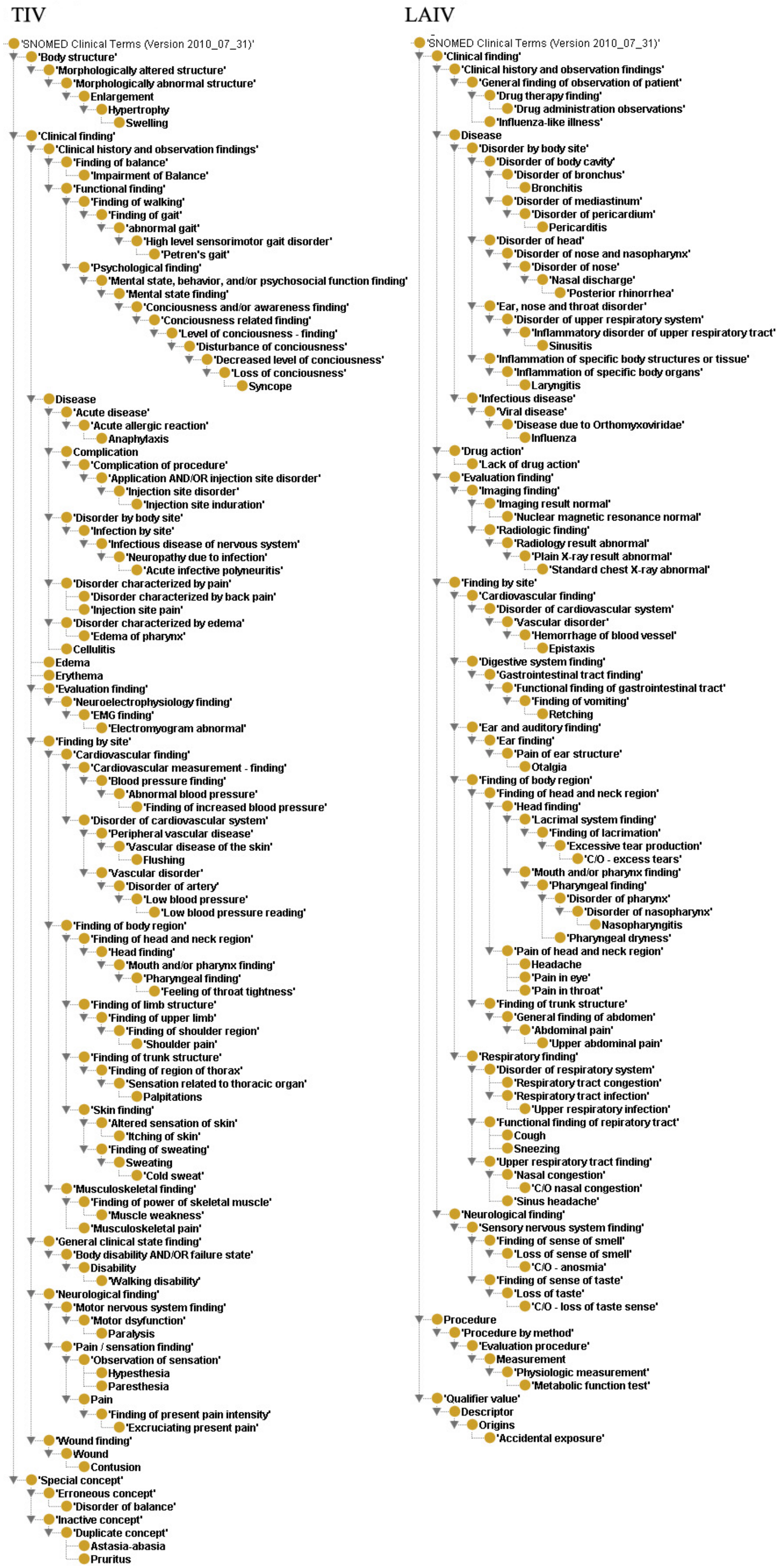
TIV

- Thing
  - 'bodily process'
    - 'pathological bodily process'
      - 'KIV-induced adverse event'
        - 'AE with an outcome of lab test abnormal'
          - 'electromyogram abnormal AE'
        - 'behavior and neurological AE'
          - 'dysarthria AE'
          - 'movement disorder AE'
            - 'paralysis AE'
            - 'reflexes decreased AE'
              - 'hyporeflexia AE'
            - 'sensation of heaviness AE'
            - 'throat tightness AE'
          - 'sensory capability AE'
            - 'chills AE'
            - 'hypoaesthesia AE'
              - 'hypoaesthesia facial AE'
              - 'hypoaesthesia oral AE'
            - 'pain AE'
              - 'chest pain AE'
              - 'musculoskeletal pain AE'
              - 'neck pain AE'
              - 'neuralgia AE'
              - 'pain in extremity AE'
              - 'shoulder pain AE'
            - 'palpitation AE'
            - 'sensation of heaviness AE'
            - 'skin burning sensation AE'
        - 'cardiovascular disorder AE'
          - 'abnormal blood pressure AE'
            - 'hypertension AE'
          - 'abnormal heartbeat AE'
            - 'increased heart rate AE'
          - 'haematoma AE'
            - 'injection-site haematoma AE'
        - 'digestive system AE'
          - 'dry mouth AE'
          - 'dysphagia AE'
        - 'eye disorder AE'
          - 'eye discharge AE'
          - 'eye irritation AE'
        - 'homeostasis AE'
          - 'abnormal fluid regulation AE'
            - 'edema AE'
              - 'local swelling AE'
              - 'pharyngeal edema AE'
              - 'tongue edema AE'
        - 'musculoskeletal system AE'
          - 'muscle adverse event'
            - 'muscle spasm AE'
              - 'laryngospasm AE'
            - 'muscular weakness AE'
        - 'nervous system AE'
          - 'abnormal cerebrospinal fluid production AE
            - 'CSF protein increased AE'
          - 'Guillain-Barre syndrome AE'
          - 'mobility decreased AE'
            - 'injected limb mobility decreased AE'
            - 'joint range of motion decreased AE'
        - 'respiratory system AE'
          - 'abnormal respiration AE'
            - 'dypsnoea AE'
        - 'skin adverse event'
          - 'hot flush AE'
          - 'pruritus AE'
            - 'eye pruritus AE'
          - 'skin discoloration AE'
            - 'flushing AE'
  - 'medical intervention'
    - 'incorrect dose administration'
      - 'accidental overdose in medical intervention'

LAIV

- Thing
  - 'bodily process'
    - 'pathological bodily process'
      - 'LAIV-induced adverse event'
        - 'AE with an outcome of lab test abnormal'
          - 'X-ray abnormal AE'
            - 'chest X-ray abnormal AE'
          - 'blood cell lab test abnormal AE'
            - 'blood creatine phosphokinase increased AE'
            - 'influenza serology positive AE'
            - 'neutrophil percentage increased AE'
          - 'computerised tomogram abnormal AE'
          - 'electrocardiogram abnormal AE'
          - 'nuclear magnetic resonance imaging brain abnormal AE'
          - 'urine analysis abnormal AE'
            - 'urine ketone body present AE'
        - 'behavior and neurological AE'
          - 'fatigue AE'
          - 'movement disorder AE'
            - 'paralysis AE'
              - 'VIIth nerve paralysis AE'
          - 'sensory capability AE'
            - 'abdominal discomfort AE'
            - 'burning sensation AE'
            - 'pain AE'
              - 'abdominal pain AE'
                - 'abdominal pain upper AE'
              - 'ear pain AE'
              - 'headache AE'
                - 'migraine AE'
                - 'sinus headache AE'
        - 'cardivasular disorder AE'
          - 'hemorrhage AE'
            - 'epistaxis AE'
        - 'digestive system AE'
          - 'dry throat AE'
          - 'retching AE'
        - 'eye disorder AE'
          - 'eye irritation AE'
          - 'photophobia AE'
        - 'gustatory system AE'
          - 'throat irritation AE'
        - 'homeostasis AE'
          - 'abnormal fluid regulation AE'
            - 'edema AE'
              - 'eyelid edema AE'
              - 'face edema AE'
        - 'infection AE'
          - 'croup infection AE'
        - 'injury and procedural complication AE'
          - 'pregnancy test positive AE'
        - 'respiratory system AE'
          - 'asthma AE'
          - 'nasal congestion AE'
            - 'sinus congestion AE'
          - 'nasal discomfort AE'
          - 'pneumonia AE'
            - 'lobar pneumonia AE'
          - 'postnasal drip AE'
          - 'respiratory system inflammation AE'
            - 'bronchitis AE'
            - 'nasopharyngitis AE'
            - 'sinus inflammation AE'
              - 'sinusitis AE'
          - 'respiratory tract congestion AE'
          - 'rhinorrhoea AE'
          - 'sneezing AE'
          - 'upper respiratory tract infection AE'
          - 'wheezing AE'
            - 'stridor AE'
        - 'skin adverse event'
          - 'pruritus AE '
            - 'pruritus generalised AE'
          - 'rash AE'
            - 'pustula rash AE'
          - 'skin discoloration AE'
            - 'purpura AE'
              - 'Henoch-Schonlein purpura AE'
        - 'social behavior AE'
          - 'acitivities of daily living impaired AE'
  - 'medical intervention'
    - 'accidental exposure in medical intervention'
    - 'drug administration'
      - 'drug exposure during pregnancy'
      - 'errored drug administration'
        - 'expired drug administration'
        - 'inappropriate schedule of drug administration'
    - 'incorrect dose administraiton '
      - 'underdose administration in medical intervention'
    - vaccination
      - 'errored vaccination'

Figure 4.8 Two groups of vaccine adverse events classified by OAE



TIV

- 'SNOMED Clinical Terms (Version 2010_07_31)'
  - 'Body structure'
    - 'Morphologically altered structure'
      - 'Morphologically abnormal structure'
        - Enlargement
          - Hypertrophy
            - Swelling
  - 'Clinical finding'
    - 'Clinical history and observation findings'
      - 'Finding of balance'
        - 'Impairment of Balance'
      - 'Functional finding'
        - 'Finding of walking'
          - 'Finding of gait'
            - 'abnormal gait'
              - 'High level sensorimotor gait disorder'
                - 'Petren's gait'
        - 'Psychological finding'
          - 'Mental state, behavior, and/or psychosocial function finding'
            - 'Mental state finding'
              - 'Conciousness and/or awareness finding'
                - 'Conciousness related finding'
                  - 'Level of conciousness - finding'
                    - 'Disturbance of conciousness'
                      - 'Decreased level of conciousness'
                        - 'Loss of conciousness'
                          - Syncope
    - Disease
      - 'Acute disease'
        - 'Acute allergic reaction'
          - Anaphylaxis
      - Complication
        - 'Complication of procedure'
          - 'Application AND/OR injection site disorder'
            - 'Injection site disorder'
              - 'Injection site induration'
      - 'Disorder by body site'
        - 'Infection by site'
          - 'Infectious disease of nervous system'
            - 'Neuropathy due to infection'
              - 'Acute infective polyneuritis'
      - 'Disorder characterized by pain'
        - 'Disorder characterized by back pain'
        - 'Injection site pain'
      - 'Disorder characterized by edema'
        - 'Edema of pharynx'
      - Cellulitis
    - Edema
    - Erythema
    - 'Evaluation finding'
      - 'Neuroelectrophysiology finding'
        - 'EMG finding'
          - 'Electromyogram abnormal'
    - 'Finding by site'
      - 'Cardiovascular finding'
        - 'Cardiovascular measurement - finding'
          - 'Blood pressure finding'
            - 'Abnormal blood pressure'
              - 'Finding of increased blood pressure'
        - 'Disorder of cardiovascular system'
          - 'Peripheral vascular disease'
            - 'Vascular disease of the skin'
              - Flushing
          - 'Vascular disorder'
            - 'Disorder of artery'
              - 'Low blood pressure'
                - 'Low blood pressure reading'
      - 'Finding of body region'
        - 'Finding of head and neck region'
          - 'Head finding'
            - 'Mouth and/or pharynx finding'
              - 'Pharyngeal finding'
                - 'Feeling of throat tightness'
        - 'Finding of limb structure'
          - 'Finding of upper limb'
            - 'Finding of shoulder region'
              - 'Shoulder pain'
        - 'Finding of trunk structure'
          - 'Finding of region of thorax'
            - 'Sensation related to thoracic organ'
              - Palpitations
        - 'Skin finding'
          - 'Altered sensation of skin'
            - 'Itching of skin'
          - 'Finding of sweating'
            - Sweating
              - 'Cold sweat'
      - 'Musculoskeletal finding'
        - 'Finding of power of skeletal muscle'
          - 'Muscle weakness'
        - 'Musculoskeletal pain'
    - 'General clinical state finding'
      - 'Body disability AND/OR failure state'
        - Disability
          - 'Walking disability'
    - 'Neurological finding'
      - 'Motor nervous system finding'
        - 'Motor dsyfunction'
          - Paralysis
      - 'Pain / sensation finding'
        - 'Observation of sensation'
          - Hypesthesia
          - Paresthesia
        - Pain
          - 'Finding of present pain intensity'
            - 'Excruciating present pain'
    - 'Wound finding'
      - Wound
        - Contusion
  - 'Special concept'
    - 'Erroneous concept'
      - 'Disorder of balance'
    - 'Inactive concept'
      - 'Duplicate concept'
        - Astasia-abasia
        - Pruritus

LAIV

- 'SNOMED Clinical Terms (Version 2010_07_31)'
  - 'Clinical finding'
    - 'Clinical history and observation findings'
      - 'General finding of observation of patient'
        - 'Drug therapy finding'
          - 'Drug administration observations'
        - 'Influenza-like illness'
    - Disease
      - 'Disorder by body site'
        - 'Disorder of body cavity'
          - 'Disorder of bronchus'
            - Bronchitis
          - 'Disorder of mediastinum'
            - 'Disorder of pericardium'
              - Pericarditis
        - 'Disorder of head'
          - 'Disorder of nose and nasopharynx'
            - 'Disorder of nose'
              - 'Nasal discharge'
                - 'Posterior rhinorrhea'
        - 'Ear, nose and throat disorder'
          - 'Disorder of upper respiratory system'
            - 'Inflammatory disorder of upper respiratory tract'
              - Sinusitis
        - 'Inflammation of specific body structures or tissue'
          - 'Inflammation of specific body organs'
            - Laryngitis
      - 'Infectious disease'
        - 'Viral disease'
          - 'Disease due to Orthomyxoviridae'
            - Influenza
    - 'Drug action'
      - 'Lack of drug action'
    - 'Evaluation finding'
      - 'Imaging finding'
        - 'Imaging result normal'
          - 'Nuclear magnetic resonance normal'
        - 'Radiologic finding'
          - 'Radiology result abnormal'
            - 'Plain X-ray result abnormal'
              - 'Standard chest X-ray abnormal'
    - 'Finding by site'
      - 'Cardiovascular finding'
        - 'Disorder of cardiovascular system'
          - 'Vascular disorder'
            - 'Hemorrhage of blood vessel'
              - Epistaxis
      - 'Digestive system finding'
        - 'Gastrointestinal tract finding'
          - 'Functional finding of gastrointestinal tract'
            - 'Finding of vomiting'
              - Retching
      - 'Ear and auditory finding'
        - 'Ear finding'
          - 'Pain of ear structure'
            - Otalgia
      - 'Finding of body region'
        - 'Finding of head and neck region'
          - 'Head finding'
            - 'Lacrimal system finding'
              - 'Finding of lacrimation'
                - 'Excessive tear production'
                  - 'C/O - excess tears'
            - 'Mouth and/or pharynx finding'
              - 'Pharyngeal finding'
                - 'Disorder of pharynx'
                  - 'Disorder of nasopharynx'
                    - Nasopharyngitis
                - 'Pharyngeal dryness'
          - 'Pain of head and neck region'
            - Headache
            - 'Pain in eye'
            - 'Pain in throat'
        - 'Finding of trunk structure'
          - 'General finding of abdomen'
            - 'Abdominal pain'
              - 'Upper abdominal pain'
      - 'Respiratory finding'
        - 'Disorder of respiratory system'
          - 'Respiratory tract congestion'
          - 'Respiratory tract infection'
            - 'Upper respiratory infection'
        - 'Functional finding of repiratory tract'
          - Cough
          - Sneezing
        - 'Upper respiratory tract finding'
          - 'Nasal congestion'
            - 'C/O nasal congestion'
          - 'Sinus headache'
    - 'Neurological finding'
      - 'Sensory nervous system finding'
        - 'Finding of sense of smell'
          - 'Loss of sense of smell'
            - 'C/O - anosmia'
        - 'Finding of sense of taste'
          - 'Loss of taste'
            - 'C/O - loss of taste sense'
  - Procedure
    - 'Procedure by method'
      - 'Evaluation procedure'
        - Measurement
          - 'Physiologic measurement'
            - 'Metabolic function test'
  - 'Qualifier value'
    - Descriptor
      - Origins
        - 'Accidental exposure'

Figure 4.9 Two groups of vaccine adverse events classified by SNOMED-CT

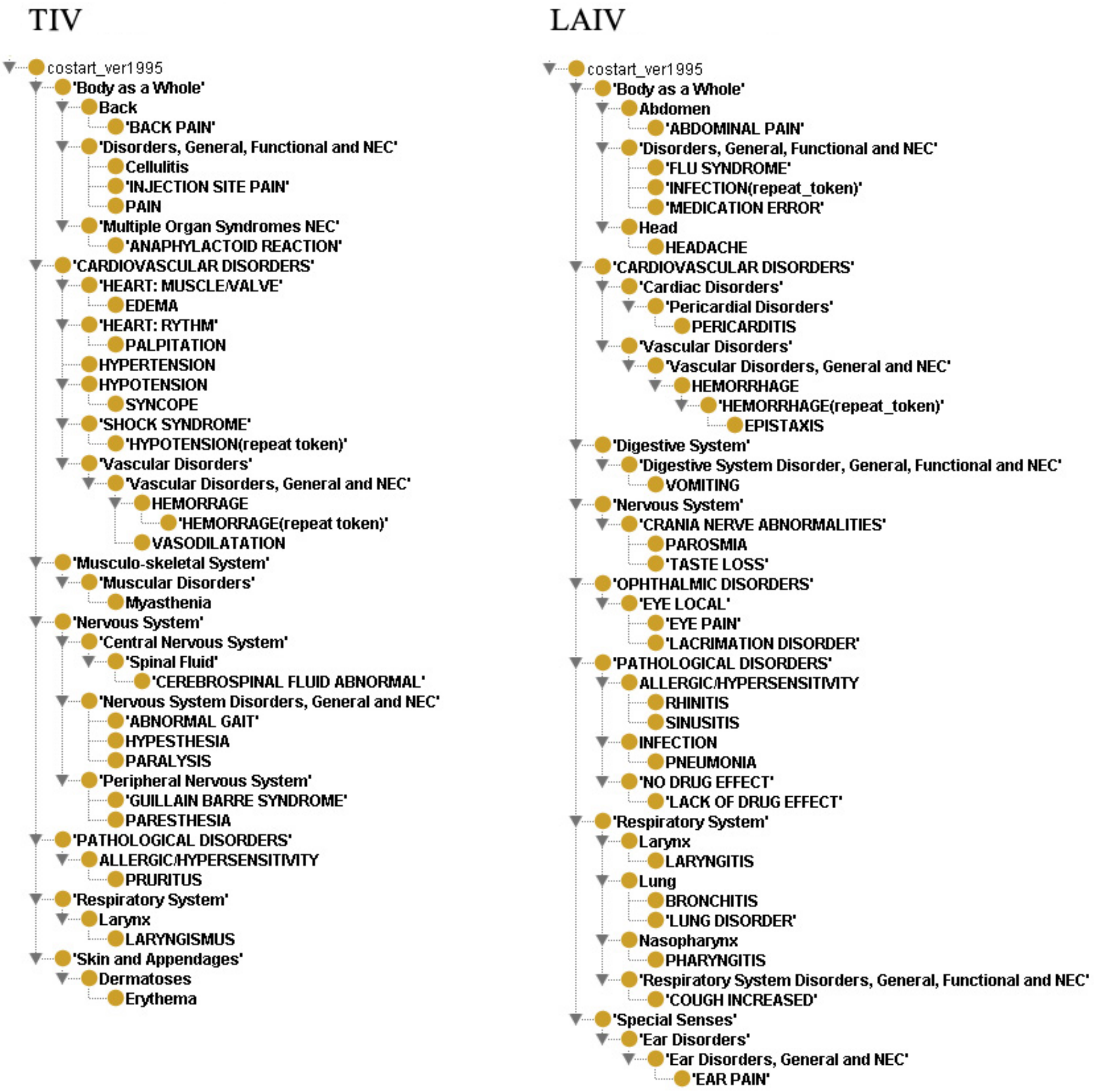


Figure 4.10 Two groups of vaccine adverse events classified by MedDRA

Chapter 5

Future Directions and Conclusions

5.1 Summary of previous sections

In this thesis, I have described the three key aspects of biomedical ontology implementation (motivation, creation, and application) that drive bioinformatics experimental science to clinical translational informatics. In doing so, combinatorial bioinformatics approach was demonstrated via two use cases; cell culture information management, and knowledge discovery from mining influenza vaccine-induced adverse event data. In chapter 2, the foundation for the standardized cell line nomenclature was laid out in the construction of the Cell Line Knowledgebase (CLKB). Cell line entry information from ATCC and the Hyper Cell Line DB was normalized and formatted to fit a central schema. Relations of derivatives were described using a simple ontology structure. The cell line nomenclature analysis from CLKB processing has brought the awareness of challenges in standardized cell line names. Issues of cross-contamination, cell line name mislabeling, and non-centralized authority for such naming system were examined.

In chapter 3, improvement of CLKB to fully implement a complete ontology for cell lines was discussed. CLKB structure was modified to conform with OBO foundry agreement, which then resulted in the development of the Cell Line Ontology (CLO). Components from external OBO-foundry ontologies that could be reused were imported to CLO. Basic Formal Ontology, Relation Ontology, and Information Artifact Ontology were used to outline the upper structure of CLO. Low-level terms were imported from NCBI Taxonomy, Cell Type Ontology, UBERON, and Ontology for Biomedical Investigation. CLO contains over 30,000 cell line entries from CLKB and Coriell Cell Bank. CLO applications in translational informatics were exemplified in information sharing and annotation in the Bioassay Ontology, and integrative data management in the practice of describing experimentally modified cells that linked information of *in vivo* cell types.

The second use case examining adverse events was presented in Chapter 4. The analysis outlined the use of the Ontology of Adverse Events (OAE) in identifying the differential AE profiles in two kinds of influenza vaccines by semantic similarity clustering. Non-normalized adverse event reports were drawn from VAERS database grouped by TIV and LAIV cohorts. The development of OAE hierarchy structure was based on pathological properties referenced by MedDRA, SNOMED CT, and the Mammalian Phenotype Ontology. Statistical and bioinformatics analysis of the VAERS records provided ranked lists of significant AEs for each group. Clustering by OAE classes revealed different sets biological activities in the two groups; nervous system and muscle disorders in TIV, and inflammation in respiratory system in LAIV. Severe adverse events such as Guillain-Barre Syndrome and its related symptoms (e.g. paralysis, myasthenia syndrome) were exclusively enriched in TIV group.

As shown throughout this thesis, integration of different data types to bring experimental and clinical principles together has become a crucial requirement in modern bioinformatics and biomedical research. This results in study that combines both discovery-driven *and* hypothesis-driven methods. Developing bioinformatics tools to build this bridge between the two models should take advantage of the information-enriched structure of biomedical ontologies. Ontology data structure also accommodates and captures the rapidly-growing amount of data. *P4 Medicine* concept[54] (Predictive, Preventive, Personalized, and Participatory) requires the transformation of traditional bioinformatics to translational informatics to achieve a practical health informatics. A framework to map across the current existing biomedical ontologies will help bringing this informatics transformation in place, but it must overcome the issue of ontology structural inconsistencies and incompatibilities[138,139].

The overall study regime of biomedical ontologies for translational informatics that bypasses the stated issues based on the *motivation, creation,* and *application* of such ontologies has resulted in a framework developed in this thesis. Ontology mapping process was based on the hypothesis that by combining the ontological knowledge in the selected set of ontologies, the linkage of information will help answering the translational informatics questions in 1) finding pathological patterns and networks in response to different treatments, 2) identifying the similarities and/or differences in health-related questions arising from point 1) (e.g. adverse event mechanism in vaccine recipients), 3) identifying the underlying mechanism of the gene regulations in such process.

Computational methods in this study have introduced a novel model of data integration that helps gearing toward the first three tiers of *P4 Medicine* (Predictive, Preventive, and Personalized) as this approach will close the gap between experimental informatics and clinical informatics leading to the next step in translational informatics in modern health science.

5.2 Applications to improved health informatics

Digital data have been the trend of modern medicine, and therefore, this calls for a well-designed information system and database management method. Designing a software application that works in such circumstance is crucial. Electronic health record (EHR) is being investigated for its potentials. This replicates the paradigm of biomedical ontology in itself that ontology was first used to *annotate* data, but then it became more apparent of its capability in assisting *computations* and *translations.* EHR was at first used to annotate patient's record in digital text format. But advance computation method has shone light on the possibility that EHR can be implemented to support (semi-)automated processing in diagnostic procedure. Equipping EHR software application with structured data such as OAE (and other clinically oriented ontologies) can provide an effective framework for inputting patient's health information, which can then be automatically linked and processed to identify health-related discovery. Data mining to identify knowledge discovery from bridging phenotypic characteristics to genotypic analysis is a general guideline as shown in Figure 6.1.

5.3 Bundling CLKB/CLO as standardized cell line nomenclature source

CLKB and CLO were one of the first attempts known to process cell line information to a standardized-format collection containing relations between the starting cell line and its derivatives. With a large number of cell line entries, CLO has been indicated as a potential source reference vocabulary for new cell line name creation. As discussed in

Chapter 2 and 3, ambiguous names caused by similarity in cell line labels are one cause of confusion due to mislabeling. Annotating cell lines based on CLO-processed information can help overcome this challenge.

Integrating external information into an ontology is another major advantage in using biomedical ontologies to expand the scope of possible applications. The ATCC Standard Development Organization has proposed the use of short tandem repeat (STR) profiling technique to authenticate cell lines in order to screen for contamination that may occur. Information linked to an ontology does not have to be in a form of classes or instances from another ontology, but it can also be textual information, URL, web object, etc. Therefore, the proposal to link STR profiling ID to its corresponding cell line entry in CLO has been discussed as the movement toward centralizing cell line entry source information. Community support and collaboration engagement are key to a successful process to this centralized source ontology for cell lines.

5.4 Maintaining OAE and CLO

As new discovery and medical inventions are made every day, it is very important to identify a mechanism required to assure that ontologies that we have established stay up to date. This is where collaboration is playing a crucial role. Many biomedical ontologies in use today are built and supported by a community that share the same goal or having their work overlaying each other's. Official websites for OAE and CLO, have been registered at http://www.oae-ontology.org/ and http://www.clo-ontology.org/ respectively. Dr. Yongqun He's lab has taken on the leadership roles of the developer groups. While maintaining the integrity of these works, the implementation of the development process will be shared among the collaborating parties. Automated assistance can also be applied to update any modified existing information, or insert new information from the established sources.

5.5 The future of OAE: OAE as a hypothesis generator

Examining post-influenza-vaccination AEs by OAE classification as described in chapter 4 revealed that, even though it is a rare incident; it was more likely that TIV vaccination would induce serious AEs such as GBS and paralysis when compared to LAIV. The finding prompts us to lay a plan for further analysis of molecular activities that may explain this higher chance of post-vaccination serious AE. As a preliminary investigation, Natural Language Processing was implemented to run the influenza vaccine-induced AE terms against PubMed abstracts to tag the AE-geneSymbol pairs that co-localized in the same sentence. Genes retrieved from this method were then analyzed by ConceptGen to identify the different Gene Ontology's biological processes or functions that were enriched in TIV and LAIV. Comparison of concepts enriched in the two groups reflected the biological activities underlying the different sets of AEs. As severe adverse events were identified exclusively in TIV, gene interactions of those that were mapped to GBS and its related symptoms were modeled using information from the MiMI interaction database. A few intermediate/hub genes were proposed to be key elements in the occurrences of GBS and other related diseases.

5.5.1 Adverse event data extraction and statistical analysis of AEs

The data analysis to extract significant adverse events was performed based on the combinatorial bioinformatics workflow as discussed in section 4.4.1 of Chapter 4. List of enriched adverse events of the two groups of influenza vaccines (TIV and LAIV) were examined. Investigation revealed that severe adverse events were enriched in TIV-induced AE list. Therefore, comprehensive analysis to identify critical biological activities was implemented on the selected severe adverse events and the related symptoms.

5.5.2 Natural language and ontology processing with UMLS

Adverse event report system undergoes the lack of support in nomenclature in VAERS and other general-purpose sources. For example, the terms ‘Fever’, ‘Increased body temperature’, and ‘Pyrexia’ appear in VAERS as different entities and records, while they all refer to the same symptom of adverse event. To make this study inclusive while accommodating for the non-standardized nomenclature issue, we first ran our lists of adverse events against Unified Medical Language System (UMLS, integrated source for terminology, classification, and coding standards) [140] to identify all synonyms for each adverse event and group them together based on their synonymous characteristics.

For each vaccine, we then processed PubMed abstracts (as of September 2009) filtering by ‘human’ or ‘Homo sapiens’ to identify gene names that appeared to be relevant to each adverse event (and its synonyms) by sentence co-localization to identify the pair of AE – gene_symbol and how many times these two components appear together in PubMed abstracts. We hypothesized that if a gene name/symbol appeared in the same sentence as an AE, there may be some relevance among the couple. For a better precision, we decided not to tag the adverse events appearing in the abstract with the human-curated gene list that corresponded to one publication, as adverse event terms can be quite general, and we may end up with irrelevant adverse event-gene tagging that would result in a false positive identification because genes tagged per paper are genes that are significant to the thesis statement of that particular paper, while adverse events appearing in that paper may be just one of many supporting methods in that paper. For each AE (and its synonym) of each vaccine, we examined the sentences that the adverse event appeared together with the presence of a gene symbol. We also mapped each gene symbol to the corresponding gene ID for normalization using programming script with NLP sentence co-localization approach. The lists of genes generated by this process were then sorted by number of occurrences in PubMed sentences to generate the candidate gene list for the gene network reconstruction of the two vaccines studied. Note that some gene symbols appearing in text can be mapped to multiple gene IDs as some gene symbols are synonymous to multiple human genes.

5.5.3 Gene-concept network reconstruction with ConceptGen and MiMI

The ranks of AE – gene_symbol pairs by number of sentences tagged from PubMed were then filtered to meet our criteria in order to minimize false positive. The filtering criteria are composed with a set of rules that 1) gene is in the top five ranking of the examined AE AND number of supporting sentences is greater than or equal to 8 or 2) in the case that some AEs are not as well-represented in PubMed abstracts that the number of supporting sentences for the top five ranking genes is less than 10, manual examination is performed to identify if the gene name tagged is related to the AE. We also include genes that overlap with the curated set of Gene2MeSH [141]. Gene2MeSH searches genes and MeSH terms that are related curated and weighted by statistics. In this case, we run each AE as a MeSH term on Gene2MeSH to identify genes that can be matched to AEs. We then cross-examined the result from Gene2MeSH tagging with our NLP list. For genes that are found in both lists, even if it may not make the criteria of being at the top 5, we include those genes in our analysis if it can be confirmed by Gene2MeSH. There are genes that exist in both gene lists even though these are from different sets of AEs per vaccine. These common 43 genes are eliminated from each list to create a true TIV- and LAIV-specific gene list. There are 130 genes specific to TIV, and 223 genes specific to LAIV.

We then ran TIV- and LAIV-specific gene lists on ConceptGen [142] to identify concept annotation and network that can be mapped by the overlapping of our gene list and the known curated gene set of the annotation. Gene network identified by MiMI[143,144] for the significant concepts were examined. Figure 6.1 summarizes the bioinformatics workflow of this study.

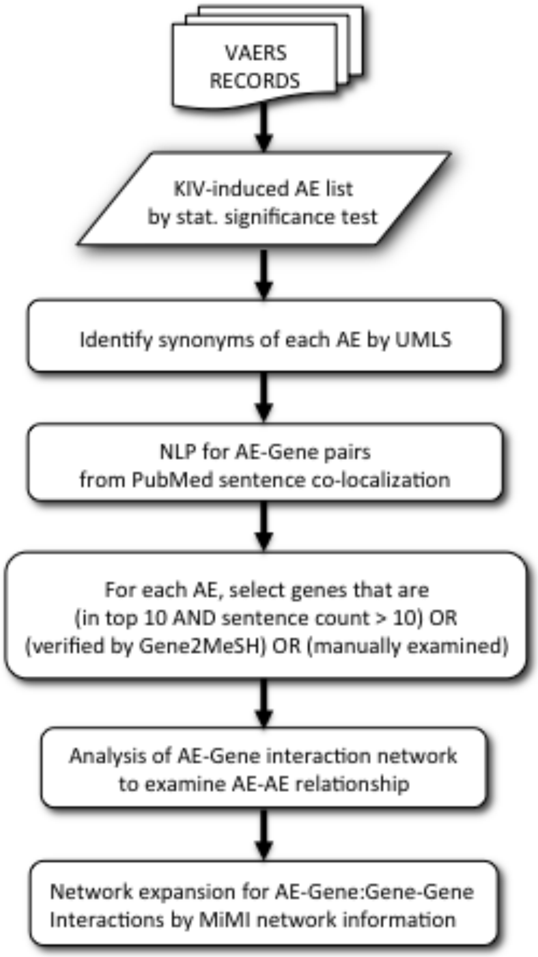

VAERS RECORDS
KIV-induced AE list by stat. significance test
Identify synonyms of each AE by UMLS
NLP for AE-Gene pairs from PubMed sentence co-localization
For each AE, select genes that are (in top 10 AND sentence count > 10) OR (verified by Gene2MeSH) OR (manually examined)
Analysis of AE-Gene interaction network to examine AE-AE relationship
Network expansion for AE-Gene:Gene-Gene Interactions by MiMI network information

Figure 5.1 Summary of the process diagram.
Sentence co-localization identifies AE and gene symbol pairs that occur together at a sentence level in PubMed abstracts. Individual AE-Gene pair counts are compiled together for each vaccine group based on AE terms that came up as specific to each group from Chi-square/PRR rankings. Lists of TIV- or LAIV-specific Chi-square/PRR ranking AEs were also reorganized to hierarchical structure based on biological relevance. TIV- or LAIV-related genes were supplied to concept enrichment analysis in attempt to detect whether gene concept enrichment can explain biological processes identified by hierarchical reorganization, while AE-gene interaction networks were expanded by gene-gene neighboring interaction information from MiMI to model for AEs that may be connected at the molecular level. This framework demonstrates assistance in hypothesis generating from piecing together gene concept enrichment, data hierarchy reorganization, molecular network analysis.

5.5.4 Preliminary results

The Proportional Reporting Ratio (PRR) score and Chi-square significance test P –value provides a criterion to create a ranked list of adverse events triggered by influenza vaccines that satisfy the statistical conditions (Chapter 4 - Methods) to be examined in this study. As will be discussed in the following section, there are severe adverse events that appear in the TIV-triggered adverse event list, and therefore will be the focus of this study. AE terms in the list were supplied to the natural language processing (NLP) workflow to identify coupling AE-gene symbol pairs that appear in literature before using these gene symbols to reconstruct the interaction network that models molecular activities that can potentially explain the occurrences of severe adverse events triggered by TIV influenza vaccines.

5.5.4.1 Statistical analysis of post-influenza-vaccination identified enriched severe adverse events in TIV, but not in LAIV

The analysis of 7,520 AE terms in 616,215 VAERS data records that may individually contain multiple AE listings (data accessed on May18th, 2011) from the statistical analysis of significant AEs that are induced by killed-inactivated influenza vaccines and live-attenuated influenza vaccine has revealed distinct sets of AEs among the two groups that are induced in different biological process clusters (Chapter 4). Further investigation of these two lists disclosed the knowledge that severe adverse events (SAEs) such as paralysis and anaphylactic reaction were present as statistically significant exclusively in TIV group as confirmed by the comparative reporting ration score and P-value in the same study. The evidence that SAEs of Guillain-Barre Syndrome, paralysis, and anaphylactic reaction are significantly more likely to be triggered by TIV in comparison to LAIV has prompted us to focus on understanding the underlying biological process that triggers these SAEs in TIV.

We hypothesized that examining these AEs based on their score rankings and gene network analysis can potentially identify an underlying system that explains the occurrence of symptoms and the connection between AEs. This hypothesis led to further analysis of molecular interaction network analysis using the workflow as summarized in the previous section. Also, interestingly, ranked list of AEs shows agreement of the occurrence of post-vaccination Guillian-Barré Syndrome (GBS) in killed inactivated Influenza vaccine recipients with a number of studies that have previously indicated (Chapter 4 – Discussion). Post-vaccination GBS is ranked among highly significant AEs in all scoring matrices of TIV group. However, the information retrieved via this study has shown no evidence of GBS in LAIV group at all. Further study to verify the relevance of post-vaccination GBS in TIV group as oppose to LAIV group should be carefully investigated.

5.5.4.2 Natural language processing of PubMed abstracts extracted gene-AE pairs of TIV-specific significant AEs

For each TIV-specific AE that was processed and mapped for its synonyms by UMLS, it was run against PubMed sentences for a co-localization with gene symbol (with tagged curated human geneID). The method resulted in a list of geneIDs (as gene symbols may be ambiguous) associated with an AE. Table 6.1 summarizes the AE-geneID pairs retrieved from PubMed tagging. Note that not every significant AE can be mapped to geneIDs as there may not be any studies for that particular AE. Over- and under-represented AEs in the biomedical research domain is a factor that contributes to this issue. Therefore, we have manually examined the results as described in the method.

Table 5.1 Summary of TIV-specific AE-geneID association by PubMed sentence co-localization.

Adverse Event	**Genes in use (by GeneID)**
Anxiety	6532, 594857, 4128, 1312, 3350, 7166, 627, 135, 121278, 1576, 1813, 672, 338, 4852, 1393
Arrhythmia	6331, 3757, 1565, 10021, 6546
Arthralgia	3077, 114548
Body temperature increased (Gene2Mesh query 'fever')	4210, 7132, 4598, 6288, 114548, 9051, 7124, 7442, 3106, 55733
Dyskinesia	1814, 1813, 1565, 6648, 3356, 1544, 1815, 1312
Dysphagia	59330, 129685, 3483, 6647, 8106
Dyspnoea	4879
Eye Pruritus	260328
Gait disturbance	3897, 26278, 5428, 7436, 497656, 619510
Glomerulonephritis	8710, 7356, 1636, 183, 60498, 185, 4036, 3075, 7040, 4868, 4358, 6402, 1585
Guillain-Barre syndrome	913, 909
Herpes zoster	3054, 7098
Hypoaesthesia	10225, 7974
Jaundice	1244, 54658, 2539, 54600
Lymphadenopathy	6392, 4838, 920, 83482, 129685
Mobility decreased	231
Muscle spasms	170302, 6792, 192, 6323
Muscular weakness	1756, 1760, 8972, 6606, 6647, 3483
Musculoskeletal pain	1436, 7555, 889, 55819, 3929, 1962
Neuropathy peripheral	5376, 4359, 2705, 9927, 4747, 57716,54332
Ocular hyperaemia	3569, 3586, 3458, 7124

Oedema	361, 1636, 7422, 358
Oedema peripheral	3630, 213, 7124
Palpitations	5710, 29855, 129685, 4838, 9961, 50951
Paraesthesia	57498, 1507, 192142, 129685, 3918
Pharyngeal oedema	450095
Pruritus generalised	3497, 1798, 2875
Rash papular	55837
Respiratory distress	6439, 21, 6435, 653509, 6440, 1636, 3508, 6436
Sepsis	7124, 54210, 929, 3569, 3586, 3146, 7099, 796, 133, 3929, 11093, 4153, 4049, 2920, 2152, 4282, 10544
Sinus congestion	474168, 51364
Swelling	292, 38, 129685, 6910, 3183
Syncope	6331, 3757, 6262, 3784, 3753
Tachycardia	6262, 845, 6331, 6530, 153, 154, 3759, 2281
Tendonitis	1507, 129685, 23481, 38, 292
Varicella	3918, 5788, 959, 1991, 129685
Vertigo	773, 619536
Visual impairment	6103, 1406, 4643, 6102, 24, 3000, 3075, 285440, 1121, 5959, 7399, 6121, 6247, 145226

Reconstruction of gene interaction network based on MiMI interaction database resulted in the overall network as shown in figure 6.2. Nodes displayed in red signify AEs that contains intermediate connections to other AEs/genes. Yellow nodes are AEs that do not connect to other AEs.

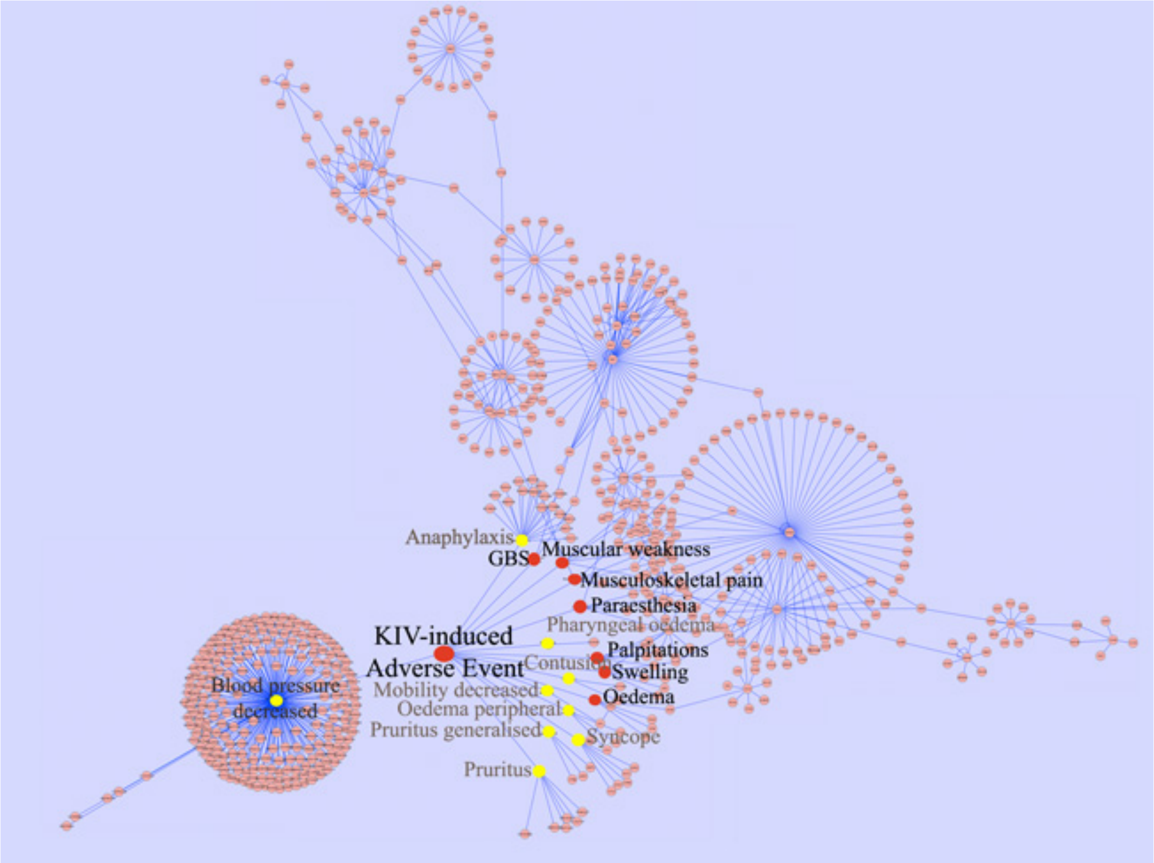


Figure 5.2 Overall gene interaction network of TIV-specific associated genes. Gene-AE network derived from TIV-specific adverse event analysis. Signified in red are adverse events detected to have some gene interactions that may indicate genes that trigger more than one AEs, or a sub-network that underlies a biological process response to TIV vaccines (high-resolution image available in online supplementary materials)

5.5.4.3 Gene concept analysis revealed different enriched concepts with differential VAEs

Enriched biological concepts identified by reconstructing NLP-processed gene-concept network reveals different significant biological processes between the two groups. The functional annotations of Gene-concept network reconstruction by *ConceptGen/MiMI* resulted in concepts identified by various sources (Biocarta, KEGG, Gene Ontology (GO), PubChem, DrugBank, etc). In this study, we examine GO Biological Processes in attempt to identify the underlying machinery of each vaccine. Interestingly, TIV data show significant concepts related to Neurological System such as Neurological System Process, and Sensory Perception, while these concepts are insignificant in LAIV data set. This may lead to explanation how *a number of motor sensory disorders (including GBS)* surface as serious AEs in inactivated-virus Influenza vaccines [111,136]). We cross-examined these concepts on both *ConceptGen* (http://conceptgen.ncibi.org/) [142] and *DAVID Bioinformatics* Tools (http://david.abcc.ncifcrf.gov/) [145]. The results from both sources are in accord with each other. LAIV data, on the other hand, show no significance in Neurological System Processing as part of the vaccine pathogen mechanism. But External Stimulus Response and Inflammatory Response are ranked significant in LAIV dataset.

The number of overlapping genes between our TIV genes result and ConceptGen's genes per each concept identified a general theme concept of central nervous system and muscle contraction system as significant concepts (with modified Fisher's exact test P-Value and enrichment testing Q-Value using FDR Benjamini method at significant level of Q-Value smaller than 0.05) (Table 6.2). The cross examination with gene concept annotations from DAVID Bioinformatics suite revealed that at the corrected Bonferroni P-Value cut-off of 0.05 cardiac processes (heart contraction, heart process, regulation of heart contraction), muscular processes (muscle system process, muscle contraction) and central nervous system processes (neurological system process, transmission of nerve impulse, sensory perception of light stimulus, visual perception, synaptic transmission, and sensory perception) are exclusively significant in TIV when compared to LAIV cohort. It should be noted that this finding is in accordance with many previous studies indicating a correlation between Influenza and central nervous system dysfunction [146,147]. It is notable that there are no serious neurological-system adverse events on our list of AEs to be studied, but our method was able to identify these underlying neurological processes. This finding may also offer explanation to other more serious vaccine-induced neurological AEs such as Guillain-Barr*é* syndrome by generating hypothesis to be tested for further study (please see discussion – expanding gene network to model gene interactions in GBS).

Table 5.2 Enriched gene concepts (GO function or process) associated with TIV- and LAIV-induced VAEs

TIV Enriched Concepts	P-Value	Q-Value	LAIV Enriched Concepts	P-Value	Q-Value
sensory perception	3.34E-05	0.0041	immune system process	1.1E-14	4.54E-12
transmission of nerve impulse	1.31E-07	2.96E-05	immune response	2.43E-12	8.61E-10
sensory perception of light stimulus	4.2E-07	8.66E-05	defense response	6.77E-18	8.39E-15
synaptic transmission	6.9E-06	1.01E-03	response to wounding	2.12E-17	1.31E-14
response to external stimulus	2.88E-05	3.76E-03	inflammatory response	1.96E-17	1.31E-14
response to chemical stimulus	0.00042	0.027	programmed cell death	2.27E-05	1.06E-03
muscle contraction	2.34E-08	7.24E-06	death	5.9E-05	0.0023
muscle system process	2.34E-08	7.24E-06	cell death	5.9E-05	0.0023
homeostatic process	7.51E-08	1.86E-05	regulation of apoptosis	1.37E-05	7.22E-04
cellular homeostasis	4.12E-06	0.00064	cell proliferation	2.01E-10	4.99E-08
cell-cell signaling	2.25E-06	3.71E-04	regulation of cell proliferation	1.66E-08	3.17E-06
behavior	6.02E-09	2.48E-06	cellular biosynthetic process	0.0011	2.83E-02
blood circulation	7.13E-12	1.77E-08	response to external stimulus	3.53E-21	8.75E-18
circulatory system process	7.13E-12	1.77E-08	response to chemical stimulus	3.99E-17	1.98E-14
visual perception	4.2E-07	8.66E-05	cell-cell signaling	3.07E-11	8.44E-09

Whilst TIV gene result suggests a strong correlation between TIV-induced AEs and central nervous system processes, LAIV gene concept annotations indicates the implication of cell-death regulations and external stimulus processes on LAIV-induced AEs. The smallest P-Value and Q-Value and the highest number of LAIV genes overlapping with concept genes in ConceptGen ranked response to external stimulus at the top of the list of significant LAIV-specific biological processes. Immune response and cell regulations are also suggested as significant LAIV-induced biological processes. With adjusted Bonferroni P-value of 0.05 or smaller, response to external stimulus is also listed as the most significant concept in DAVID gene functional annotations. DAVID result can also be clustered into three major functional groups; response to stimulus (response to stimulus, response to external stimulus, response to chemical stimulus), inflammatory and immune responses (defense response, response to wounding, inflammatory response, immune system process, immune system development), and cell regulations (cell proliferation, cell motility, death, cell death, regulations of programmed cell death and apoptosis). The strong correlation between LAIV-induced AEs and the annotated concepts may be explained by the nature of LAIV being a live vaccine, and therefore LAIV may trigger vital immune and cell regulation responses.

In conclusion, gene-concept enrichment analysis has demonstrated great correspondence with the processes in biological systems highlighted by semantic similarity clustering as discussed in the previous section.

5.5.4.4 **Gene network analysis predicted candidate genes underlying related molecular activities of severe adverse events in TIV**

Severe adverse events with association to nervous system and motor coordination enriched in TIV-specific AE set include paralysis, GBS, and myasthenia. Identifying the overall interactions in the network depicted in figure 6.2 led to the attempt to propose sub-network activities among those GBS-related AEs that connect to each other. Using NLP method revealed molecular interactions with references to public database (MiMI) in order to model a gene network that may explain the underlying molecular activities responding to TIV vaccines as shown in figure 6.3. The model was constructed based on the interaction information as listed in table 6.3.

Table 5.3 Summary of GBS related AE molecular network reconstruction data (summarized from Cytoscape input file)

AE or GeneID	**Relation**	**Associated component**
Syncope	is_a	Adverse event
Swelling	is_a	Adverse event
Pruritus generalised	is_a	Adverse event
Pruritus	is_a	Adverse event
Pharyngeal oedema	is_a	Adverse event
Paraesthesia	is_a	Adverse event
Palpitations	is_a	Adverse event
Oedema peripheral	is_a	Adverse event
Oedema	is_a	Adverse event
Musculoskeletal pain	is_a	Adverse event
Muscular weakness	is_a	Adverse event
Mobility decreased	is_a	Adverse event
Guillain-Barre syndrome	is_a	Adverse event

Anaphylaxis	is_a	Adverse event
18663330	relates_to	Pruritus
18094266	relates_to	Contusion
17081716	relates_to	Anaphylaxis
17024848	relates_to	Anaphylaxis
15985820	relates_to	Anaphylaxis
15753886	relates_to	Anaphylaxis
15025389	relates_to	Anaphylaxis
14616362	relates_to	Pruritus
11422133	relates_to	Anaphylaxis
11132737	relates_to	Pruritus
10535881	relates_to	Pruritus
10321563	relates_to	Pruritus
9571942	relates_to	Contusion
8686958	relates_to	Anaphylaxis
8623140	relates_to	Anaphylaxis
8278629	relates_to	Anaphylaxis
8032237	relates_to	Anaphylaxis
7755201	relates_to	Anaphylaxis
7596227	relates_to	Pruritus
7484432	relates_to	Anaphylaxis
6808133	relates_to	Anaphylaxis
3471098	relates_to	Anaphylaxis
2871754	relates_to	Anaphylaxis
2653575	relates_to	Anaphylaxis
793410	relates_to	Anaphylaxis
450095	relates_to	Pharyngeal oedema
192142	relates_to	Paraesthesia

129685	interacts_with	6872, 6881, 6882, 6883, 6873, 6874, 6877, 6878, 6879, 6889
129685	relates_to	Palpitations, Paraesthesia, Swelling, Palpitations
57716	interacts_with	7430, 1821, 7430, 5906
57498	relates_to	Paraesthesia
57498	interacts_with	4916, 10093, 2261, 6326, 7046, 26140
55819	relates_to	Musculoskeletal pain
50951	relates_to	Palpitations
29855	relates_to	Palpitations
26278	interacts_with	81, 9463, 81628, 30834
9961	relates_to	Palpitations
8972	relates_to	Muscular weakness
8972	interacts_with	4143, 279, 4143, 8972, 4255, 653361, 654817, 5329, 6476
8972	relates_to	Muscular weakness
7555	interacts_with	10236, 27350, 55299, 7818, 1660, 10969, 2079, 3183, 10236, 220988, 3182, 11100, 10642, 10643, 5813, 27316, 6119, 6147, 6122, 6160, 6164, 6165, 25873, 6217, 6223, 6189, 26156, 10492
7555	relates_to	Musculoskeletal pain
7555	interacts_with	10492, 1660, 3183, 6189, 3183, 6189
7436	interacts_with	348, 1191, 1600, 4023, 4043, 5328, 5649, 5054, 9784
7422	relates_to	Oedema
7124	relates_to	Oedema peripheral
6910	relates_to	Swelling
6648	interacts_with	6790, 51608, 84064, 57670, 79159, 5867, 6189
6647	interacts_with	596
6647	relates_to	Muscular weakness
6606	interacts_with	596, 10236, 7917, 10492
6606	relates_to	Muscular weakness

6606	interacts_with	7917, 596, 55791, 1207, 8161, 10980, 11218, 1660, 2091, 8880, 2661, 50628, 25929, 79833, 79760, 10236, 3192, 3609, 3837, 114823, 27257, 25804, 23658, 11157, 51690, 4686, 22916, 54433, 26578, 54433, 26578, 5430, 10248, 10419, 51808, 51639, 8487, 6606, 6607, 6628, 6632, 6633, 6634, 6635, 6636, 6637, 10073, 6813, 6814, 10492, 96764, 073, 7157, 9094, 79084, 9406, 1660
6331	relates_to	Syncope
6262	relates_to	Syncope
5710	relates_to	Palpitations
5428	interacts_with	5428, 11232, 6742
5376	interacts_with	821, 4359, 5824
4838	relates_to	Palpitations
4747	interacts_with	29117, 55755, 1107, 2902, 8988, 8898, 4644, 4744, 4747, 51588, 5585, 5901, 6709, 7248, 7431, 7532
4359	interacts_with	54984, 5376, 335, 820, 929, 3078, 3929, 6227
3929	relates_to	Musculoskeletal pain
3918	interacts_with	649, 1294, 2199, 3914, 4313, 4811
3918	relates_to	Paraesthesia
3897	interacts_with	286, 287, 160, 1173, 1272, 6900, 1457, 7430, 3678, 3685, 4478, 4684, 1463, 8650, 8682, 5621, 10048, 5962, 6195, 6196, 7430, 5962
3784	relates_to	Syncope
3757	relates_to	Syncope
3753	relates_to	Syncope
3630	relates_to	Oedema peripheral
3497	relates_to	Pruritus generalised
3483	relates_to	Muscular weakness
3483	interacts_with	3479, 3485, 3488
3356	interacts_with	56899, 1742, 8665, 2752, 2767, 2770, 2776, 3717, 4130, 10573, 4716, 4832, 4916, 5445, 5536, 6197, 4916, 2770

3183	relates_to	Swelling
2875	relates_to	Pruritus generalised
2705	interacts_with	6714, 801, 2705, 2706, 4950, 5566, 5578, 6714, 801
1962	interacts_with	6342
1962	relates_to	Musculoskeletal pain
1815	interacts_with	54102, 4690, 2885, 6714, 54102, 2554, 2885, 3765, 4690, 6714, 3765
1814	interacts_with	54102, 2316, 4690, 2036, 23413, 54102, 1933, 1937, 2035, 2036, 2037, 2316, 23413, 10755, 2770, 2885, 8777, 4690, 5962, 2885, 10755, 5962, 2035, 2770
1813	interacts_with	54102, 8618, 2316, 2036, 23413, 10755, 135, 8618, 93664, 801, 54102, 1813, 2035, 2036, 2316, 23413, 10755, 2771, 2773, 2781, 2890, 2891, 3763, 3765, 4905, 5074, 84687, 6755, 3765, 93664, 2035, 801
1798	relates_to	Pruritus generalised
1760	relates_to	Muscular weakness
1760	interacts_with	1822, 6310, 10658, 5348, 11337, 3316, 5350, 4659, 5894, 5916, 6667, 56893
1760	relates_to	Muscular weakness
1756	interacts_with	8618
1756	relates_to	Muscular weakness
1756	interacts_with	70, 8618, 93664, 1605, 8525, 27185, 1756, 1821, 1837, 1838, 356, 3768, 3761, 3889, 3856, 5239, 137868, 6640, 6641, 6645, 54212, 54221, 7402, 1821, 93664
1636	relates_to	Oedema
1565	interacts_with	54658, 1660, 54578, 114, 10978, 54205, 1576, 4519, 1660, 2709, 5255, 8858, 5699, 54648, 54578
1544	interacts_with	54658, 54578, 522, 5447, 7172
1507	relates_to	Paraesthesia
1436	interacts_with	2885, 867, 1435, 1436, 2534, 9402, 2885, 3635, 3636, 4067, 5295, 5296, 5921, 6464, 8651, 9021,

		6654, 8563, 7525
1436	relates_to	Musculoskeletal pain
1312	interacts_with	1636, 4143, 1636, 191, 1208, 9516, 4143, 4144, 5720, 9319, 22803
925	interacts_with	926, 909, 919, 915, 916, 917, 925, 926, 3105, 3106, 3107, 3133, 3134, 3135, 27040, 3932, 3956, 5788, 6426, 140890, 25942, 6955, 6957, 28639
913	relates_to	Guillain-Barre syndrome
913	interacts_with	567
913	relates_to	Guillain-Barre syndrome
910	interacts_with	909, 5660
909	relates_to	Guillain-Barre syndrome
909	interacts_with	567, 910, 911, 925
889	interacts_with	5906, 2130, 9270, 5906, 5970
889	relates_to	Musculoskeletal pain
567	interacts_with	926, 925, 3133, 3107, 7917, 919, 3105, 2, 162, 164, 8907, 10053, 1174, 8905, 130340, 57, 7917, 811, 11126, 909, 910, 912, 919, 916, 917, 11033, 136227, 2217, 53826, 3077, 3106, 3107, 3133, 3134, 3135, 3309, 3803, 3804, 3805, 3806, 3807, 3811, 3812, 3824, 11024, 10859, 10288, 55690, 11087, 6892, 6955, 6957, 28639, 7305, 7414, 910,
361	relates_to	Oedema
358	relates_to	Oedema
292	relates_to	Swelling
231	interacts_with	4086, 7030
231	relates_to	Mobility decreased
213	relates_to	Oedema peripheral
38	relates_to	Swelling

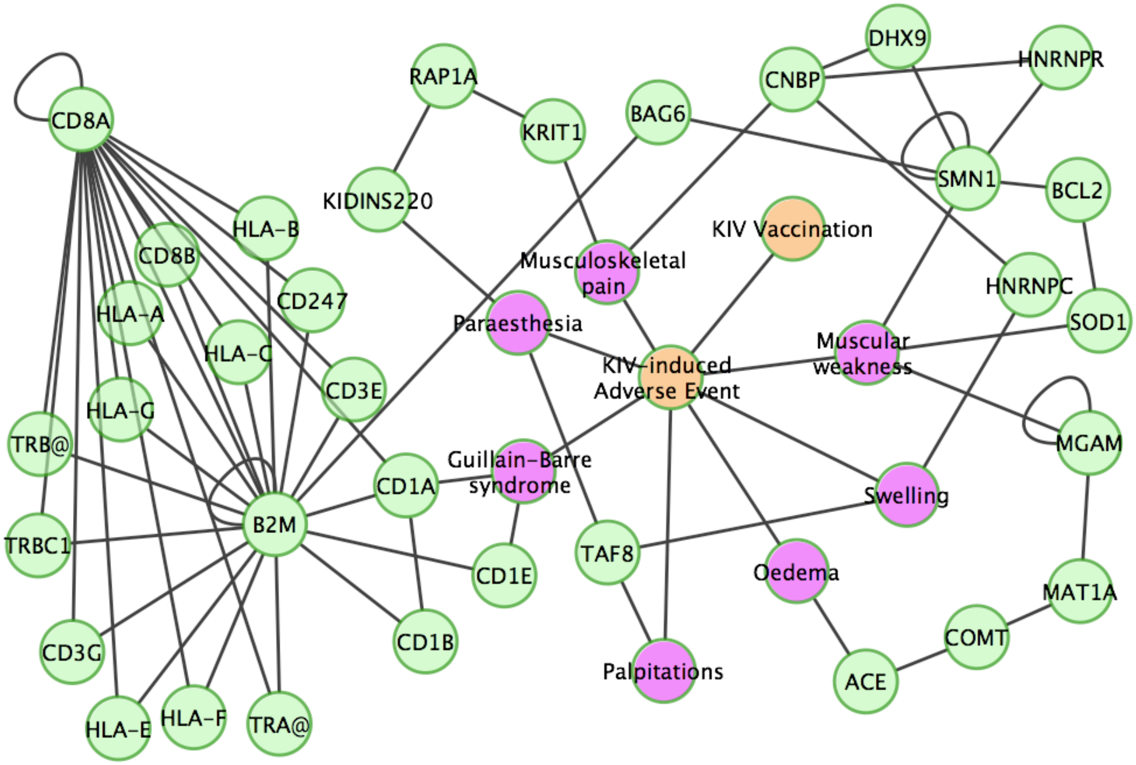


Figure 5.3 TIV-induced GBS-related gene activities proposed by ConceptGen/MiMI interactions.
GBS as linked to muscular weakness indicated the potential subnetwork "muscular weakness-SMN1-BCG6-B2M-CD1-GBS" as a key molecular cluster. CD1 is known for its association with GBS. However, other genes, like B2M (associated with CD8+ T cells) are also important as connected with confirmed interaction data.

Some of the hub or intermediate genes as shown in figure 6.3 have been indicated in literature of their association to biological functions such as; B2M with its mutation that would result in hypercatabolic hyperproteinemia, SMN1 with its mutation that was associated with spinal muscular atrophy, and ACE that functioned in conversion of angiotensin I to peptide angiotensin II which is a potent vasopressor and aldosterone-stimulating peptide controlling blood pressure and fluid-electrolyte balance.

In our investigation of gene interaction of severe adverse events, it does not matter that CD1A and CD1E in association to GBS may contradict with each other in the published literature as both genes interact with B2M which is a hub gene connecting to other genes within the same sub-network. There was, however, SMN1 that showed up in the result with the connection to GBS via molecular interactions. We then explored the published information of both hub genes and found out that B2M encodes a serum protein associated with major histocompatibility complex (MHS) class I heavy chain on the suface of nearly all nucleated cells with its mutation triggering hypercatabolic hypoproteinemia (http://www.ncbi.nlm.nih.gov/gene/567)[148,149]. Mutations in telomeric copy of SMN1 were associated with spinal muscular atrophy (while mutations in centromeric copy of SMN1 did not lead to disease, but may be a modifier of disease caused by mutation in the telomeric copy) (http://www.ncbi.nlm.nih.gov/gene/6606) [150]. ACE functions in encoding an enzyme that catalyzes the conversion of angiotensin I into an active form of peptide angiotensin II. Angiotensin II is a potent vasopressor and aldosterone-stimulating peptide that controls blood pressure and fluid-electrolyte balance. This enzyme plays a key role in the renin-angiotensin system. Many studies have associated the presence or absence of a 287 bp Alu repeat element in this gene with the levels of circulating enzyme or cardiovascular pathophysiologies (http://www.ncbi.nlm.nih.gov/gene/1636)[151,152].

5.5.5 OAE as a potential hypothesis generator

The combinatorial bioinformatics framework with NLP processing as shown in this study helps generating hypothesis. Gene interactions demonstrated in figure 5.2 and 5.3 propose a model of activities that may be underlying the process of Influenza TIV-induced GBS. The pathways identified with this method depict symptoms in nervous and muscular system and probable cause in immune system, which has been hypothesized in many studies as discussed in chapter 4. Pathways and processes suggested by gene interactions include CD8 T-Cell process, metabolisms of various enzymes and sugars, muscle and heart contraction, potassium and calcium transports and nervous system development. Verification of this hypothesized Influenza TIV-induced GBS gene interactions by laboratory protocols is encouraged for further investigation. We have described here the future work plan based on preliminary results that OAE with other bioinformatics approaches construct a hypothesis generator that can suggest a potential hypothesis for further translational informatics investigation.

5.6 Conclusions: Ontology services to bridge to translational research

As described throughout this thesis, we have demonstrated the use of biomedical ontology to drive from basic bioinformatics research towards translational research by creating and utilizing CLO and OAE. Integrative approaches combining bioinformatics methods to biomedical ontologies have been validated of the capability to overcome challenges in multidisciplinary practice in this thesis. The framework of multi-ontology integration in CLO case study, and the combinatorial bioinformatics workflow in OAE case study have depicted the potentials of ontology services to translational informatics. The novel combinatorial framework as shown in chapter 2-4 can be adapted and modified to answer other translational research questions.

References

1. Sagiv Y, Yannakakis M (1980) Equivalences among relational expressions with the union and difference operator JACM 27: 633-655.
2. Zhang S, Bodenreider O (2007) Experience in Aligning Anatomical Ontologies. International journal on Semantic Web and information systems 3: 1-26.
3. Burgun A, Mougin F, Bodenreider O (2009) Two approaches to integrating phenotype and clinical information. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2009: 75-79.
4. Tenenbaum JD, Whetzel PL, Anderson K, Borromeo CD, Dinov ID, et al. (2011) The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. Journal of biomedical informatics 44: 137-145.
5. Bodenreider O (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of medical informatics: 67-79.
6. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ (2004) The NLM Indexing Initiative's Medical Text Indexer. Studies in health technology and informatics 107: 268-272.
7. Sarntivijai S, Ade AS, Athey BD, States DJ (2008) A bioinformatics analysis of the cell line nomenclature. Bioinformatics 24: 2760-2766.
8. Zhang D, Roderer NK, Huang G, Zhao X (2006) Developing a UMLS-based indexing tool for health science repository system. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium: 1157.
9. Ozgur A, Xiang Z, Radev DR, He Y (2011) Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. Journal of biomedical semantics 2 Suppl 2: S8.
10. Dickson S, Pouchard L, Ward R, Atkins G, Cole M, et al. (2005) Linking human anatomy to knowledge bases: a visual front end for electronic medical records. Studies in health technology and informatics 111: 586-591.
11. Blaschke C, Leon EA, Krallinger M, Valencia A (2005) Evaluation of BioCreAtIvE assessment of task 2. BMC bioinformatics 6 Suppl 1: S16.
12. Couto FM, Silva MJ, Lee V, Dimmer E, Camon E, et al. (2006) GOAnnotator: linking protein GO annotations to evidence text. Journal of biomedical discovery and collaboration 1: 19.
13. Crangle CE, Zbyslaw A (2004) Identifying gene ontology concepts in natural-language text. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference 4: 2821-2823.
14. Daraselia N, Yuryev A, Egorov S, Mazo I, Ispolatov I (2007) Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC bioinformatics 8: 243.

15. Srinivasan P, Qiu XY (2007) GO for gene documents. BMC bioinformatics 8 Suppl 9: S3.
16. Hersh WR, Greenes RA (1990) Information retrieval in medicine: state of the art. MD computing : computers in medical practice 7: 302-311.
17. Brandt C, Nadkarni P (2001) Web-based UMLS concept retrieval by automatic text scanning: a comparison of two methods. Computer methods and programs in biomedicine 64: 37-43.
18. Lowe CI, Wright JL, Bearn DR (2001) Computer-aided Learning (CAL): an effective way to teach the Index of Orthodontic Treatment Need (IOTN)? Journal of orthodontics 28: 307-311.
19. Ruiz JG, Mintzer MJ, Issenberg SB (2006) Learning objects in medical education. Medical teacher 28: 599-605.
20. Whetzel PL, Parkinson H, Stoeckert CJ, Jr. (2006) Using ontologies to annotate microarray experiments. Methods in enzymology 411: 325-339.
21. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, et al. (2006) The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 22: 866-873.
22. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic acids research 35: D747-750.
23. Kato K, Yamashita R, Matoba R, Monden M, Noguchi S, et al. (2005) Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. Nucleic acids research 33: D533-536.
24. Cole CL, Kanter AS, Cummens M, Vostinar S, Naeymi-Rad F (2004) Using a terminology server and consumer search phrases to help patients find physicians with particular expertise. Studies in health technology and informatics 107: 492-496.
25. Lee M, Wang W, Yu H (2006) Exploring supervised and unsupervised methods to detect topics in biomedical text. BMC bioinformatics 7: 140.
26. Bodenreider O, Burgun A (2004) Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. Studies in health technology and informatics 107: 327-331.
27. Bodenreider O, Zhang S (2006) Comparing the representation of anatomy in the FMA and SNOMED CT. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium: 46-50.
28. Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome biology 6: R21.
29. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic acids research 39: D507-513.
30. McDonald CJ, Overhage JM, Dexter P, Takesue BY, Dwyer DM (1997) A framework for capturing clinical data sets from computerized sources. Annals of internal medicine 127: 675-682.

31. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, et al. (2007) The SAGE Guideline Model: achievements and overview. Journal of the American Medical Informatics Association : JAMIA 14: 589-598.
32. Bouhaddou O, Warnekar P, Parrish F, Do N, Mandel J, et al. (2008) Exchange of computable patient data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): terminology mediation strategy. Journal of the American Medical Informatics Association : JAMIA 15: 174-183.
33. Kahn CE, Jr., Santos A, Thao C, Rock JJ, Nagy PG, et al. (2007) A presentation system for just-in-time learning in radiology. Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology 20: 6-16.
34. Dolin RH, Alschuler L, Behlen F, Biron PV, Boyer S, et al. (1999) HL7 document patient record architecture: an XML document architecture based on a shared information model. Proceedings / AMIA Annual Symposium AMIA Symposium: 52-56.
35. Lee Y, Supekar K, Geller J (2006) Ontology integration: experience with medical terminologies. Computers in biology and medicine 36: 893-919.
36. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, et al. (2008) caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. Journal of biomedical informatics 41: 106-123.
37. Fridsma DB, Evans J, Hastak S, Mead CN (2008) The BRIDG project: a technical report. Journal of the American Medical Informatics Association : JAMIA 15: 130-137.
38. Blake JA, Bult CJ (2006) Beyond the data deluge: data integration and bio-ontologies. Journal of biomedical informatics 39: 314-320.
39. Joubert M, Dufour JC, Aymard S, Falco L, Fieschi M (2005) Designing and implementing health data and information providers. International journal of medical informatics 74: 133-140.
40. Bergeron E, Simons R, Linton C, Yang F, Tallon JM, et al. (2007) Canadian benchmarks in trauma. The Journal of trauma 62: 491-497.
41. Lieberman MI, Ricciardi TN, Masarie FE, Spackman KA (2003) The use of SNOMED CT simplifies querying of a clinical data warehouse. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium: 910.
42. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275-1283.
43. Wolting C, McGlade CJ, Tritchler D (2006) Cluster analysis of protein array results via similarity of Gene Ontology annotation. BMC bioinformatics 7: 338.
44. Rubin DL, Dameron O, Bashir Y, Grossman D, Dev P, et al. (2006) Using ontologies linked with geometric models to reason about penetrating injuries. Artificial intelligence in medicine 37: 167-176.
45. Knapp S What's in a name? A history of taxonomy.
46. Darwin C, Carroll J (2003) On the origin of species by means of natural selection: Broadview Press.
47. Wikipedia (2011) Metaphysics (Aristotle).

48. W3C (2004) OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004.
49. Weiss A, Wiskocil RL, Stobo JD (1984) The role of T3 surface molecules in the activation of human T cells: a two-stimulus requirement for IL 2 production reflects events occurring at a pre-translational level. Journal of immunology 133: 123-128.
50. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 25: 1251-1255.
51. Hoehndorf R, Oellrich A, Dumontier M, Kelso J, Rebholz-Schuhmann D, et al. (2010) Relations as patterns: bridging the gap between OBO and OWL. BMC bioinformatics 11: 441.
52. Blonde W, Mironov V, Venkatesan A, Antezana E, De Baets B, et al. (2011) Reasoning with bio-ontologies: using relational closure rules to enable practical querying. Bioinformatics 27: 1562-1568.
53. Ceusters W, Smith B (2010) A unified framework for biomedical terminologies and ontologies. Studies in health technology and informatics 160: 1050-1054.
54. Bradley WG, Golding SG, Herold CJ, Hricak H, Krestin GP, et al. (2011) Globalization of P4 medicine: predictive, personalized, preemptive, and participatory--summary of the proceedings of the Eighth International Symposium of the International Society for Strategic Studies in Radiology, August 27-29, 2009. Radiology 258: 571-582.
55. Intel Excerpts from A Conversation with Gordon Moore: Moore's Law.
56. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research 30: 207-210.
57. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, et al. (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic acids research 31: 94-96.
58. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, et al. (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic acids research 39: D1002-1004.
59. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. Nucleic acids research 37: D868-872.
60. Brazma A (2009) Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges. TheScientificWorldJournal 9: 420-423.
61. Krause A, Combaret V, Iacono I, Lacroix B, Compagnon C, et al. (2005) Genome-wide analysis of gene expression in neuroblastomas detected by mass screening. Cancer letters 225: 111-120.
62. Yannakakis M, Gavril F (1980) Edge dominating sets in graphs. SIAM Journal on Applied Mathematics 38: 364-372.
63. Kalfoglou Y, Schorlemmer M (2003) Ontology mapping: the state of the art. The Knowledge Engineering Review 18: 1-31.

64. Madhavan J, Bernstein PA, Domingos P, Halevy AY. Representing and reasoning about mappings between domain models; 2002.
65. Tian Y, McEachin RC, Santos C, States DJ, Patel JM (2007) SAGA: a subgraph matching tool for biological graphs. Bioinformatics 23: 232-239.
66. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, et al. (2005) Relations in biomedical ontologies. Genome biology 6: R46.
67. Simon J, Dos Santos M, Fielding J, Smith B (2006) Formal ontology for natural language processing and the integration of biomedical databases. International journal of medical informatics 75: 224-231.
68. Yeh I, Karp PD, Noy NF, Altman RB (2003) Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics 19: 241-248.
69. Noy NF, Musen MA (2000) Using PROMPT Ontology-Comparison Tools in the EON Ontology Alignment Contest. Stanford Medical Informatics, Stanford University.
70. Hill DP, Blake JA, Richardson JE, Ringwald M (2002) Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. Genome research 12: 1982-1991.
71. Degtyarenko K, Hastings J, de Matos P, Ennis M (2009) ChEBI: an open bioinformatics and cheminformatics resource. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] Chapter 14: Unit 14 19.
72. Johnson HL, Cohen KB, Baumgartner WA, Jr., Lu Z, Bada M, et al. (2006) Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing: 28-39.
73. Sarntivijai S, Ade AS, Athey BD, States DJ (2007) The Cell Line Ontology and its use in tagging cell line names in biomedical text. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium: 1103.
74. Sarntivijai S, Zuoshuang X, Meehan TF, Diehl AD, Vempati U, et al. Cell Line Ontology: Redesigning the Cell Line Knowledgebase to Aid Integrative Translational Informatics; 2011 July 28-30, 2011; Buffalo, N.Y.
75. Liu H, Hu ZZ, Torii M, Wu C, Friedman C (2006) Quantitative assessment of dictionary-based protein named entity tagging. Journal of the American Medical Informatics Association : JAMIA 13: 497-507.
76. Schulz S, Beisswanger E, Wermter J, Hahn U (2006) Towards an upper level ontology for molecular biology. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium: 694-698.
77. Rinaldi F, Schneider G, Kaljurand K, Hess M, Romacker M (2006) An environment for relation mining over richly annotated corpora: the case of GENIA. BMC bioinformatics 7 Suppl 3: S3.
78. Nelson-Rees WA, Flandermeyer RR, Hawthorne PK (1974) Banded marker chromosomes as indicators of intraspecies cellular contamination. Science 184: 1093-1096.

79. Nelson-Rees WA (2001) Responsibility for truth in research. Philosophical transactions of the Royal Society of London Series B, Biological sciences 356: 849-851.

80. Drexler HG, Quentmeier H, Dirks WG, Uphoff CC, MacLeod RA (2002) DNA profiling and cytogenetic analysis of cell line WSU-CLL reveal cross-contamination with cell line REH (pre B-ALL). Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 16: 1868-1870.

81. Drexler HG, Uphoff CC, Dirks WG, MacLeod RA (2002) Mix-ups and mycoplasma: the enemies within. Leukemia research 26: 329-333.

82. MacLeod RA, Dirks WG, Reid YA, Hay RJ, Drexler HG (1997) Identity of original and late passage Dami megakaryocytes with HEL erythroleukemia cells shown by combined cytogenetics and DNA fingerprinting. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 11: 2032-2038.

83. Kerrigan L, Nims RW (2011) Authentication of human cell-based products: the role of a new consensus standard. Regenerative medicine 6: 255-260.

84. Manniello A, Ruzzon T (1996) Cell Line Data Base and HyperCLDB. Biotech Knowledge Sources 9.

85. Parodi B, Aresu O, Manniello A, Romano P (1993) Human and Animal Cell Lines Catalogue Editrice abc – Officine Grafiche, Genova.

86. Romano P, Aresu O, Iannotta B, Manniello A, Parodi B, et al. (1993) Interlab Project Databases: an effort towards the needs of a wider body of unskilled users. . Binary 5: 7.

87. Noy NF, M. C, R.W. G, Knublauch H, Tu SW, et al. Protégé-2000: an open source ontology-development and knowledge-acquisition environment 2003. pp. 953.

88. Noy NF, Sintek M, Decker S, M. C, Gerguson RW, et al. (2000) Creating Semantic Web Contents with Protégé-2000. Intelligent Systems, IEEE 6: 60-71.

89. Arp R, Smith B. Function, Role, and Disposition in Basic Formal Ontology.; 2008. Nature Precedings.

90. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y (2010) OntoFox: web-based support for ontology reuse. BMC research notes 3: 175.

91. Liang XF, Li L, Liu DW, Li KL, Wu WD, et al. (2011) Safety of influenza A (H1N1) vaccine in postmarketing surveillance in China. The New England journal of medicine 364: 638-647.

92. Schurer SC, Vempati U, Smith R, Southern M, Lemmon V (2011) BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. Journal of biomolecular screening 16: 415-426.

93. Scherer WF, Syverton JT, Gey GO (1953) Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. The Journal of experimental medicine 97: 695-710.

94. Haber P, Sejvar J, Mikaeloff Y, DeStefano F (2009) Vaccines and Guillain-Barre syndrome. Drug Saf 32: 309-323.

95. Chen RT, Rastogi SC, Mullen JR, Hayes SW, Cochi SL, et al. (1994) The Vaccine Adverse Event Reporting System (VAERS). Vaccine 12: 542-550.

96. Botsis T, Ball R (2011) Network analysis of possible anaphylaxis cases reported to the US vaccine adverse event reporting system after H1N1 influenza vaccine. Studies in health technology and informatics 169: 564-568.
97. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R (2011) Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. Journal of the American Medical Informatics Association : JAMIA 18: 631-638.
98. Geier DA, Geier MR (2006) A meta-analysis epidemiological assessment of neurodevelopmental disorders following vaccines administered from 1994 through 2000 in the United States. Neuro endocrinology letters 27: 401-413.
99. Woo EJ, Ball R, Burwen DR, Braun MM (2008) Effects of stratification on data mining in the US Vaccine Adverse Event Reporting System (VAERS). Drug safety : an international journal of medical toxicology and drug experience 31: 667-674.
100. Geier MR, Geier DA (2004) A case-series of adverse events, positive re-challenge of symptoms, and events in identical twins following hepatitis B vaccination: analysis of the Vaccine Adverse Event Reporting System (VAERS) database and literature review. Clinical and experimental rheumatology 22: 749-755.
101. Varricchio F, Iskander J, Destefano F, Ball R, Pless R, et al. (2004) Understanding vaccine safety information from the Vaccine Adverse Event Reporting System. Pediatr Infect Dis J 23: 287-294.
102. Evans SJ, Waller PC, Davis S (2001) Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and drug safety 10: 483-486.
103. Heeley E, Waller P, Moseley J (2005) Testing and implementing signal impact analysis in a regulatory setting: results of a pilot study. Drug safety : an international journal of medical toxicology and drug experience 28: 901-906.
104. Egberts AC, Meyboom RH, van Puijenbroek EP (2002) Use of measures of disproportionality in pharmacovigilance: three Dutch examples. Drug safety : an international journal of medical toxicology and drug experience 25: 453-458.
105. DuMouchel W (1999) Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System. The American Statistician 53.
106. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, et al. (1998) A Bayesian neural network method for adverse drug reaction signal generation. European journal of clinical pharmacology 54: 315-321.
107. He Y, Xiang Z, Sarntivijai S, Toldo L, Ceusters W. AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events. In: Smith B, editor; 2011 28-30 July, 2011; Buffalo, NY. pp. 309.
108. Brown EG (2003) Methods and pitfalls in searching drug safety databases utilising the Medical Dictionary for Regulatory Activities (MedDRA). Drug safety : an international journal of medical toxicology and drug experience 26: 145-158.
109. Hill AB (1965) The Environment and Disease: Association or Causation? Proc R Soc Med 58: 295-300.

110. Saunders NR, Habgood MD, Dziegielewska KM (1999) Barrier mechanisms in the brain, I. Adult brain. Clinical and experimental pharmacology & physiology 26: 11-19.

111. Evans D, Cauchemez S, Hayden FG (2009) "Prepandemic" immunization for novel influenza viruses, "swine flu" vaccine, Guillain-Barre syndrome, and the detection of rare severe adverse events. J Infect Dis 200: 321-328.

112. Lee GM, Greene SK, Weintraub ES, Baggs J, Kulldorff M, et al. (2011) H1N1 and seasonal influenza vaccine safety in the vaccine safety datalink project. American journal of preventive medicine 41: 121-128.

113. Maquet D, Croisier JL, Dupont C, Moutschen M, Ansseau M, et al. (2010) Fibromyalgia and related conditions: electromyogram profile during isometric muscle contraction. Joint, bone, spine : revue du rhumatisme 77: 264-267.

114. Yikilmaz A, Doganay S, Gumus H, Per H, Kumandas S, et al. (2010) Magnetic resonance imaging of childhood Guillain-Barre syndrome. Child's nervous system : ChNS : official journal of the International Society for Pediatric Neurosurgery 26: 1103-1108.

115. Baxter R, Lewis N, Bakshi N, Vellozzi C, Klein NP (2012) Recurrent Guillain-Barre Syndrome Following Vaccination. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America.

116. Lee SJ, Kim YO, Woo YJ, Kim MK, Nam TS, et al. (2012) Neurologic adverse events following influenza A (H1N1) vaccinations in children. Pediatrics international : official journal of the Japan Pediatric Society.

117. Andrews N, Stowe J, Al-Shahi Salman R, Miller E (2011) Guillain-Barre syndrome and H1N1 (2009) pandemic influenza vaccination using an AS03 adjuvanted vaccine in the United Kingdom: self-controlled case series. Vaccine 29: 7878-7882.

118. Choe YJ, Cho H, Kim SN, Bae GR, Lee JK (2011) Serious adverse events following receipt of trivalent inactivated influenza vaccine in Korea, 2003-2010. Vaccine 29: 7727-7732.

119. Dieleman J, Romio S, Johansen K, Weibel D, Bonhoeffer J, et al. (2011) Guillain-Barre syndrome and adjuvanted pandemic influenza A (H1N1) 2009 vaccine: multinational case-control study in Europe. BMJ 343: d3908.

120. Sejvar JJ, Pfeifer D, Schonberger LB (2011) Guillain-barre syndrome following influenza vaccination: causal or coincidental? Current infectious disease reports 13: 387-398.

121. Verity C, Stellitano L, Winstone AM, Andrews N, Stowe J, et al. (2011) Guillain-Barre syndrome and H1N1 influenza vaccine in UK children. Lancet 378: 1545-1546.

122. Williams SE, Pahud BA, Vellozzi C, Donofrio PD, Dekker CL, et al. (2011) Causality assessment of serious neurologic adverse events following 2009 H1N1 vaccination. Vaccine 29: 8302-8308.

123. Burwen DR, Ball R, Bryan WW, Izurieta HS, La Voie L, et al. (2010) Evaluation of Guillain-Barre Syndrome among recipients of influenza vaccine in 2000 and 2001. American journal of preventive medicine 39: 296-304.

124. McNeil MM, Arana J, Stewart B, Hartshorn M, Hrncir D, et al. (2012) A cluster of nonspecific adverse events in a military reserve unit following pandemic influenza A (H1N1) 2009 vaccination-Possible stimulated reporting? Vaccine.

125. Vellozzi C, Broder KR, Haber P, Guh A, Nguyen M, et al. (2010) Adverse events following influenza A (H1N1) 2009 monovalent vaccines reported to the Vaccine Adverse Event Reporting System, United States, October 1, 2009-January 31, 2010. Vaccine 28: 7248-7255.

126. Evans D, Cauchemez S, Hayden FG (2009) "Prepandemic" immunization for novel influenza viruses, "swine flu" vaccine, Guillain-Barre syndrome, and the detection of rare severe adverse events. The Journal of infectious diseases 200: 321-328.

127. Vellozzi C, Burwen DR, Dobardzic A, Ball R, Walton K, et al. (2009) Safety of trivalent inactivated influenza vaccines in adults: background for pandemic influenza vaccine safety monitoring. Vaccine 27: 2114-2120.

128. Juurlink DN, Stukel TA, Kwong J, Kopp A, McGeer A, et al. (2006) Guillain-Barre syndrome after influenza vaccination in adults: a population-based study. Archives of internal medicine 166: 2217-2221.

129. Izurieta HS, Haber P, Wise RP, Iskander J, Pratt D, et al. (2005) Adverse events reported following live, cold-adapted, intranasal influenza vaccine. JAMA : the journal of the American Medical Association 294: 2720-2725.

130. Kao CD, Chen JT, Lin KP, Shan DE, Wu ZA, et al. (2004) Guillain-Barre syndrome coexisting with pericarditis or nephrotic syndrome after influenza vaccination. Clinical neurology and neurosurgery 106: 136-138.

131. Geier MR, Geier DA, Zahalsky AC (2003) Influenza vaccination and Guillain Barre syndrome small star, filled. Clinical immunology 107: 116-121.

132. Lasky T, Terracciano GJ, Magder L, Koski CL, Ballesteros M, et al. (1998) The Guillain-Barre syndrome and the 1992-1993 and 1993-1994 influenza vaccines. The New England journal of medicine 339: 1797-1802.

133. Niu M, Ball R (2009) Adverse events after anthrax vaccination reported to the Vaccine Adverse Event Reporting System (VAERS), 1990-2007 [Vaccine 2009;27:290-297]. Vaccine 27: 6654-6655.

134. Die-Kacou H, Yavo JC, Kakou KA, Kamagate M, Balayssac E, et al. (2009) [Post-vaccinal adverse effects monitoring during national campaign of vaccination against measles in Cote d'Ivoire]. Bull Soc Pathol Exot 102: 21-25.

135. Souayah N, Nasar A, Suri MF, Qureshi AI (2007) Guillain-Barre syndrome after vaccination in United States a report from the CDC/FDA Vaccine Adverse Event Reporting System. Vaccine 25: 5253-5255.

136. Haber P, Slade B, Iskander J (2007) Letter to the Editor. Guillain-Barre Syndrome(GBS) after vaccination reported to the United States Vaccine Adverse Event Reporting System(VAERS) in 2004. Vaccine 25: 8101.

137. Souayah N, Nasar A, Suri MF, Qureshi AI (2009) Guillain-Barre syndrome after vaccination in United States: data from the Centers for Disease Control and Prevention/Food and Drug Administration Vaccine Adverse Event Reporting System (1990-2005). Journal of clinical neuromuscular disease 11: 1-6.

138. Hasse P, Volker J (2008) Ontology Learning and Reasoning - Dealing with uncertainty and Inconsistency. In: CesarG. da Costa P, C. dA, Fanizzi N, Laskey

KB, Laskey KJ et al., editors. Uncertainty Reasoning for the Semantic Web I, ISWC International Workshops, URSW 2005-2007: Springer-Verlag Berlin Heidelberg. pp. 366-384.

139. Klein M. Combining and relating ontologies: an analysis of problems and solutions; 2001.

140. (NLM) NLoM (2009) Unified Medical Language System (UMLS). Bathesda (MD): National Library of Science.

141. Ade AW, ZC; States, DJ (2007) Gene2MeSH [Internet]. 2007 Mar. ed. Ann Arbor (MI).

142. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, et al. (2010) ConceptGen: a gene set enrichment and gene set relation mapping tool. Bioinformatics 26: 456-463.

143. Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, et al. (2007) Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. Nucleic acids research 35: D566-571.

144. Tarcea VG, Weymouth T, Ade A, Bookvich A, Gao J, et al. (2009) Michigan molecular interactions r2: from interacting proteins to pathways. Nucleic acids research 37: D642-646.

145. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

146. Toovey S (2008) Influenza-associated central nervous system dysfunction: a literature review. Travel Med Infect Dis 6: 114-124.

147. Piyasirisilp S, Hemachudha T (2002) Neurological adverse events associated with vaccination. Curr Opin Neurol 15: 333-338.

148. Jun Y, Ahn K (2011) Tmp21, a novel MHC-I interacting protein, preferentially binds to Beta2-microglobulin-free MHC-I heavy chains. BMB reports 44: 369-374.

149. Carroll IR, Wang J, Howcroft TK, Singer DS (1998) HIV Tat represses transcription of the beta 2-microglobulin promoter. Molecular immunology 35: 1171-1178.

150. Ruggiu M, McGovern VL, Lotti F, Saieva L, Li DK, et al. (2011) A role for SMN exon 7 splicing in the selective vulnerability of motor neurons in Spinal Muscular Atrophy. Molecular and cellular biology.

151. Simoni J, Simoni G, Moeller JF, Tsikouris JP, Wesson DE (2007) Evaluation of angiotensin converting enzyme (ACE)-like activity of acellular hemoglobin. Artificial cells, blood substitutes, and immobilization biotechnology 35: 191-210.

152. Ferreira AJ, Santos RA (2005) Cardiovascular actions of angiotensin-(1-7). Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica [et al] 38: 499-507.