***a)***

**Distribution Agreement**

Signature:

12/10/2014

Wayne A Harris Date

**b) Designing an integrated analytic database for lymphoma patient research using an open source toolset: Design and evaluation of elements for the creation of the Georgia Patient Analytic Lymphoma Registry (GA-PAL).**

By

Wayne A. C. Harris

Master of Science in Public Health

Department of Biostatistics, Rollins School of Public Health, Emory University

_________________________________________ [Chair’s signature]

Dr. Lance Waller, Ph. D.

Committee Chair

# c) Designing an integrated analytic database for lymphoma patient research using an open source toolset: Design and evaluation of elements for the creation of the Georgia Patient Analytic Lymphoma Registry (GA-PAL).

By

Wayne A. C. Harris

B. Sc

University of Toronto

1994

Thesis Committee Chair: Dr. Lance Waller, Ph. D.

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health

in the Department of Biostatistics

2014

# Abstract

Designing an integrated analytic database for lymphoma patient research using an open source toolset: Design and evaluation of elements for the creation of the Georgia Patient Analytic Lymphoma Registry (GA-PAL).

By Wayne A. C. Harris

## *d) Abstract*

Increasingly, informatics is having a significant impact on the management, analysis, and reporting of health data. As the field has matured, improved tools have evolved for these purposes and identified greater application in the healthcare research setting. Still, significant challenges remain particularly in the collection, integration, and analysis of health data. The complexity of unstructured data stored in huge data silos at healthcare institutions and lack of standardization contribute to the challenges. Another consideration is the steadily and exponentially growing stream of data that is becoming harder to manage and interpret. These challenges present a level of complexity that is difficult to overcome. In this project, we describe methods to use existing data integration tools to construct a lymphoma patient database and constructed an ontology to link ICD-9 coded electronic health record data with ICD-O-3 coded cancer registry data. The Georgia Patient Analytic Lymphoma Registry database (GA-PAL) is based on an open source analytic, semantically driven informatics platform, Eureka Clinical Analytics, under development here at Emory University. This platform leverages a suite of applications to provide the desired functionality. Protégé (http://protege.stanford.edu, Stanford University) is the ontology management component. Data extraction and transformation is achieved by PROTEMPA a temporal data abstraction technology. All of the data is finally imported into I2B2, a database platform, where data can be queried using ontology concepts as well as derived or user defined variables. We created a database of 12491 patients with defined diagnosis of lymphoma by ICD-9 codes from 1992-2012. A simple query of this data set for patients receiving RCHOP chemotherapy regimen produced a subset of 3082 patients. This conflicted with the data we also received from the hospital cancer registry that indicated there were about 4500 confirmed cases of lymphoma diagnosed during the same period.

Although challenges still remain to achieving full functionality, the use of this open source solution to a prevailing problem shows great promise. Our work here draws upon the previous work done to develop the LEAD database architecture based on the caBIG platform(Huang et al. 2009).

**COVER PAGE**

# e) Designing an integrated analytic database for lymphoma patient research using an open source toolset: Design and evaluation of elements for the creation of the Georgia Patient Analytic Lymphoma Registry (GA-PAL).

By

Wayne A. C. Harris

B. Sc

University of Toronto

1994

Thesis Committee Chair: Dr. Lance Waller, Ph. D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health

in the Department of Biostatistics

2014

# Acknowledgements

My sincerest thanks to Dr. Lance Waller and Dr. Christopher Flowers who helped me see things through to the end. I went into bonus time and you kept the clocks running. I also have to show sincere appreciation to Dr. Vicki Hertzberg with whom this journey into health informatics  began, and who was earnestly supportive through this experience.

My family has been my eternal source of support and I wish that I continue to grow and become my best for them. My mother Carol is one of my greatest critics and greatest supporters and I only wish to make you proud. I particularly wish my grandmother was here to see me finally complete this work. I am sure she would tell me “ Boy, why you took so long?” in her sweet Trinidadian accent. I am sad that you are not here to see me “mumsy” but I know that you would tell me .. ‘now keep going’.

To my wife Lan and baby Sophia. I am glad that you have put up with my grumpiness and neglect the last month especially and been enourmously patient with me. I owe you both and look forward to spending more time with you both. We are entering a period of great changes in our lives and I look forward to sharing that with you.

Finally I just wish to express my deep appreciation to everyone that has supported me along this path and I hope that I can keep on it and do better…

**Table of Contents**

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

Wayne Harris 12/10/14 5:35 PM
**Formatted**

## 1. Introduction

### *A. Background Information*

Hematologic malignancies, which include the acute and chronic leukemias, Hodgkin's disease (now called Hodgkin's lymphoma), non-Hodgkin’s lymphoma (NHL) and myeloma, account for 9% of cancer cases diagnosed in the US. They occur less commonly than some solid tumors but represent a significant disease burden in the population. These malignancies represent a large and heterogeneous collection of cancers reflecting the complexity of the normal hematopoietic and immune systems of which they are derived.

*Non-Hodgkin Lymphoma (NHL)*

According to Surveillance, Epidemiology, and End Result (SEER) data (http://seer.cancer.gov/statfacts), in 2014 there will be an estimated 70,800 new cases of NHL (or 4.3% of all new cancer cases), and 18,990 deaths due to NHL (3.2% of all cancer deaths). The 5-year survival expected in 2014 is 69.3% (2004-2010 data). About 2.1% (lifetime risk) of men and women in the US will be diagnosed with NHL in their lifetime (SEER 2008-2010 data). (Surveillance Research Program 2014) About one quarter (27.8%) of newly diagnosed cases are diagnosed with local stage disease and the age adjusted incidence for all cases is 23.8/100,000 in men and 16.3/100,000 in women; 75.2% of patients are diagnosed after the age of 54 and 9.3% of diagnosed patients are over age 84 (the median age at diagnosis is 66). However, unlike leukemia, NHL is not common in pediatric patients under age 20 (1.6%). Although death rates due to NHL have been falling 2.7% per year since 2001, it remains the 8th most common cause of death in the US. Approximately 85% of all NHLs are of B cell origin and the remaining 15% are T cell origin however this dichotomy belies the complex heterogeneity of the disease, e.g., general survival

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

statistics for NHL are not really helpful because survival depends heavily on the type of lymphoma. The numbers for Hodgkin's disease and myeloma reflect similar patterns as seen for leukemia and NHL but to a lower degree as disease rates for these are lower (see Table 1).

Table 1. Summary of Incidence and Mortality Due to Hematologic Neoplasms Expected in 2014 (data compiled from Cancer Stat Fact Sheets http://seer.cancer.gov/statfacts/).

| **2014 Statistics** | **Estimated New Cases** | **% Of All New Cancer Cases** | **Estimated Cancer Deaths** | **% Of All Cancer Deaths** |
|---|---:|---:|---:|---:|
| Leukemia | 52,380 | 3.1 | 24,090 | 4.1 |
| Lymphoma (Hodgkin's) | 9,190 | 0.6 | 1,180 | 0.2 |
| Lymphoma (non-Hodgkin) | 70,800 | 4.3 | 18,990 | 3.2 |
| Myeloma | 24,050 | 1.4 | 11,090 | 1.9 |
| **Total** | **156,420** | **9.4** | **55,350** | **9.4** |

**Analyzing Disease Patterns**

A number of classification systems have been developed to categorize lymphoid malignancies in clinically and biologically relevant ways to refine our ability to diagnose and treat these cancers.  For example, the Revised European-American Lymphoma (REAL) classification published in 1994 defines clinic-pathological entities based on a combination of morphology, immunophenotype, genetic abnormalities and clinical features (Morton et al. 2007).  More than 90% of lymphoid malignancies can be classified using this approach.  The World Health Organization's (WHO) International Classification of Diseases (ICD) represents the standard diagnostic system for the reporting, storage, and retrieval of data for epidemiology, health management, and clinical purposes. (Morton et al. 2007)

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

The WHO's ICD system is used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and health records. Reported conditions are designated by codes in the classification schema which undergo periodic revision to update new understanding of disease biology, diagnosis and treatment. A special ICD coding schema also exists specifically for oncology, ICD-O, which is based on the same approach as the REAL classification thus allowing healthcare systems to capture granular information about the biological and clinical characteristics of a neoplasm. The standard ICD coding cannot capture histological information about a cancer whereas the ICO-O schema can easily distinguish histologically distinct neoplasms by means of integrated coding for morphology (tumor cell type behavior and differentiation) and topography (site and sub site). The ICD coding schema is in its 10th revision (ICD-10) despite most clinical systems still relying on the ICD-9 codes to capture clinical data. (Morton et al. 2007) The ICD-O codes are in their 3rd revision (ICD-O-3), categorize lymphomas according to the cell of origin (B cells, T cells, or NK cells) and incorporate disease morphology, immunophenotype, and genetic and clinical features to define subtype. Because of its greater specificity and granularity, the ICD-O-3 coding system is used to capture data in cancer registries. In addition to improvements in disease coding and capture, improved treatment modalities along with improved diagnostic techniques have improved patient outcomes for hematologic neoplasms (this is reflected in the chart on the SEER Stat facts and figures - http://seer.cancer.gov/statfacts/html/leuks.html). However there are still areas where additional research is needed to address challenges with respect to diagnosis and treatment of lymphoid malignancies, such as improving screening or finding better treatments since one in ten cancer patients die of one of these cancers and they are so heterogeneous in nature.

Data about cancer incidence, diagnosis and treatment in the US is facilitated by a system of disease registries that are funded at the national level and managed at the state level. Through ongoing cancer surveillance, the timely and systematic collection and analysis of data on new cancer cases, extent of disease, screening, treatment, survival and mortality, a steady stream of data is collected via health

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

providers to the state run registries. The program is formally called the Surveillance, Epidemiology and End Results program (SEER), and the data they collect is used to examine trends over time, find disease patterns by region or groups of people, and/or show whether screening and other prevention measures are having an impact. (American Cancer Society, http://www.cancer.gov)

There are two general streams of data defining registries, hospital-based data and population-based data. Hospital registries may be part of a treatment facility's cancer program and provide complex data used to evaluate patient care within the hospital. These data are effective for monitoring care and educating care providers, and may also be pooled to provide information about the course of cancer and comparative effectiveness of treatments among providers, patient subsets and regions. The population-based registries, such as those based at state health departments, aggregate data from multiple reporting facilities within their geographic region like hospitals, community clinics and doctors' offices, nursing homes and cancer treatment and care facilities to collect information on all cases diagnosed in their region. The data collected allow the determination of disease incidence and mortality by region across the country, by demographic characteristics of the individuals, and by diagnosis time so cancer trends and the impact on certain communities or cohorts can be observed and inform cancer prevention and control programs. (American Cancer Society, http://www.cancer.gov)

Subsequently, the SEER databases have detailed sociodemographic information in addition to details about disease histology, treatments and survival.  However, they contain only limited clinical data and specific clinical events are not as well documented in the cancer registry. The advantage to the registry databases is that they contain data that have been manually validated and verified for the clinical diagnosis rather than depending on billing codes and therefore are more reliable. Conversely, hospital registries such as the Emory Cancer Registry (ECR), the cancer registry for Emory Healthcare hospitals, is a clinical database comprised of clinical data, laboratory data and treatment response data which are derived from separate system. (i.e. HealthQuest: hospital data, IDX: clinic data, EeMR/Cerner Powerchart: patient electronic medical records, Pharmnet:pharmacy data) (Huang et al. 2009)

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

Because the ICD-O coding schema is so comprehensive in its ability to capture data about cancers and to communicate information about cancers it has become the de facto categorization methodology for cancer registries and has facilitated better monitoring and surveillance of cancer in populations. The Emory Cancer Registry, like many provider registries must abstract data from systems that codify disease using the ICD-9 billing codes, requires data providers to convert the ICD 9 coding system to ICD-O. This usually occurs during the manual review of the patient record in the abstraction process explained in more detail below.

State cancer registries like the Georgia Comprehensive Cancer Registry (GCCR), located in Atlanta at the Rollins School of Public Health at Emory University, collect data on all cancer cases in a state for the purpose of epidemiologic monitoring of disease. Such registries capture granular data on the diagnosis of disease, patient demographics and survival data on all patients being treated for cancer. In the case of the GCCR, most of their data comes from provider institutions like Emory Healthcare that are major treatment centers and a focal point for patients seeking care in their state. The state cancer registries, however, do not collect data directly. Instead, the provider institutions typically have a cancer registry team that evaluates patient data to detect and locate potential cancer patients that enter their system. In the case of Emory healthcare, treatment encounters generate ICD9 billing codes and HL7 transactions codes for encounter (Health Level-7 or HL7 refers to a set of international standards for transfer of clinical and administrative data between hospital information systems (http://www.HL7.org)). Codes that potentially correspond with diagnosis, history or treatment of cancer are flagged for manual review. A registrar then manually reviews the patient record and abstracts diagnosis and administrative data if indeed that patient is found to have been diagnosed or treated for cancer.  The cancer patients' data are then reported via specific protocols and formats to the state cancer registry. Since the state cancer registries are mandated to collect data for the purposes of evaluating disease incidence and epidemiologic surveillance, treatment data and other clinical details are not reported to the registry. This reduces the utility of the disease registry for purposes of comparative effectiveness studies and quality and outcomes measurement. Still,

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

all of the clinical data are captured, even if not reported, in the healthcare provider's clinical transaction data records. In the case of Emory Healthcare, these data are collected in the transactional data mart called the clinical data warehouse or CDW.

### Clinical Data Warehouse

Within the Emory University Healthcare system (EHC), all patient data from the clinics and hospitals are captured into one central repository for patient data, the clinical data warehouse (CDW). A clinical data warehouse is a repository of historical health care data organized for reporting and analysis. A CDW facilitates data access by having data from multiple sources in one place with linked, easily searchable data. However, the data are largely unstructured and organized primarily for administrative reporting or for patient monitoring and management purposes, being applied mostly to patient management and billing (Lyman et al. 2008). One of the key new approaches to improving research and outcomes will be to more fully leverage the incredible volumes and varieties of data being collected into clinical data warehouses. These data have enormous potential to improve performance, measurement and health care quality, as well as generate new hypothesis and insight that may lead to more effective diagnostic capabilities, treatments and health outcomes for patients. However, accessing, organizing and analyzing these data to support research and quality improvement projects are persistent and evolving challenges (Lyman et al. 2008).

The data in the CDW are stored in discrete data elements and are structured to provide data in the following contexts: patient, provider, encounter, and location. The database also includes the following from the following systems:

- Data from legacy administrative systems (for both hospital and clinic)
- Electronic medical records
- Laboratory results
- Pharmacy

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

- Clinical trials
- Genomics and microarray databases

Data can be divided into the following subject areas:

- Clinic Billing and Registration
- Clinic Appointments and Scheduling
- Hospital Billing and Registration
- Medical Record Abstract
- Diagnosis/Procedures
- Emergency Department
- Clinical Laboratory Results
- Cytogenetics
- Orders
- Radiology Reports
- Medication Administration
- Prescriptions
- Power forms
- Clinical Documentation

Since the CDW contains protected health information (PHI), and to protect the real time performance of the CDW, which is part of the operational system of EHC, access to the data in the CDW by researchers, can currently only be obtained through a system analyst with the appropriate credentials.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

### *B. Problem Statement*

Increasingly, electronic systems are becoming standard in collecting, managing and eventually mining data to increase and improve our knowledge in health and health care. There is a pressing need to be able to share information and develop tools whereby data can be managed structured and analyzed automatically as new systems and policies in place and continuing to come online threaten to overwhelm the entire system with the sheer volume of data that will now be available to a researcher. It is at once both boon and curse because without adequate tools to mine the data, the value of the raw unstructured data in these huge data containers never gets realized. Clearly, the technology to structure, manage and evaluate data needs to be improved so that researchers won't have to depend on less secure, less sophisticated methods like desktop hard drives and portables storage media and simple Excel files to store and manage their research data. One challenge, as indicated above, is that the CDW data are simply too cumbersome and sensitive to allow researchers direct access. The current procedure for researchers to access data in the CDW data in our healthcare system involves providing an analyst with descriptive and phenotypic information about the patient subsets they want to select from the data mart.  The analyst then structures the database queries and data maps to extract the required data set. The process is inefficient because it requires dedicated IT resources to provide services to the entire research community and the analysts are not domain specialists.  Beyond that, clinical data are not the only important data that are useful in understanding disease. Both clinical and epidemiologic data contribute significantly to understanding disease. Epidemiologic data provides a context in which disease occurs and resolves, and the clinical data provides a details about disease progression and the effects of care and treatment. So in a sense, the public health data provides information about the health ecology and the clinical data provides information about the health process. Both are therefore required to have a more complete understanding and finding methods to reduce the impact of disease and illness in populations.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

### C. *Purpose Statement*

There have been many efforts to organize health data and create tools to manage it such that research and information sharing were facilitated. The most notable of these was the caBIG initiative sponsored by the National Cancer Institute (NCI), which unfortunately was resolved to be unsuccessful (NIH Board of Scientific Advisors (BSA) Ad Hoc Working Group report published March 2011). With the failure of the caBIG initiative, discussed further below (Foley 2011), to yield benefits and advantages to the research community in creating an informatics platform to manage cancer research and data nationally, there remains a pressing need to develop tools to extract data from unstructured records and integrate them with other data sources to create new data registries of structured and interpretable data that will enable better research and discovery. By leveraging previous work using existing resources that were developed for the caBIG platform to develop a data integration tool (Huang et al. 2009), we aim to design a new system that will facilitate the integration of existing databases using a highly extensible and modifiable open source platform that will structure data, allow the transformation of data to create user defined variables that will lend better interpretation of the data.

The high level objective of this project involves the creation of and demonstration of the use of an ontology based system to integrate patient level data with administrative and demographic data from multiple sources to create an electronic health record (EHR)-linked data set that could then be queried in novel ways. The fully operational system is intended to integrate data from the transaction-based CDW that contained all patient data from the hospitals and clinics in the Emory Healthcare System (EHC), administrative data from the Georgia Discharge Data System (GDDS) and Medicare and epidemiologic data from the GA-SEER cancer registry. We also hoped to include patient data from manually derived sources via REDCap forms (Research Electronic Data Capture) (Harris et al. 2009) and other cancer care providers in the state of Georgia, including Kaiser Permanente Group (KPG) and Georgia Cancer

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

Specialists (GCS), that would contribute to creation of a state wide comprehensive profile of lymphoma patients in the state of Georgia. These providers along with Emory healthcare primarily serve Metro Atlanta, North and Central Georgia where most lymphoma cases are reported, as shown in the figure below.

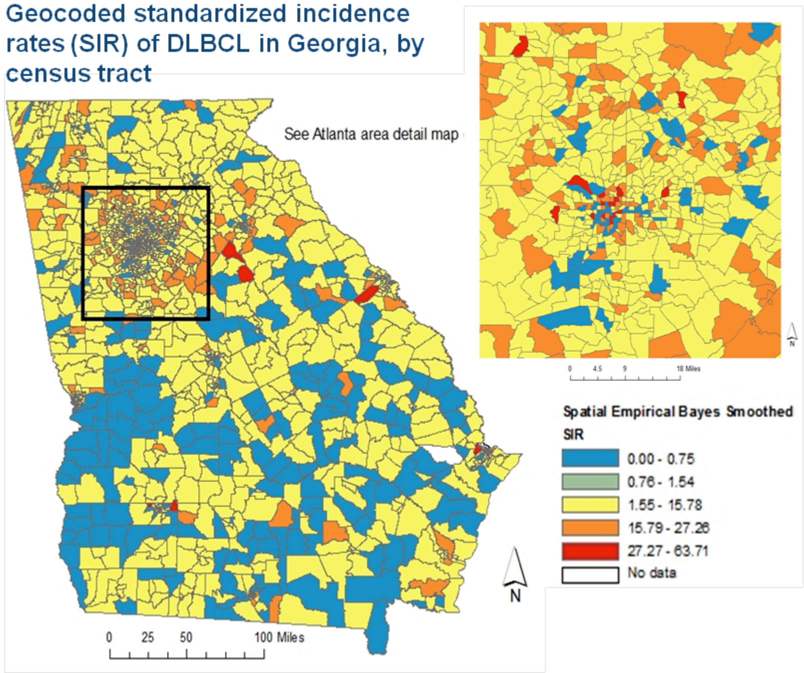

Geocoded standardized incidence rates (SIR) of DLBCL in Georgia, by census tract
See Atlanta area detail map
0 4.5 9 18 Miles
Spatial Empirical Bayes Smoothed SIR
0.00 - 0.75
0.76 - 1.54
1.55 - 15.78
15.79 - 27.26
27.27 - 63.71
No data
N
N
0 25 50 100 Miles

Figure 1. Geocoded standardized incidence rates for lymphoma in the state of Georgia demonstrating the concentration of cases in the state of Georgia the majority of cases are concentrated to the north of the state and would comprise the treatment population for the large care providers like Emory Healthcare, Kaiser Permanente Group and Georgia Cancer Specialists whose services are focused in the Metro Atlanta, North and Central Georgia regions.

### *D. Significance Statement*

Improving clinical outcomes and the quality of care patients receive are two significant goals of the healthcare system on which technology is now having a huge impact as better tools are increasingly

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

available. Health informatics is the scientific discipline that applies technological tools to the health domain to systematically acquire process and use data to develop and share new information and knowledge in medicine and health. (Kuziemsky and Lau 2010). One of the most important requirements for health information systems will be to integrate the huge volumes of data already being collected that may contain particular details that can benefit our insight and understanding of cancer. Lymphomas are a heterogeneous group of cancers that require the development of focused research and clinical approaches for histological subtypes. Integration of existing multiple data sources at the patient level will contribute to understanding the biological variability in the pathogenesis of lymphomas and their responses to treatment and will promote the development of innovative treatment strategies. In order to expedite the development of innovative clinical and therapeutic strategies for lymphoma, the oncology informatics team of the Winship Cancer Institute has been developing means to integrate existing clinical information into database systems that support cancer research. The long-term goal of this project is to broaden the scope of data integration to include data sources beyond the clinical setting, new types of data that will enrich the data for analysis and improve our understanding of cancer populations. One of the major challenges to data integration in the biomedical data management community is the vast numbers of ways similar or identical concepts are described in different information systems. It is critical to have methods to minimize semantic conflicts so that data can be better interpreted by machines.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

E. Definition of Terms

| | |
|---|---|
| AIW | Analytical Information Warehouse |
| BSA | Board of Scientific Advisors |
| caBIG | Cancer BioInformatics Grid |
| CDW | Clinical Data Warehouse |
| CT | Computed Tomography |
| CVRG | Cardiovascular Research Grid |
| DLBCL | Diffuse Large B Cell Lymphoma |
| ECR | Emory Cancer Registry |
| EHC | Emory HealthCare |
| EHR | Electronic Health Record |
| EMR | Electronic Medical Record |
| ETL | Extract Transform Load |
| FL | Follicular Lymphoma |
| GA-PAL | Georgia Patient Analytic Lymphoma Registry |
| GCCR | Georgia Comprehensive Cancer Registry |
| GCS | Georgia Cancer Specialists |
| GDDS | Georgia Discharge Data System |
| HL7 | Health Level 7 |
| ICD | International Classification of Diseases |
| KPG | Kaiser Permanente Group |
| LEAD | Lymphoma Enterprise Architecture Database |
| LLDB | LLDB Large linked Databases |
| MCL | Mantle Cell Lymphoma |
| NHL | non-Hodgkin's lymphoma |
| NIH | National Institutes of Health |
| PET | Positron Emitted Tomography |
| PHI | Public Health Information |
| REAL | Revised European-American Lymphoma |
| REDCap | Research Electronic Data Capture |
| SEER | Surveillance, Epidemiology, and End Result |
| WHO | World Health Organization |

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

**2. Review of previous data integration efforts**

A. Large Linked Databases (LLDB)

Previously, large linked databases (LLDB) were thought to be the solution to data integration. Earlier efforts to create a large linked database focused on cancer successfully combined the independent legacy databases within the Emory healthcare system into one system (Graiser et al. 2007). However, the query strategies developed to identify a cohort of follicular lymphoma (FL) patients, a challenging histological diagnosis, in the database using common ICD-O and ICD-9 codes and text searches of the electronic medical record (EMR) demonstrated the weakness of using ICD-9 codes over ICD-O codes. The ICD-9 coded query strategy was far less specific and sensitive in detecting this patient population. This effort also revealed a set of significant limitations in the use of LLDBs for medical research, namely:

1. Relying on coded outcomes using ICD9 diagnosis codes can lead to significant inaccuracies partly because these codes are frequently assigned by personnel unfamiliar with the patient, disease or procedure being coded.
2. The use of patient identifiers such as social security numbers to link data across heterogeneous databases can lead to data integrity problems caused by data entry errors, incomplete data entry, or inconsistent practices
3. Some LLDBs capture identical data points from multiple sources, thus compounding the inaccuracies unique to each data source

Consequently, the variability among the systems in the way the same disease may have been coded in a single institution can be magnified when comparing data between different institutions. Therefore, this

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

approach would not be effective for integrating or comparing data across institutions and the populations they serve.

### B. Cancer BioInformatics Grid (caBIG)

CaBIG, the Cancer Biomedical Informatics Grid, was a program launched by the NCI in 2004 to address the rapidly growing demands for tools to meet the needs of researchers that wanted to make use of the large volumes of data becoming available as technology was beginning to have a large impact on health data (Fromer 2012). The goal of the project was to create a technology platform that enabled information sharing by providing free and open access tools data and infrastructure to facilitate connectivity, collaboration and interoperability among research communities and information systems (Fromer 2012). The development of the caBIG platform created a new suite of informatics tools that used a semantic methodology to categorize data and promote sharing of data (Huang et al. 2009). The tool covered clinical trials management systems, tissue banks, pathology tools, integrated cancer research architecture, semantic vocabularies, common data elements (CDEs), data sharing and intellectual capital. This development also created a common, extensible informatics platform to integrate diverse data types and support interoperable analytics tools. Some of the key successes were pointed out in the 2011 NIH report published by the BSA Ad Hoc Working Group and included creation of a common set of data standards for data exchange and integration, the creation of a controlled vocabulary set (a standard set of terms and their established definition), and establishing a consensus that there was a need to overcome traditional boundaries and enable multi-organizational data, information and knowledge sharing. Controlled vocabularies were important for data integration because they provide a harmonized set of terms and definitions. The elements of a controlled vocabulary are called concepts, which can be defined explicitly or implicitly (Kohler et al. 2003). When relation data are added to these concepts we create an ontology whereby the knowledge in a given domain, the concepts and relationships between them, can be represented in a graph. The figure below shows the logic model for the concepts (nodes) and relations

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

(edges) in the clinical cancer care setting that was used to create a new data registry based on the caBIG platform. This is described below. Where caBIG was criticized in the report was the fact that it became too much of a software enterprise leading to the development of over 70 applications, but did not focus enough on creating a sufficient user base and being amenable to the needs of those users. Because of its technology-centric approach to data sharing it became too complex and unfocused (Fromer 2012).

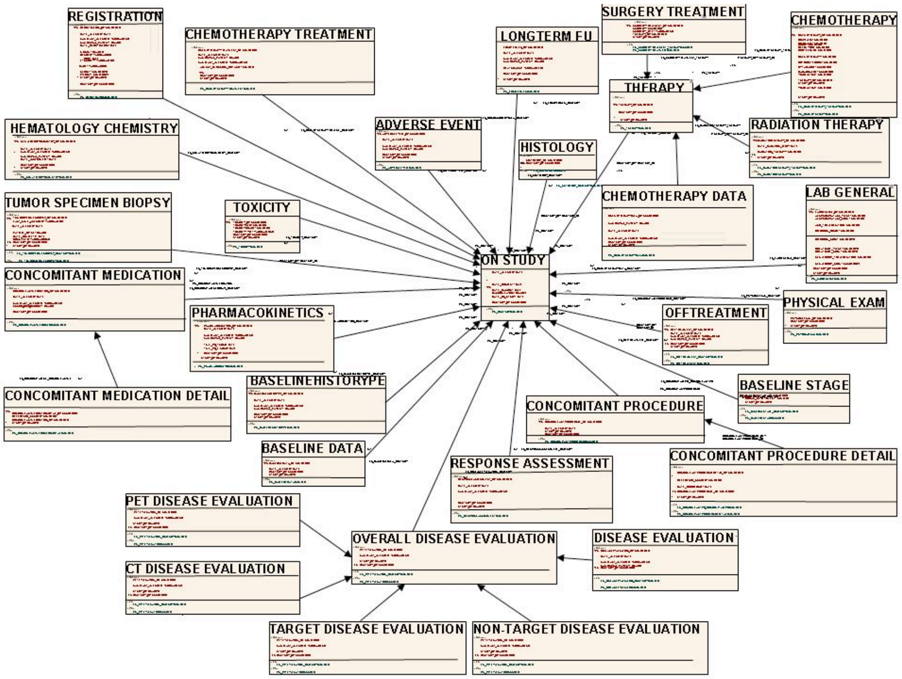


Figure 2. Logic model for caBIG-compliant LEAD clinical database. This figure shows the relationships between the key data elements in the cancer care setting (Huang et al. 2009).

## C. The Lymphoma Enterprise Architecture Database (LEAD)

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

In response to these challenges, our group built a new system called the Lymphoma Enterprise Architecture Database (LEAD) designed to integrate clinical and biomedical data at the patient level. This allowed patient focused integration of our institution's clinical trials, cancer registry, clinical, administrative and pharmacy data within a single database. We also showed in previous work (Huang et al. 2009) that the data elements and structures in LEAD database could be used for institutional studies linking data across legacy systems. Also, the patient data (once stored within the LEAD database) can be shared and reused based on the standards of the caBIG architecture. In this figure below, required variables for the LEAD database architecture are shown in a simple model to demonstrate how the data from different data sources were integrated into a new data model. The elements are organized under higher-level concepts or classes and color-coded to indicate the source.

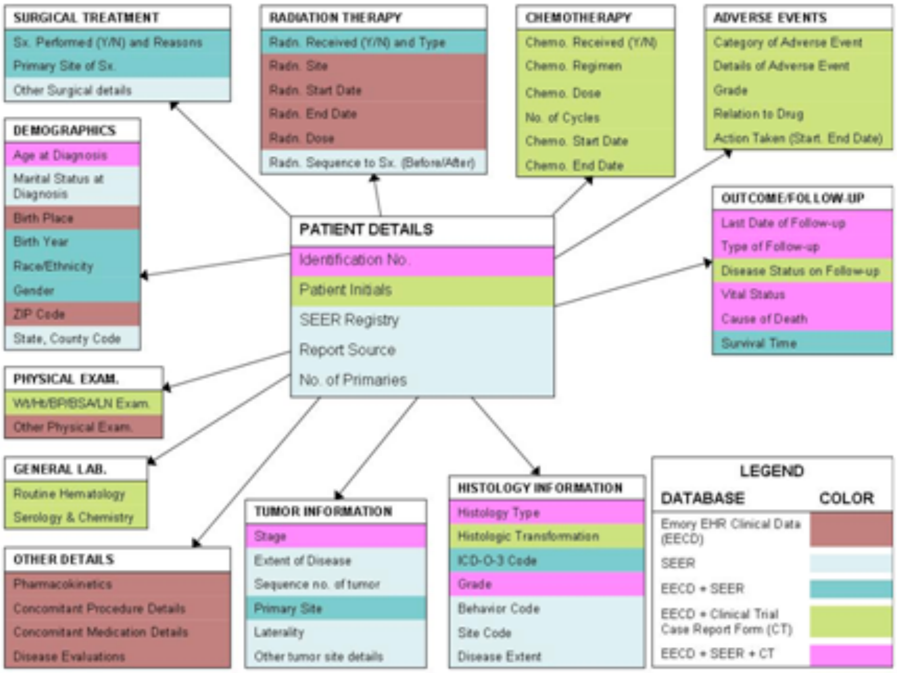


Figure 3. Relations between entities (concepts) and classes (categories) from all data sources in the LEAD database architecture are shown. The data categories are color coded to indicate data source. Color codes are shown in the legend. Abbreviations: EECO, Emory Electronic Health Record (HER) Clinical Data. SEER, Surveillance, Epidemiology and End Results Data. CT, Emory clinical trials data (Huang et al. 2009).

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

D. Summary of the Current Problem

With the failure of the caBIG platform, our group sought creation of a new system that borrowed on the successes of the LEAD database to create an ontology based system that would integrate patient level clinical and demographic data from clinical systems and cancer patient registries for epidemiologic research. Our objective was to produce a comprehensive database of lymphoma patients in the Emory Healthcare system containing diagnostic, administrative, procedural, demographic and outcome data. The ultimate goal is to create a model that could be expanded to integrate data from other healthcare providers and state registries and create a state-wide lymphoma database permitting the epidemiologic investigation of incidence, treatment disparities, patterns of care, and outcomes across the state of Georgia for patients treated with hematologic malignancies (leukemia, lymphoma and myeloma).

## 3. Methodology

### A. Introduction - Georgia Patient Analytic Lymphoma Registry (GA-PAL)

In the interest of improving diagnoses/treatment modalities and health outcomes for patients with lymphoid malignancies, mining of data in the CDW for hypothesis generation and discovery is an important process. Our research group has been working on ways to integrate existing clinical information into database systems that support cancer research. As stated earlier, the data structures in the CDW are not conducive to granular or nuanced biomedical research so the effort has been underway to create an adjunct database system to facilitate research that would hopefully lead to innovative clinical and therapeutic strategies for lymphoma. We are currently working on a system to integrate data from the Emory CDW (used for process and quality initiatives) with discharge data from the Georgia Discharge Data system (GDDS), Medicare data and Georgia cancer registry data from the Georgia Comprehensive

Cancer Registry (GCCR) to permit a comprehensive analysis of related patient data to develop an integrated perspective of lymphoma disease and treatment in Georgia.  Figure 4 below shows the schema for the overall system with data sources indicated. The accessible data that are established are in regular uncolored boxes, but the data where there may be some access challenges due to governance rules are shown in color. The elements circled in yellow represent data to which we will have limited access. The data sources circled in red are unlikely to provide raw data access as they would be proprietary to the provider. We envision this system capturing data from in-network hospitals and clinics belonging to Emory Healthcare, epidemiologic data from the Georgia Comprehensive Cancer Registry, discharge data and administrative data from the state health administrative systems (available through GDDS and Medicare linked SEER data via the state cancer registry-circled yellow) system, manually captured clinical data using the REDCap database system and potentially other clinical care providers in the state (i.e., Kaiser Permanente Group (KPG) and Georgia Cancer Specialists (GCS)-circled red). These data will not be presented in real time but refreshed in the database and are expected to be refreshed on an annual or semi-annual basis.  This system will leverage previous work to develop a new approach to creating research databases from large unstructured data stores.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right:  0.25"

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

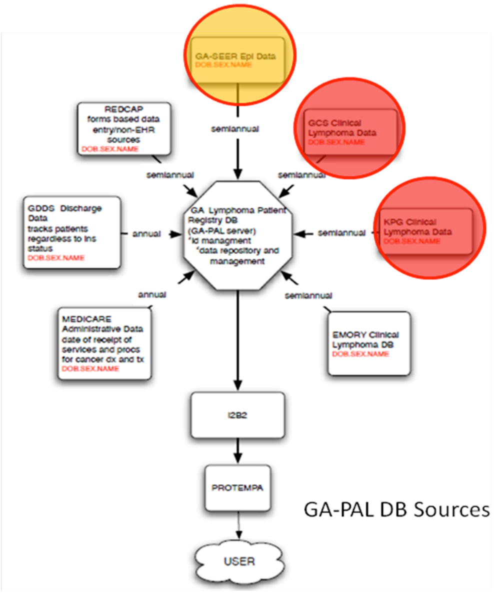


Figure 4. Schematic of the planned designed of the GA-PAL database. The providers are circled in red because we may not have access to proprietary raw data from these partners. The GA-SEER/GCCR element is circled in yellow because our access will likely be limited to data seen at our own institution due to privacy and consent concerns.

### B. Requirements

Our primary objectives were to demonstrate local integration of EHC clinical patient data with epidemiologic data from the GCCR and show how we could query this database to answer questions not answerable by the original data and existing systems.

Our initial strategy was to show that we can extract cohort specific data from the CDW into a new database and merge this with registry data from the GCCR. This newly created limited data set (LDS) would be defined by the patients' record for having a history of cancer and diagnosis of lymphoma indicated by ICD codes. The integration of data would be facilitated by the existing system ontology in

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

the CDW and new domain or application ontologies to map data between the ICD-O-3 coded registry data and the ICD9 coded data from the clinical systems. What we report on in this paper is extraction of a lymphoma defined patient set into a new database and the refinement of this database into a subset of patients under a specific cancer therapy to demonstrate the facility of our system in accessing and querying research data. Our long-term goals were to test the new database to address two specific categories of use cases that are relevant to the types of studies we plan to perform:

I. Clinical Use Cases

- What is the frequency of CT scans and PET scans in patients with follicular lymphoma (FL), mantle cell lymphoma (MCL) and diffuse large B cell lymphoma (DLBCL)?
  - At diagnosis
  - During treatment phase
  - At the end of treatment
  - During surveillance
- What are the patient diagnoses contained in the database?
- What treatment or chemotherapy are the patients receiving and what is the timing of those treatments?
- Did the patients get CT or PET scanned? If so what type and what timing?

II. Epidemiologic Use Cases

- Where does the patient live?
- What is the patient’s insurance status? Was there ever a lapse?
- What is the patient’s family history of cancer?
- What is the primary site of the patient’s cancer?
- What treatment regimen did the patient receive?

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

- Were there any adverse events?
- Did the patient have B-symptoms?
- What was the outcome of treatment?
- Was there relapse?

If we are successful in addressing needs in both of these domains we will have made a significant step towards creating a comprehensive tool for clinical research, epidemiologic research, and discovery in oncology informatics. Answering these questions requires creation of new algorithms to define new elements like chemotherapy regimen, treatment phase and periodicity because these are not directly coded in the native data structures. With the new system we are developing, we are able to implement semantic data structures to examine patient data using an open source platform that permits creation of algorithms within the system to define newly derived variables that facilitate knowledge driven data mining that is more accessible to non-traditional users of information systems. Thus researchers can directly access the system and data rather than having to depend on system analysts. To illustrate the potential of the system, we show that we can capture cohorts of patients based on disease phenotype or treatment characteristics and use that data to ask initially simple questions about the target patient subset.

### *C. Proposed Solution*

In developing solutions, we partnered with the software engineers at Emory University's Center for Comprehensive Informatics to provide specifications and functionality goals for the tool. A key element of our solution is the Eureka Clinical Analytics platform, an open source project already under development. It was conceived as part of the vision of the Cardiovascular Research Grid (CVRG) to create tools that enable researchers to analyze and manipulate their biomedical research data in the cloud (Post et al. 2013b). This application is essentially the graphical user interface for the Analytic

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

Information Warehouse (AIW), a previously developed software package that produces data registries from the CDW. The AIW allows the application of a semantic layer over the existing clinical data to map the structure of that data to a conceptual, ontology-based data model that is database-agnostic because it only requires that there is a schema mapping the ontology elements to the data source . Periodically all of the transactional data from the multiple databases of the hospital and clinics of the healthcare system that are stored in the CDW are cloned into the AIW which transforms the data by applying algorithms and data models that can generate derived variables that can be used to classify and categorize data. The process to build the AIW data mart is called Extract Transform and Load or ETL, and where this occurs is called the ETL layer. There are three parts and three components to the ETL layer of the AIW. First is the ontological abstraction. Second is the transformation of the data by PROTEMPA, a previously designed tool to categorize data. Finally, third is the load of the data, raw, transformed and derived into a new database in I2B2, a database system designed to be used for healthcare data. In which terms or variables can be queried according to concepts contained within an application ontology.

## EXTRACT

In the first step in the ETL process, a virtual data model (VDM) or abstraction ontology created by a data modeler using the Protégé (Stanford University) ontology editing software to model the concepts that data are to be mapped to in the domain with the data elements from the CDW. A separate XML document directly maps the ontology concept to the specific data element location in the CDW. Subsequently, the largely unstructured data in the CDW, which may have a schema that has nothing to do with how researchers conceive the data, can now have a conceptual structure overlain upon it.(Post et al. 2011; Post et al. 2013a)

## TRANSFORM

While the abstraction ontology developed in Protégé, together with the mapping data in the XML file is used to identify data elements to be extracted from the CDW. Another previously developed software

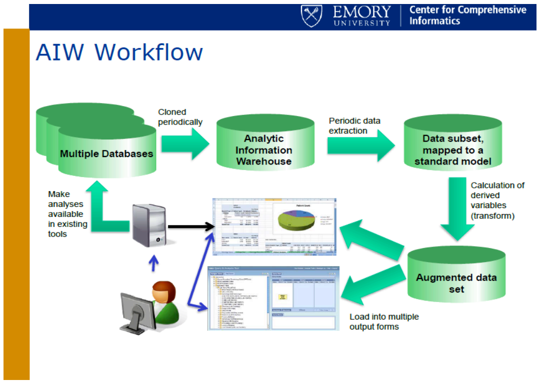
Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

application, PROTEMPA (Post and Harrison 2007), provides the backbone by which extraction and transformation of the data occur. This program allows defining trends and states by defining temporal patterns in sequential data that help specify a state or trend in the data, or assigning temporality. Any algorithms to create new or derived variables in the data are applied here as well (Post et al. 2011).

LOAD

Finally, the augmented data that now have some structure applied to them via the mapping ontology and temporal transformations, are loaded into a new database in I2B2 (Post et al. 2011). I2B2 is a widely adopted clinical research data warehousing system that allows investigation of large data sets by leveraging concepts. Rather than writing SQL queries to the DB, a user can create cohorts or subsets of patients by applying concepts and report that data out as a dataset or a file (Post et al. 2013a). The SQL code that generates all the SQL queries to extract the data and produce these results is generated automatically in the AIW. There is the preexisting functionality to create rules that define derived variables in the data that are executed via PROTEMPA and loaded into I2B2 as user defined variables. So while most of the process is invisible to the user, user defined variables that are loaded into I2B2 as additional concepts are user generated. The result is the creation of an augmented dataset via the Eureka Clinical Analytics web browser interface. Users can specify new variables that are derived from the data via PROTEMPA, which runs in the background, and have these derived variables exported to I2B2 where the data can then be queried to create a limited data set that can then be downloaded for analysis in a standard biostatistical package like SAS. An overview of the process is shown in the workflow diagram in Figure 5 below.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"



EMORY UNIVERSITY | Center for Comprehensive Informatics

AIW Workflow

Multiple Databases

Cloned periodically

Analytic Information Warehouse

Periodic data extraction

Data subset, mapped to a standard model

Calculation of derived variables (transform)

Augmented data set

Load into multiple output forms

Make analyses available in existing tools

Figure 5. General workflow for producing an augmented data set from raw data in the AIW that users may then analyze or access in I2B2.

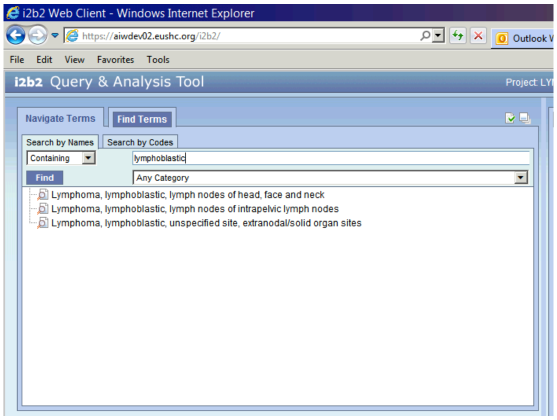
As stated above, the Eureka Clinical Analytics package and all of its components were already under development prior to the creation of this project. However, its development is still pre-production. I worked with Dr. Christopher Flowers as the domain expert providing the specifications and requirements for the development of the tool and performed all of the testing of the application. Changes that were made to the system were implemented by the software engineers of the Center for Comprehensive Informatics at Emory University. To improve the overall functionality and performance of the analytics package and improve its performance, several platform, function, and process, we made the following specific improvements:

- ❖ Platform Improvements
  - ➢ Direct Data connection
    - ▪ During the course of this project it became apparent that the cost of having duplicate copies of the CDW, one for production and one for the AIW, would become prohibitive as data storage needs are increasing exponentially. This forced us to develop the ability of Eureka to access the CDW directly. This functionality has advantages beyond negating the need to have a full clone of the entire CDW in the AIW.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

- Query optimization
  - One of the early problems discovered with the AIW was inefficiency in the extraction process due to the automatic query process generated by PROTEMPA to extract data from the CDW. The software developers that are our partners in creating this system addressed this by optimizing the query generation process to drastically improve efficiencies.
- New features
  - Term search in I2B2 ontology

We made key improvements to the front end in I2B2 to improve functionality and user experience. The first of which was a term search that that allowed users to quickly find the terms they wanted rather than search through the entire ontology tree. This made it much easier to find desired terms for queries (screenshot below in Figure 5).



Containing: lymphoblastic

Any Category

Lymphoma, lymphoblastic, lymph nodes of head, face and neck
Lymphoma, lymphoblastic, lymph nodes of intrapelvic lymph nodes
Lymphoma, lymphoblastic, unspecified site, extranodal/solid organ sites

Figure 6. Screenshot of the ontology term search module in I2B2

- Functional Improvements-
  - Cohort creation with patient list upload

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

- Though not currently available on the front end of the package, we added the facility to create a cohort of patients based on a pre-existing list of patients by uploading a spreadsheet containing the list or target patients to enable creation of patient subsets that would minimize the data extracted from the CDW into I2B2. This is particularly important to reduce the process times for data extraction by pulling data on required variables for a subset of patients rather than the entire patient set.

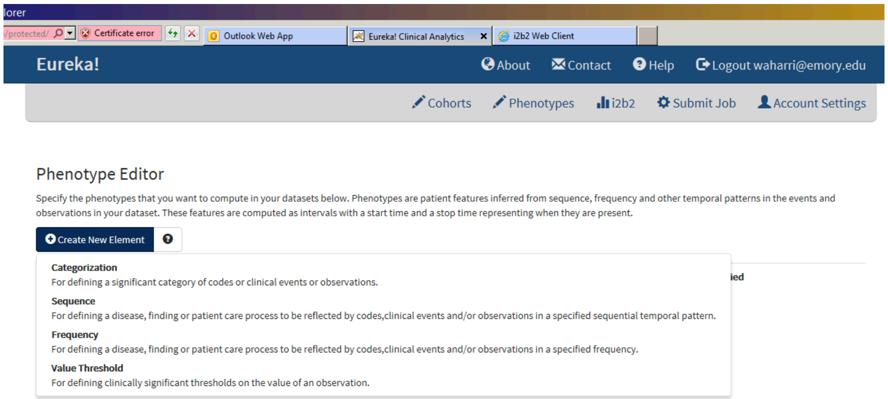
➢ Cohort creation using concepts

- We also added the ability to create a new patient list that defines cohorts of patients prior to extracting data to I2B2 to reduce the size of the dataset to be extracted and thereby reduce processing times. Full implementation is still under development.

➢ New phenotype defining functionality

One our most useful feature additions to the system includes a set of new tools to create more complex user-defined phenotypes characterized by their temporal sequence, their value above or below a threshold, their categorization according to one or several concepts, or their frequency. This is particularly important in the oncology setting where complex phenotypes for chemotherapy treatment regimens or diagnostic sequence in the treatment and monitoring stages of the disease are particularly important (Figure 7).

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"



Phenotype Editor

Specify the phenotypes that you want to compute in your datasets below. Phenotypes are patient features inferred from sequence, frequency and other temporal patterns in the events and observations in your dataset. These features are computed as intervals with a start time and a stop time representing when they are present.

**Categorization**
For defining a significant category of codes or clinical events or observations.

**Sequence**
For defining a disease, finding or patient care process to be reflected by codes,clinical events and/or observations in a specified sequential temporal pattern.

**Frequency**
For defining a disease, finding or patient care process to be reflected by codes,clinical events and/or observations in a specified frequency.

**Value Threshold**
For defining clinically significant thresholds on the value of an observation.

Figure 7. Screenshot of the phenotyping editor in Eureka that allows for producing user defined or derived phenotypes in the I2B2 database.

- Process Improvements
  - Target patient subset identification
    - In looking at solutions to accessing data in the GCCR without requiring informed patient consent or dealing with issues of data governance, we discovered that we could access the cancer registry data being sent to the GCCR via the in-house cancer registry group within our institution. This greatly facilitates our data extraction process because we now have access to a validated list of cancer patients with identifying data and don't have to rely on ICD9 codes or pathology reports to create a list. The Emory Cancer Registry group reviews patient records for all patients coded for cancer diagnosis in the electronic medical record and confirms that diagnosis using pathology reports and annotations before abstracting data on those patients that have a confirmed diagnosis.

  - Creation of ICD mapping ontology

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

- I designed a new ontology data model and functioning ontology for matching and linking disease specific data that may be classified using different classification coding system. The process of building this ontology is described in the following section.

D. Building the Mapping Ontology

As the abstraction ontology demonstrated in the ETL process described earlier, ontologies can facilitate data extraction and integration. Different disease specific databases may represent their data using ICD9 codes as we do in our clinical systems at Emory, or ICD-O-3 codes as is done in typically in cancer registries. If we wish to integrate from both of these sources we need to have a means of mapping data to a common model. One of the most effective ways to do this is using a domain ontology for the different coding systems. This approach provides a great advantage because ontologies are amenable to modification and updating to incorporate new knowledge without affecting the underlying data. Also, ontologies are highly extensible meaning that if we wish to extend its functionality to include alternate terms, obsoleted terms, and many to one or one to many definitions; this can easily be done by adding or changing terms and relationships in the ontology.

In this case, I first started with a hierarchical classification schema based on the REAL/WHO classification system on which the ICD-O-3 system is based (APPENDIX II). This was similar to the data triples that we generated from the ICD-9 hierarchy (Appendix I). I then extended the data in these tables by adding mapped ICD9 and ICD10 codes for diseases specified using the SEER Hematopoietic and Lymphoid Neoplasm Database (http:// http://seer.cancer.gov/seertools/hemelymph) to cross map all of the classification data and building database triples in an excel worksheet using Microsoft Excel. We note that some of the terms in the ICD9 coding system became obsolete in the development of the later ICD10 and ICDO schemas and were not included in the ontology to avoid overlapping of concepts and circularity. Some of the ICDO codes represented expansion of detail on some disease categories

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

represented in the ICD9 coding thus having a one to many rather than a one to one relationship. This and other aspects of the multiple ontologies we would need to create for our integration processes produce a challenge for the I2B2 development team because its functionality does not currently support multiple ontologies.

The data sets we built in Microsoft Excel identify IS-A relationships between ICD codes, disease definitions, and disease categories up the tree to the superclass of neoplasms.  The IS-A relationship identifies parent-child relationships between terms and facilitates codification of knowledge and the relationships between terms in a computer readable structure. This table was the source used to build the ontology data model (Figure 8 below) and the ontology using Protégé 4.3 (http://protege.stanford.edu, Stanford University) allowing us to structure the ontology in a standard software format ontologies.

The resulting ontology data model, with concepts (blue circles), the relation between concepts (blue arrows), and concept properties (red boxes), is shown below (Figure 8) and is structured according to disease code and clinical description. These were structured in separate classes as cancers can be defined clinically in situ with respect to tissue or site, or by their ICD code that may not be unique for that particular disease.

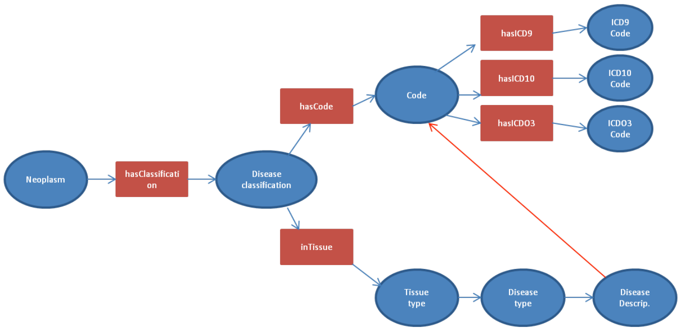
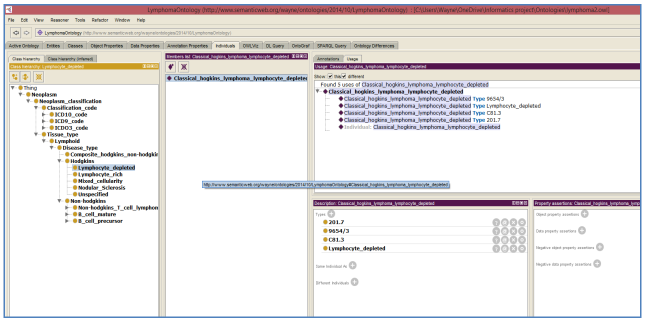


Figure 8. Schematic of the ontology data model for mapping of lymphoma disease classifications. Here we show the hierarchical structure of the ontology and the cross mapping between the disease definitions and the ICD code class.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

The classes represent the categorical groups defining the data and only the disease descriptions are reported as instances in the ontology. So there is a matching class and instance for disease descriptions. The concept properties, or object properties, show the relation between a class and its subclass. The linkage between instances is provided by the disease description class and ICD classes. As shown in Figure 9, using the example of a disease description of  'classical Hodgkin's lymphoma, lymphocyte depleted', we can see that the instance is itself linked to the 4 data types of 'disease description', 'ICD9 code', 'ICD10 code', and 'ICDO3' code producing 5 uses of the term in the ontology.



LymphomaOntology (http://www.semanticweb.org/wayne/ontologies/2014/10/LymphomaOntology) : [C:\Users\Wayne\OneDrive\Informatics project\Ontologies\lymphoma2.owl]

LymphomaOntology (http://www.semanticweb.org/wayne/ontologies/2014/10/LymphomaOntology)

Class hierarchy: Lymphocyte_depleted

```
Thing
  Neoplasm
    Neoplasm_classification
      Classification_code
        ICD10_code
        ICD9_code
        ICDO3_code
      Tissue_type
        Lymphoid
          Disease_type
            Composite_hodgkins_non-hodgkir
            Hodgkins
              Lymphocyte_depleted
              Lymphocyte_rich
              Mixed_cellularity
              Nodular_Sclerosis
              Unspecified
            Non-hodgkins
              Non-hodgkins_T_cell_lymphon
              B_cell_mature
              B_cell_precursor
```

Members list: Classical_hogkins_lymphoma_lymphocyte_depleted

Classical_hogkins_lymphoma_lymphocyte_deplete

http://www.semanticweb.org/wayne/ontologies/2014/10/LymphomaOntology#Classical_hogkins_lymphoma_lymphocyte_depleted

Usage: Classical_hogkins_lymphoma_lymphocyte_depleted

```
Found 5 uses of Classical_hogkins_lymphoma_lymphocyte_depleted
Classical_hogkins_lymphoma_lymphocyte_depleted
    Classical_hogkins_lymphoma_lymphocyte_depleted Type 9654/3
    Classical_hogkins_lymphoma_lymphocyte_depleted Type Lymphocyte_depleted
    Classical_hogkins_lymphoma_lymphocyte_depleted Type C81.3
    Classical_hogkins_lymphoma_lymphocyte_depleted Type 201.7
    Individual: Classical_hogkins_lymphoma_lymphocyte_depleted
```

Description: Classical_hogkins_lymphoma_lymphocyte_depleted

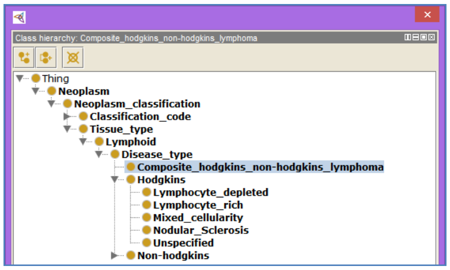
Types
- 201.7
- 9654/3
- C81.3
- Lymphocyte_depleted

Property assertions: Classical_hogkins_lymphoma_lymp

Figure 9 Screenshot of the Protégé application showing the ontology and attribute data for the disease description instance and related details.

Figures 10 and 11 below show the ontology structure and ontology model for Hodgkin's lymphoma produced directly from the new ontology in Protégé. Here the instance of the disease description is not showing because the ontology structure and model only show classes in the ontology. This demonstrates the hierarchical 'is-A' relation between the classes and subclasses, a standard feature of ontologies.  In this way, the elements and instances of the ontology, a specific disease description in this case, is given a clear taxonomic definition, which can include inheritance of the higher-level descriptions in the hierarchy.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"



Class hierarchy: Composite_hodgkins_non-hodgkins_lymphoma

- Thing
  - Neoplasm
    - Neoplasm_classification
      - Classification_code
      - Tissue_type
        - Lymphoid
          - Disease_type
            - Composite_hodgkins_non-hodgkins_lymphoma
            - Hodgkins
              - Lymphocyte_depleted
              - Lymphocyte_rich
              - Mixed_cellularity
              - Nodular_Sclerosis
              - Unspecified
            - Non-hodgkins

Figure.10 Screen shot of the ontology hierarchy in Protégé.

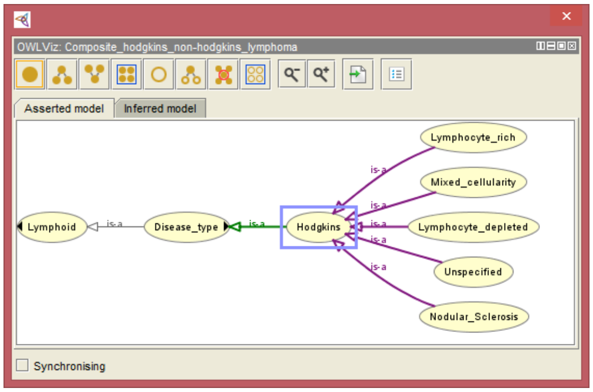
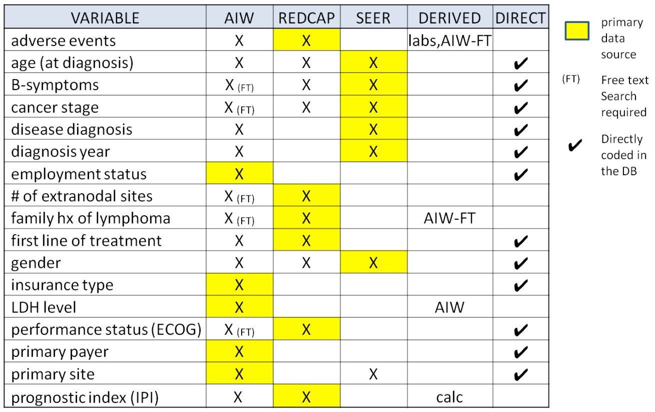


Figure 11. This screen shot shows the relation model for the ontology in protégé for Hodgkin's disease

E. Analysis of variables for data integration

To further evaluate the ability of the proposed patient analytic database to perform as planned, we did an analysis of the data sources for the GA-PAL database (GCCR, AIW, and REDCap forms) to determine if key research variables for outcome analyses were available directly in the database or would be have to be derived from the data after purported integration with AIW data, REDCap form data and registry data. We found (as shown in Table 2) that most of the data is directly coded in the database with very little derivation required.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"



| VARIABLE | AIW | REDCAP | SEER | DERIVED | DIRECT |
|---|---|---|---|---|---|
| adverse events | X | 🟨 X | | labs,AIW-FT | |
| age (at diagnosis) | X | X | 🟨 X | | ✔ |
| B-symptoms | X (FT) | X | 🟨 X | | ✔ |
| cancer stage | X (FT) | X | 🟨 X | | ✔ |
| disease diagnosis | X | | 🟨 X | | ✔ |
| diagnosis year | X | | 🟨 X | | ✔ |
| employment status | 🟨 X | | | | ✔ |
| # of extranodal sites | X (FT) | 🟨 X | | | |
| family hx of lymphoma | X (FT) | 🟨 X | | AIW-FT | |
| first line of treatment | X | 🟨 X | | | ✔ |
| gender | X | X | 🟨 X | | ✔ |
| insurance type | 🟨 X | | | | ✔ |
| LDH level | 🟨 X | | | AIW | |
| performance status (ECOG) | X (FT) | 🟨 X | | | ✔ |
| primary payer | 🟨 X | | | | ✔ |
| primary site | 🟨 X | | X | | ✔ |
| prognostic index (IPI) | X | 🟨 X | | calc | |

🟨 primary data source

(FT) Free text Search required

✔ Directly coded in the DB

Table 2. List of key variables required for outcome studies of lymphoma patients with their source indicated. If variables are not present at source they must be derived from the raw data. The yellow boxes indicate the primary data source for those research variables. "FT" indicates that the data are stored as free text. The check marks indicate data that is directly encoded in the DB. Variables that are not checked must be derived from existing data.

These data represent the combination of different data types from each data source to form an augmented data set as demonstrated for GCCR data and AIW data below where, in this case, data are assumed to be matched using patient demographics (i.e. patient name, age, date of birth, city, social security number, and gender). The data matching facilitates merging of the epidemiologic data from the registry and the clinical setting. (Figure 11 below)

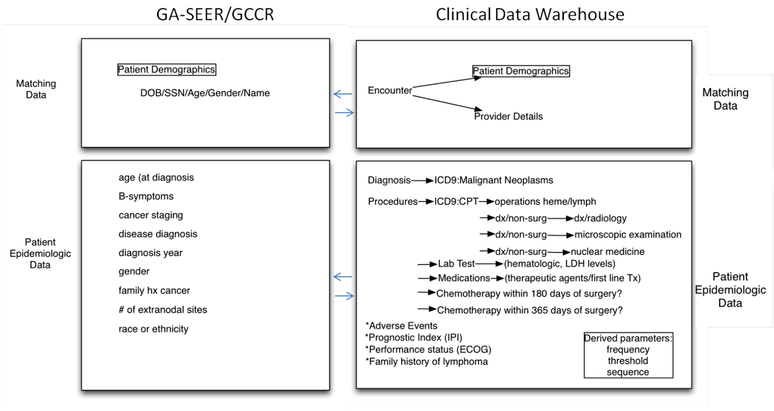

GA-SEER/GCCR
Clinical Data Warehouse
Matching Data
Patient Demographics
DOB/SSN/Age/Gender/Name
Encounter
Patient Demographics
Provider Details
Matching Data
Patient Epidemiologic Data
age (at diagnosis
B-symptoms
cancer staging
disease diagnosis
diagnosis year
gender
family hx cancer
# of extranodal sites
race or ethnicity
Diagnosis→ICD9:Malignant Neoplasms
Procedures→ICD9:CPT→operations heme/lymph
→dx/non-surg→dx/radiology
→dx/non-surg→microscopic examination
→dx/non-surg→nuclear medicine
→Lab Test→(hematologic, LDH levels)
→Medications→(therapeutic agents/first line Tx)
→Chemotherapy within 180 days of surgery?
→Chemotherapy within 365 days of surgery?
*Adverse Events
*Prognostic Index (IPI)
*Performance status (ECOG)
*Family history of lymphoma
Derived parameters: frequency threshold sequence
Patient Epidemiologic Data

Figure 12. Chart showing examples of joining data types and patient data types to be integrated from the GCCR and the AIW.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

Finally, we examined 3 current databases used for concurrent studies at our institution and compared their data types with those of the state cancer registry to evaluate how likely we may be able to match data from those databases to the registry to augment those data with cancer registry data so they can be used for studies beyond which those databases were designed. The DLBCL database likely has too little data detail to enable matching to registry data, however, data from the LungSPORE and NAACCR databases may be amenable to augmentation (Table 3). These studies are all clinical studies for which very little epidemiologic data are being collected. Here again, the rich demographic data in the state cancer registry can be used to not only supplement the data available on the study population, but also to link these data to other epidemiologic data within the SEER database, perhaps even other providers advancing our goal of creating a comprehensive and complete data set for these cancer patients.

Table 3. Comparison of common data types in existing cancer research databases at our institution and the cancer registry

| Candidate fields | GA registry | DLBCL | NAACCR | LungSPORE |
|---|---|---|---|---|
| **Patient-Confidential** | | | | |
| Name—Last | X | | | X |
| Name—First | X | | | X |
| Medical Record Number | X | | | X |
| Social Security Number (MEDICARE LINK) | X | | | |
| **Record ID** | | | | |
| Patient ID Number (same?) | X | | X? | |
| Registry Type | X | | X | |
| NAACCR Record Vision | X | | X | |
| Registry ID (which one?) | X | | X? | |
| FIN Coding System | X | | X | |
| Tumor Record Number | X | | X | |
| Patient System ID- hosp | X | | X | |
| NPI—Registry ID (same?) | X | | X? | |
| **Demographic** | | | | |
| Race Coding Sys—Current | X | X | | X |
| Sex (GENDER) | X | X | | X |

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

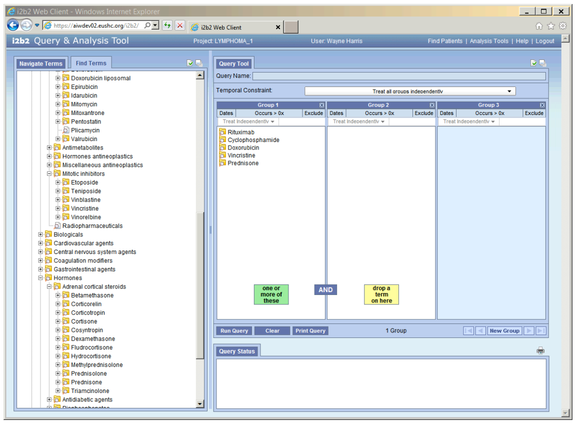
| | | | | |
|---|---|---|---|---|
| Age at Dx | X | | | |
| Birth date | X | | | X |
| **Cancer Diagnosis** | | | | |
| Primary Site (definition?) | X | X? | | |
| Grade | X | | | X |
| **Stage/Prognostic Factors** | | | | |
| SEER Summary Stage 2000 (definitions?) | X | X? | | X |

**4. Results**

Using the methods described above we were able to generate a data set for lymphoma patients within the Emory healthcare system and also a patient cohort defined using I2B2 concepts having received the RCHOP chemotherapy regimen.

Using ICD-9 codes for all lymphoma diagnoses to query all patient data in the AIW covering the period 1992-2012, a subset of 12,491 patients was created. This patient subset was then queried for having received the drug combination for RCHOP chemotherapy (rituximab, cyclophosphamide, vincristine, doxorubicin and prednisone). This was done using the drag and drop functionality of I2B2 where the relevant concepts are simply dragged into the query window, then the query submitted (shown in Figure13 below).

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"



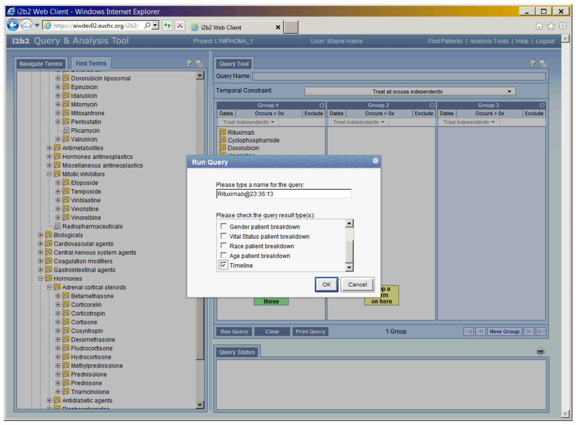
Query Name:

Temporal Constraint: Treat all groups independently

Group 1: Rituximab, Cyclophosphamide, Doxorubicin, Vincristine, Prednisone

Figure 13. Screenshot of I2B2 screen showing the creation of the query-defining patients on RCHOP chemotherapy.

RCHOP regimen would be a proxy for patients that were diagnosed with disease that typically receives this treatment. These would include follicular lymphoma (FL), mantle cell lymphoma (MCL), diffuse large B cell lymphoma (DLBCL) and primary cutaneous B cell lymphoma (PCBCL). (See Appendix III)



Run Query

Please type a name for the query: Rituximab@23:36:13

Please check the query result type(s):

☐ Gender patient breakdown
☐ Vital Status patient breakdown
☐ Race patient breakdown
☐ Age patient breakdown
☑ Timeline

Figure 14. Screenshot of I2B2 screen showing the creation query box that defines the result set required by the user.

Specifying the result set that is required then runs the query. In this case we called the timeline data so we could see temporal data for the drug regimen administered to patients. (See Figure 14 above)  A

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

heterogeneous set of 3082 patients was created, some patients clearly not appearing to fit the phenotype of having received all drugs together. As shown in Figure 14 below, while some patients do appear to have received the RCHOP chemotherapy combination as defined above, even in repeated instances (circled), many did not appear to have received the five drugs at all. (See Figure 15 below) Another piece of evidence that the methodology needs refining is the fact that the ECR data for 1992-2012 reported only about 4,500 patients total as having been diagnosed with some form of lymphoma in EHC hospitals. It is quite unlikely that 68% of the patients will receive RCHOP. These data would suggest that our simple query strategy is grossly overestimating the number of lymphoma patients receiving RCHOP, and perhaps even the total number of lymphoma patients. Since these results are driven by disease definitions using ICD-9 billing codes, it also strongly supports the assertion that the ICD-9 billing codes are simply insufficient to directly process research data and results in an automated fashion.

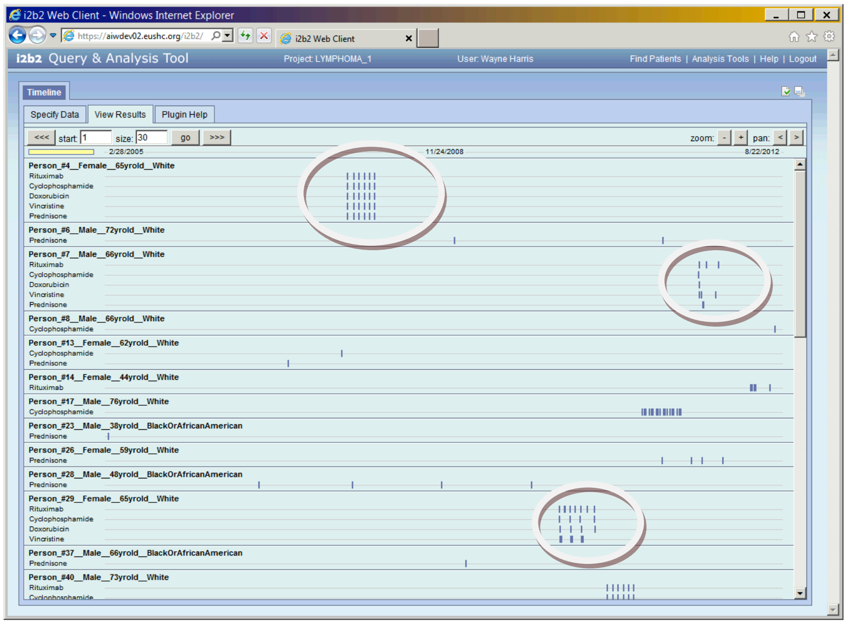

i2b2 Web Client - Windows Internet Explorer
https://aiwdev02.eushc.org/i2b2/
i2b2 Web Client
i2b2 Query & Analysis Tool
Project: LYMPHOMA_1
User: Wayne Harris
Find Patients | Analysis Tools | Help | Logout
Timeline
Specify Data
View Results
Plugin Help
<<<
start: 1
size: 30
go
>>>
zoom: - +
pan: < >
2/28/2005
11/24/2008
8/22/2012
Person_#4__Female__65yrold__White
Rituximab
Cyclophosphamide
Doxorubicin
Vincristine
Prednisone
Person_#6__Male__72yrold__White
Prednisone
Person_#7__Male__66yrold__White
Rituximab
Cyclophosphamide
Doxorubicin
Vincristine
Prednisone
Person_#8__Male__66yrold__White
Cyclophosphamide
Person_#13__Female__62yrold__White
Cyclophosphamide
Prednisone
Person_#14__Female__44yrold__White
Rituximab
Person_#17__Male__76yrold__White
Cyclophosphamide
Person_#23__Male__38yrold__BlackOrAfricanAmerican
Prednisone
Person_#26__Female__59yrold__White
Prednisone
Person_#28__Male__48yrold__BlackOrAfricanAmerican
Prednisone
Person_#29__Female__65yrold__White
Rituximab
Cyclophosphamide
Doxorubicin
Vincristine
Person_#37__Male__66yrold__BlackOrAfricanAmerican
Prednisone
Person_#40__Male__73yrold__White
Rituximab
Cyclophosphamide

Figure 15. Screenshot of I2B2 screen showing the results of the chemotherapy query. The patient result circled is believed to be indicative of patients on the RCHOP chemotherapy having been indicated in the CDW as having received all 5 drugs simultaneously.

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

## 5. Discussion

In preparation for performing various studies on merged clinical administrative and epidemiologic data in the proposed GA-PAL database, we collaborated with the Eureka Clinical Analytics development team of engineers and developers at Emory University's CCI to refine the platform, function and process of the system to produce a robust open source tool that is amenable to clinical, translational and epidemiologic studies on cohorts of lymphoma patients in the state of Georgia. One of the key objectives was to create a system that gives more independence to researchers in collecting, transforming and querying their datasets using the I2B2 platform. Our tool moves toward allowing researchers to participate in the entire process of specifying data, extracting and transforming it, then querying that data for further cohort creation and subsequent analysis. We developed a key set of features that will enhance the performance and usability of the Eureka Clinical Analytics package allowing it to produce user-defined cohorts and datasets that are exported into an I2B2 database. The objectives achieved in this effort were to demonstrate the functionality of the tool in creating the dataset and using data query to refine or subset the data to a target population. We found that we were able to create the user cohort and identify a target subset, however, when we compared this to a validated lymphoma patient list from the hospital's cancer registry, it was apparent that our query strategy needed to be refined to be more selective. It's also quite likely that our cohort creation strategy needs to be improved as it depends solely on ICD-9 codes, which may not be especially effective. Our original strategy in obtaining data from the state cancer registry was to get validated patient data where patient diagnoses were abstracted from the patient record and confirmed. This reduces our reliance on methods using ICD-9 codes to identify patient cohorts. To that end, we designed and implemented an ICD mapping ontology to facilitate matching diagnoses between data obtained from the cancer registry and data existing in the analytic database. Since we were able to get access to patient lists from our local cancer registry to enable lymphoma patient cohort definitions, our strategy needs to be changed to delineate what we can do now with the data readily at hand to create

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

an EHC lymphoma registry, subsequent data quality studies evaluating how well disease diagnoses are coded in our database. It will still be useful to incorporate state registry data as this will be useful in creating linkages to data outside of EHC such as linkage to discharge and Medicare data to which these have already been linked. It may also be useful in gleaning data on EHC patients that was generated outside of our network, such as, at other provider institutions. Our ontology was also built in such a way that it could be extended to add additional elements and serve as a more general ontology for lymphoma or other integration strategies.

Our demonstration of the utility of the new or improved functionality of the system to create cohorts of patients or define patient treatment phenotypes by which new augmented patient sets can be created shows some promise but there is still a lot of refinement to be done to make the system more stable and produce better algorithms to better select target patient cohorts. As we make better progress in this effort we will be able to validate the effectiveness of our approaches by testing our downstream clinical and epidemiologic use cases. If we are successful we will vastly improve the research workflow for lymphoma and also for other research endeavors that share this model, and make a significant step towards developing a tool that will facilitate broad scale data integration for more comprehensive clinical and epidemiologic studies.

Wayne Harris 12/10/14 5:30 PM
**Deleted:**

## References

*Foley J (2011) Report Blasts Problem-Plagued Cancer Research Grid InformationWeel.*
*Fromer M (2012) SIDEBAR: Lessons Learned from caBIG. The ASCO Post 3(6)*
*Graiser M, et al. (2007) Development of query strategies to identify a histologic lymphoma subtype in a large linked database system. Cancer informatics 3:149-58*
*Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG (2009) Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of biomedical informatics 42(2):377-81 doi:10.1016/j.jbi.2008.08.010*
*Huang T, Shenoy PJ, Sinha R, Graiser M, Bumpers KW, Flowers CR (2009) Development of the Lymphoma Enterprise Architecture Database: a caBIG Silver level compliant system. Cancer informatics 8:45-64*

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

*Kohler J, Philippi S, Lange M (2003) SEMEDA: ontology based semantic integration of biological databases. Bioinformatics 19(18):2420-7*

*Kuziemsky CE, Lau F (2010) A four stage approach for ontology-based health information system design. Artificial intelligence in medicine 50(3):133-48 doi:10.1016/j.artmed.2010.04.012*

*Lyman JA, Scully K, Harrison JH, Jr. (2008) The development of health care data warehouses to support data mining. Clinics in laboratory medicine 28(1):55-71, vi doi:10.1016/j.cll.2007.10.003*

*Morton LM, et al. (2007) Proposed classification of lymphoid neoplasms for epidemiologic research from the Pathology Working Group of the International Lymphoma Epidemiology Consortium (InterLymph). Blood 110(2):695-708 doi:10.1182/blood-2006-11-051672*

*Post A, et al. (2011) A Temporal Abstraction-based Extract, Transform and Load Process for Creating Registry Databases for Research. AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science 2011:46-50*

*Post AR, Harrison JH, Jr. (2007) PROTEMPA: a method for specifying and identifying temporal sequences in retrospective data for patient selection. Journal of the American Medical Informatics Association : JAMIA 14(5):674-83 doi:10.1197/jamia.M2275*

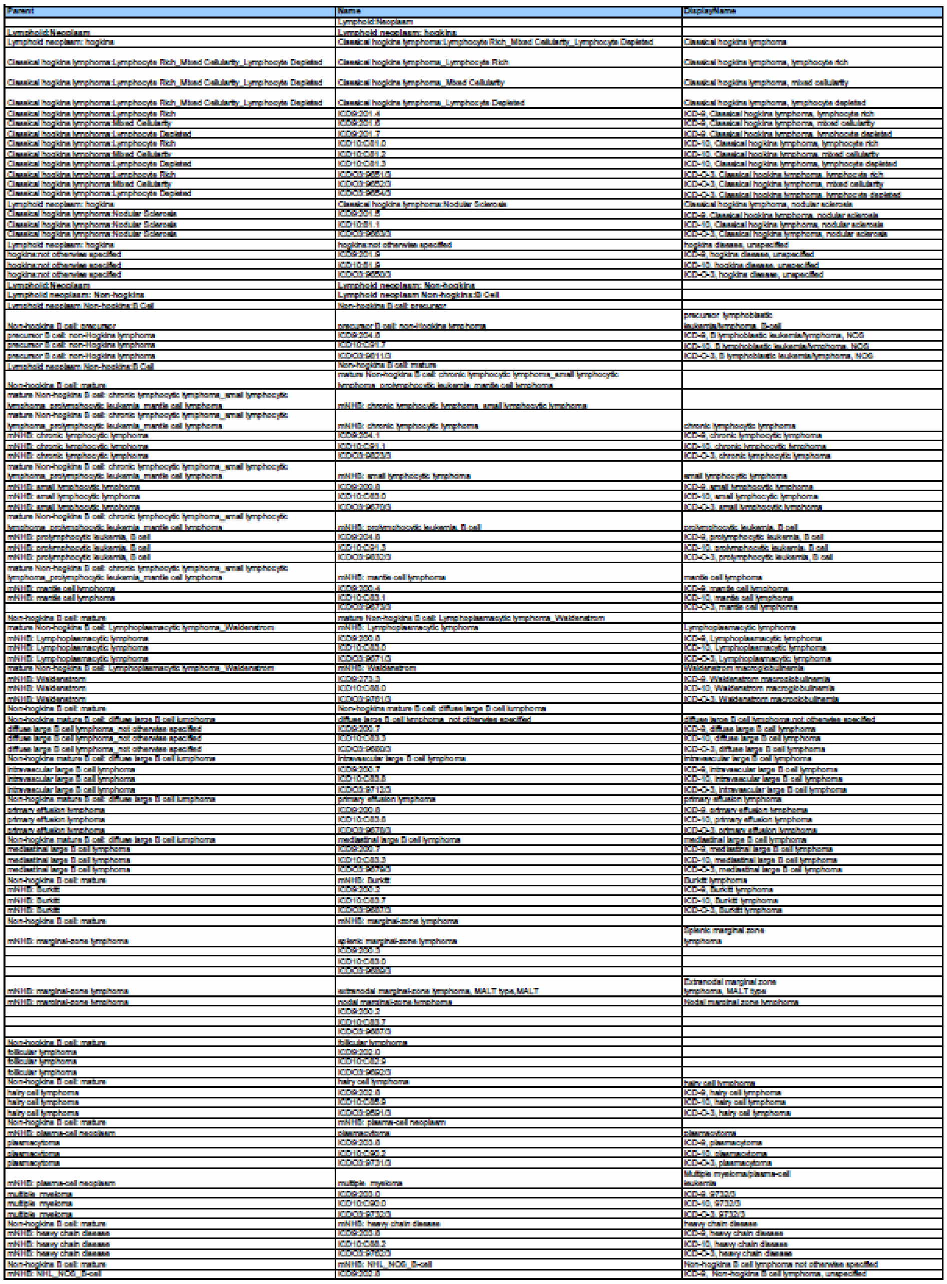
*Post AR, et al. (2013a) Semantic ETL into i2b2 with Eureka! AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science 2013:203-7*

*Post AR, et al. (2013b) Temporal abstraction-based clinical phenotyping with Eureka! AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2013:1160-9*

*Surveillance Research Program N (2014) SEER Stat Fact Sheets: Non-Hodgkin Lymphoma. In. http://seer.cancer.gov/statfacts/html/nhl.html Accessed march 14 2014 2014*

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

# Appendix I - triples



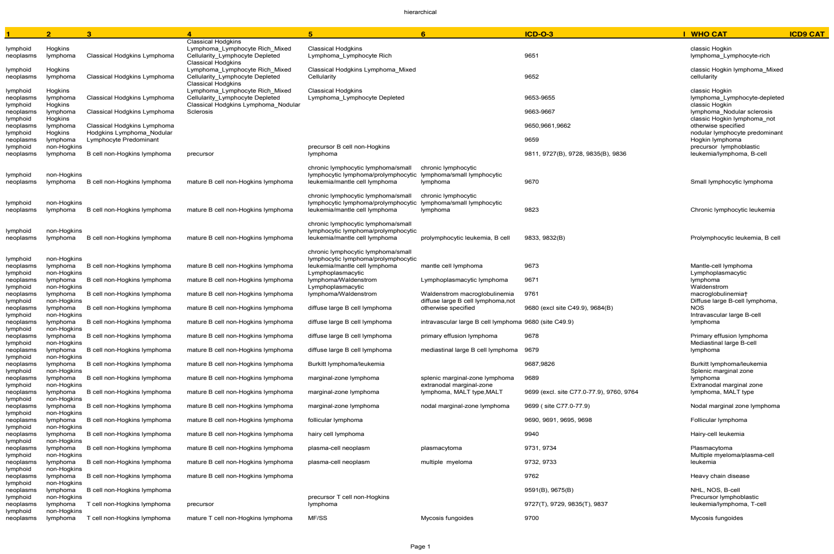
| Parent | Name | DisplayName |
|---|---|---|
| | Lymphoid:Neoplasm | |
| Lymphoid:Neoplasm | Lymphoid neoplasm: hogkins | |
| Lymphoid neoplasm: hogkins | Classical hogkins lymphoma:Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical hogkins lymphoma |
| Classical hogkins lymphoma:Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical hogkins lymphoma_Lymphocyte Rich | Classical hogkins lymphoma, lymphocyte rich |
| Classical hogkins lymphoma:Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical hogkins lymphoma_Mixed Cellularity | Classical hogkins lymphoma, mixed cellularity |
| Classical hogkins lymphoma:Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical hogkins lymphoma_Lymphocyte Depleted | Classical hogkins lymphoma, lymphocyte depleted |
| Classical hogkins lymphoma:Lymphocyte Rich | ICD9:201.4 | ICD-9, Classical hogkins lymphoma, lymphocyte rich |
| Classical hogkins lymphoma:Mixed Cellularity | ICD9:201.6 | ICD-9, Classical hogkins lymphoma, mixed cellularity |
| Classical hogkins lymphoma:Lymphocyte Depleted | ICD9:201.7 | ICD-9, Classical hogkins lymphoma, lymphocyte depleted |
| Classical hogkins lymphoma:Lymphocyte Rich | ICD10:C81.0 | ICD-10, Classical hogkins lymphoma, lymphocyte rich |
| Classical hogkins lymphoma:Mixed Cellularity | ICD10:C81.2 | ICD-10, Classical hogkins lymphoma, mixed cellularity |
| Classical hogkins lymphoma:Lymphocyte Depleted | ICD10:C81.3 | ICD-10, Classical hogkins lymphoma, lymphocyte depleted |
| Classical hogkins lymphoma:Lymphocyte Rich | ICDO3:9651/3 | ICD-O-3, Classical hogkins lymphoma, lymphocyte rich |
| Classical hogkins lymphoma:Mixed Cellularity | ICDO3:9652/3 | ICD-O-3, Classical hogkins lymphoma, mixed cellularity |
| Classical hogkins lymphoma:Lymphocyte Depleted | ICDO3:9654/3 | ICD-O-3, Classical hogkins lymphoma, lymphocyte depleted |
| Lymphoid neoplasm: hogkins | Classical hogkins lymphoma:Nodular Sclerosis | Classical hogkins lymphoma, nodular sclerosis |
| Classical hogkins lymphoma:Nodular Sclerosis | ICD9:201.5 | ICD-9, Classical hogkins lymphoma, nodular sclerosis |
| Classical hogkins lymphoma:Nodular Sclerosis | ICD10:81.1 | ICD-10, Classical hogkins lymphoma, nodular sclerosis |
| Classical hogkins lymphoma:Nodular Sclerosis | ICDO3:9663/3 | ICD-O-3, Classical hogkins lymphoma, nodular sclerosis |
| Lymphoid neoplasm: hogkins | hogkins:not otherwise specified | hogkins disease, unspecified |
| hogkins:not otherwise specified | ICD9:201.9 | ICD-9, hogkins disease, unspecified |
| hogkins:not otherwise specified | ICD10:81.9 | ICD-10, hogkins disease, unspecified |
| hogkins:not otherwise specified | ICDO3:9650/3 | ICD-O-3, hogkins disease, unspecified |
| Lymphoid:Neoplasm | Lymphoid neoplasm: Non-hogkins | |
| Lymphoid neoplasm: Non-hogkins | Lymphoid neoplasm Non-hogkins:B Cell | |
| Lymphoid neoplasm Non-hogkins:B Cell | Non-hogkins B cell: precursor | |
| Non-hogkins B cell: precursor | precursor B cell: non-Hogkins lymphoma | precursor lymphoblastic leukemia/lymphoma, B-cell |
| precursor B cell: non-Hogkins lymphoma | ICD9:204.0 | ICD-9, B lymphoblastic leukemia/lymphoma, NOS |
| precursor B cell: non-Hogkins lymphoma | ICD10:C91.7 | ICD-10, B lymphoblastic leukemia/lymphoma, NOS |
| precursor B cell: non-Hogkins lymphoma | ICDO3:9811/3 | ICD-O-3, B lymphoblastic leukemia/lymphoma, NOS |
| Lymphoid neoplasm Non-hogkins:B Cell | Non-hogkins B cell: mature | |
| Non-hogkins B cell: mature | mature Non-hogkins B cell: chronic lymphocytic lymphoma_small lymphocytic lymphoma_prolymphocytic leukemia_mantle cell lymphoma | |
| mature Non-hogkins B cell: chronic lymphocytic lymphoma_small lymphocytic lymphoma_prolymphocytic leukemia_mantle cell lymphoma | mNHB: chronic lymphocytic lymphoma_small lymphocytic lymphoma | |
| mature Non-hogkins B cell: chronic lymphocytic lymphoma_small lymphocytic lymphoma_prolymphocytic leukemia_mantle cell lymphoma | mNHB: chronic lymphocytic lymphoma | chronic lymphocytic lymphoma |
| mNHB: chronic lymphocytic lymphoma | ICD9:204.1 | ICD-9, chronic lymphocytic lymphoma |
| mNHB: chronic lymphocytic lymphoma | ICD10:C91.1 | ICD-10, chronic lymphocytic lymphoma |
| mNHB: chronic lymphocytic lymphoma | ICDO3:9823/3 | ICD-O-3, chronic lymphocytic lymphoma |
| mature Non-hogkins B cell: chronic lymphocytic lymphoma_small lymphocytic lymphoma_prolymphocytic leukemia_mantle cell lymphoma | mNHB: small lymphocytic lymphoma | small lymphocytic lymphoma |
| mNHB: small lymphocytic lymphoma | ICD9:200.8 | ICD-9, small lymphocytic lymphoma |
| mNHB: small lymphocytic lymphoma | ICD10:C83.0 | ICD-10, small lymphocytic lymphoma |
| mNHB: small lymphocytic lymphoma | ICDO3:9670/3 | ICD-O-3, small lymphocytic lymphoma |
| mature Non-hogkins B cell: chronic lymphocytic lymphoma_small lymphocytic lymphoma_prolymphocytic leukemia_mantle cell lymphoma | mNHB: prolymphocytic leukemia, B cell | prolymphocytic leukemia, B cell |
| mNHB: prolymphocytic leukemia, B cell | ICD9:204.8 | ICD-9, prolymphocytic leukemia, B cell |
| mNHB: prolymphocytic leukemia, B cell | ICD10:C91.3 | ICD-10, prolymphocytic leukemia, B cell |
| mNHB: prolymphocytic leukemia, B cell | ICDO3:9832/3 | ICD-O-3, prolymphocytic leukemia, B cell |
| mature Non-hogkins B cell: chronic lymphocytic lymphoma_small lymphocytic lymphoma_prolymphocytic leukemia_mantle cell lymphoma | mNHB: mantle cell lymphoma | mantle cell lymphoma |
| mNHB: mantle cell lymphoma | ICD9:200.4 | ICD-9, mantle cell lymphoma |
| mNHB: mantle cell lymphoma | ICD10:C83.1 | ICD-10, mantle cell lymphoma |
| | ICDO3:9673/3 | ICD-O-3, mantle cell lymphoma |
| Non-hogkins B cell: mature | mature Non-hogkins B cell: Lymphoplasmacytic lymphoma_Waldenstrom | |
| mature Non-hogkins B cell: Lymphoplasmacytic lymphoma_Waldenstrom | mNHB: Lymphoplasmacytic lymphoma | Lymphoplasmacytic lymphoma |
| mNHB: Lymphoplasmacytic lymphoma | ICD9:200.8 | ICD-9, Lymphoplasmacytic lymphoma |
| mNHB: Lymphoplasmacytic lymphoma | ICD10:C83.0 | ICD-10, Lymphoplasmacytic lymphoma |
| mNHB: Lymphoplasmacytic lymphoma | ICDO3:9671/3 | ICD-O-3, Lymphoplasmacytic lymphoma |
| mature Non-hogkins B cell: Lymphoplasmacytic lymphoma_Waldenstrom | mNHB: Waldenstrom | Waldenstrom macroglobulinemia |
| mNHB: Waldenstrom | ICD9:273.3 | ICD-9, Waldenstrom macroglobulinemia |
| mNHB: Waldenstrom | ICD10:C88.0 | ICD-10, Waldenstrom macroglobulinemia |
| mNHB: Waldenstrom | ICDO3:9761/3 | ICD-O-3, Waldenstrom macroglobulinemia |
| Non-hogkins B cell: mature | Non-hogkins mature B cell: diffuse large B cell lumphoma | |
| Non-hogkins mature B cell: diffuse large B cell lumphoma | diffuse large B cell lymphoma_not otherwise specified | diffuse large B cell lymphoma not otherwise specified |
| diffuse large B cell lymphoma_not otherwise specified | ICD9:200.7 | ICD-9, diffuse large B cell lymphoma |
| diffuse large B cell lymphoma_not otherwise specified | ICD10:C83.3 | ICD-10, diffuse large B cell lymphoma |
| diffuse large B cell lymphoma_not otherwise specified | ICDO3:9680/3 | ICD-O-3, diffuse large B cell lymphoma |
| Non-hogkins mature B cell: diffuse large B cell lumphoma | Intravascular large B cell lymphoma | Intravascular large B cell lymphoma |
| Intravascular large B cell lymphoma | ICD9:200.7 | ICD-9, Intravascular large B cell lymphoma |
| Intravascular large B cell lymphoma | ICD10:C83.8 | ICD-10, Intravascular large B cell lymphoma |
| Intravascular large B cell lymphoma | ICDO3:9712/3 | ICD-O-3, Intravascular large B cell lymphoma |
| Non-hogkins mature B cell: diffuse large B cell lumphoma | primary effusion lymphoma | primary effusion lymphoma |
| primary effusion lymphoma | ICD9:200.8 | ICD-9, primary effusion lymphoma |
| primary effusion lymphoma | ICD10:C83.8 | ICD-10, primary effusion lymphoma |
| primary effusion lymphoma | ICDO3:9678/3 | ICD-O-3, primary effusion lymphoma |
| Non-hogkins mature B cell: diffuse large B cell lumphoma | mediastinal large B cell lymphoma | mediastinal large B cell lymphoma |
| mediastinal large B cell lymphoma | ICD9:200.7 | ICD-9, mediastinal large B cell lymphoma |
| mediastinal large B cell lymphoma | ICD10:C83.3 | ICD-10, mediastinal large B cell lymphoma |
| mediastinal large B cell lymphoma | ICDO3:9679/3 | ICD-O-3, mediastinal large B cell lymphoma |
| Non-hogkins B cell: mature | mNHB: Burkitt | Burkitt lymphoma |
| mNHB: Burkitt | ICD9:200.2 | ICD-9, Burkitt lymphoma |
| mNHB: Burkitt | ICD10:C83.7 | ICD-10, Burkitt lymphoma |
| mNHB: Burkitt | ICDO3:9687/3 | ICD-O-3, Burkitt lymphoma |
| Non-hogkins B cell: mature | mNHB: marginal-zone lymphoma | |
| mNHB: marginal-zone lymphoma | splenic marginal-zone lymphoma | Splenic marginal zone lymphoma |
| | ICD9:200.3 | |
| | ICD10:C83.0 | |
| | ICDO3:9689/3 | |
| mNHB: marginal-zone lymphoma | extranodal marginal-zone lymphoma, MALT type,MALT | Extranodal marginal zone lymphoma, MALT type |
| mNHB: marginal-zone lymphoma | nodal marginal-zone lymphoma | Nodal marginal zone lymphoma |
| | ICD9:200.2 | |
| | ICD10:C83.7 | |
| | ICDO3:9687/3 | |
| Non-hogkins B cell: mature | follicular lymphoma | |
| follicular lymphoma | ICD9:202.0 | |
| follicular lymphoma | ICD10:C82.9 | |
| follicular lymphoma | ICDO3:9690/3 | |
| Non-hogkins B cell: mature | hairy cell lymphoma | hairy cell lymphoma |
| hairy cell lymphoma | ICD9:202.8 | ICD-9, hairy cell lymphoma |
| hairy cell lymphoma | ICD10:C85.9 | ICD-10, hairy cell lymphoma |
| hairy cell lymphoma | ICDO3:9591/3 | ICD-O-3, hairy cell lymphoma |
| Non-hogkins B cell: mature | mNHB: plasma-cell neoplasm | |
| mNHB: plasma-cell neoplasm | plasmacytoma | plasmacytoma |
| plasmacytoma | ICD9:203.8 | ICD-9, plasmacytoma |
| plasmacytoma | ICD10:C90.2 | ICD-10, plasmacytoma |
| plasmacytoma | ICDO3:9731/3 | ICD-O-3, plasmacytoma |
| mNHB: plasma-cell neoplasm | multiple myeloma | Multiple myeloma/plasma-cell leukemia |
| multiple myeloma | ICD9:203.0 | ICD-9, 9732/3 |
| multiple myeloma | ICD10:C90.0 | ICD-10, 9732/3 |
| multiple myeloma | ICDO3:9732/3 | ICD-O-3, 9732/3 |
| Non-hogkins B cell: mature | mNHB: heavy chain disease | heavy chain disease |
| mNHB: heavy chain disease | ICD9:203.8 | ICD-9, heavy chain disease |
| mNHB: heavy chain disease | ICD10:C88.2 | ICD-10, heavy chain disease |
| mNHB: heavy chain disease | ICDO3:9762/3 | ICD-O-3, heavy chain disease |
| Non-hogkins B cell: mature | mNHB: NHL_NOS_B-cell | Non-hogkins B cell lymphoma not otherwise specified |
| mNHB: NHL_NOS_B-cell | ICD9:202.8 | ICD-9, Non-hogkins B cell lymphoma, unspecified |

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

| | | |
|---|---|---|
| mNHB: NHL_NOS_B-cell | ICD10:C85.9 | ICD-10, Non-hogkins B cell lymphoma, unspecified |
| mNHB: NHL_NOS_B-cell | ICDO3:9591/3 | ICD-O-3, Non-hogkins B cell lymphoma, unspecified |
| Lymphoid neoplasm: Non-hogkins | Lymphoid neoplasm Non-hogkins:T Cell | Non-hogkins T cell lymphoma |
| Lymphoid neoplasm Non-hogkins:T Cell | NHT:precursor | |
| NHT:precursor | precursor T cell non-Hogkins lymphoma | Precursor lymphoblastic<br>leukemia/lymphoma, T-cell |
| precursor T cell non-Hogkins lymphoma | ICD9:204.0 | ICD-9, Precursor T-cell lymphoblastic leukemia |
| precursor T cell non-Hogkins lymphoma | ICD10:C91.7 | ICD-10, Precursor T-cell lymphoblastic leukemia |
| precursor T cell non-Hogkins lymphoma | ICDO3:9837/3 | ICD-O-3, Precursor T-cell lymphoblastic leukemia |
| Lymphoid neoplasm Non-hogkins:T Cell | mNHT | |
| mNHT | mNHT:Mycosis fungoides_Sezary syndrome | |
| mNHT:Mycosis fungoides_Sezary syndrome | Mycosis fungoides | Mycosis fungoides |
| Mycosis fungoides | ICD9:202.1 | ICD-9, Mycosis fungoides |
| Mycosis fungoides | ICD10:C84.0 | ICD-10, Mycosis fungoides |
| Mycosis fungoides | ICDO3:9700/3 | ICD-O-3, Mycosis fungoides |
| mNHT:Mycosis fungoides_Sezary syndrome | Sezary syndrome | Sezary syndrome |
| Sezary syndrome | ICD9:202.2 | ICD-9, Sezary syndrome |
| Sezary syndrome | ICD10:C84.1 | ICD-10, Sezary syndrome |
| Sezary syndrome | ICDO3:9701/3 | ICD-O-3, Sezary syndrome |
| mNHT | mNHT: peripheral T cell lymphoma | |
| mNHT: peripheral T cell lymphoma | peripheral T cell lymphoma,not otherwise specified | Peripheral T-cell lymphoma, NOS |
| peripheral T cell lymphoma,not otherwise specified | ICD9:202.7 | ICD-9,Peripheral T-cell lymphoma, NOS |
| peripheral T cell lymphoma,not otherwise specified | ICD10:C84.4 | ICD-10, Peripheral T-cell lymphoma, NOS |
| peripheral T cell lymphoma,not otherwise specified | ICDO3:9702/3 | ICD-O-3, Peripheral T-cell lymphoma, NOS |
| mNHT: peripheral T cell lymphoma | angioimmunoblastic | Angioimmunoblastic T-cell lymphoma |
| angioimmunoblastic | ICD9:202.7 | ICD-9, Angioimmunoblastic T-cell lymphoma |
| angioimmunoblastic | ICD10:C84.4 | ICD-10, Angioimmunoblastic T-cell lymphoma |
| angioimmunoblastic | ICDO3:9705/3 | ICD-O-3, Angioimmunoblastic T-cell lymphoma |
| mNHT: peripheral T cell lymphoma | subcutaneous panniculitis | Subcutaneous panniculitis-like T-cell lymphoma |
| subcutaneous panniculitis | ICD9:202.8 | ICD-9, subcutaneous panniculitis |
| subcutaneous panniculitis | ICD10:C84.5 | ICD-10, subcutaneous panniculitis |
| subcutaneous panniculitis | ICDO3:9708/3 | ICD-O-3, subcutaneous panniculitis |
| mNHT: peripheral T cell lymphoma | anaplastic large cell | Anaplastic large-cell lymphoma, T-cell or null-cell type |
| anaplastic large cell | ICD9:200.6 | ICD-9, anaplastic large cell |
| anaplastic large cell | ICD10:C83.3 | ICD-10, anaplastic large cell |
| anaplastic large cell | ICDO3:9714/3 | ICD-O-3, anaplastic large cell |
| mNHT: peripheral T cell lymphoma | hepatosplenic | Hepatosplenic T-cell lymphoma |
| hepatosplenic | ICD9:202.8 | ICD-9, Hepatosplenic T-cell lymphoma |
| hepatosplenic | ICD10:C84.5 | ICD-10, Hepatosplenic T-cell lymphoma |
| hepatosplenic | ICDO3:9716/3 | ICD-O-3, Hepatosplenic T-cell lymphoma |
| mNHT: peripheral T cell lymphoma | enteropathy | Enteropathy-type T-cell lymphoma |
| enteropathy | ICD9:202.8 | ICD-9, Enteropathy-associated T-cell lymphoma |
| enteropathy | ICD10:C84.5 | ICD-10, Enteropathy-associated T-cell lymphoma |
| enteropathy | ICDO3:9717/3 | ICD-O-3, Enteropathy-associated T-cell lymphoma |
| mNHT: peripheral T cell lymphoma | cutaneous T, not otherwise specified | Cutaneous T-cell lymphoma, NOS |
| cutaneous T, not otherwise specified | ICD9:202.8 | ICD-9, Cutaneous T-cell lymphoma, NOS |
| cutaneous T, not otherwise specified | ICD10:C84.5 | ICD-9, Cutaneous T-cell lymphoma, NOS |
| cutaneous T, not otherwise specified | ICDO3:9709/3 | ICD-9, Cutaneous T-cell lymphoma, NOS |
| mNHT: peripheral T cell lymphoma | primary cutaneous anaplastic large-cell lymphoma | Primary cutaneous anaplastic large-cell lymphoma |
| primary cutaneous anaplastic large-cell lymphoma | ICD9:202.7 | ICD-9, Primary cutaneous anaplastic large-cell lymphoma |
| primary cutaneous anaplastic large-cell lymphoma | ICD10:C84.4 | ICD-10, Primary cutaneous anaplastic large-cell lymphoma |
| primary cutaneous anaplastic large-cell lymphoma | ICDO3:9718/3 | ICD-O-3, Primary cutaneous anaplastic large-cell lymphoma |
| mNHT | mNHT: ATLL | Adult T-cell leukemia/lymphoma |
| mNHT: ATLL | ICD9:204.8 | ICD-9, Adult T-cell leukemia/lymphoma |
| mNHT: ATLL | ICD10:C91.5 | ICD-10, Adult T-cell leukemia/lymphoma |
| mNHT: ATLL | ICDO3:9827/3 | ICD-O-3, Adult T-cell leukemia/lymphoma |
| mNHT | mNHT: NK/T-cell lymphoma | NK/T-cell lymphoma, nasaltype/ aggressive NK-cell leukemia |
| mNHT: NK/T-cell lymphoma | ICD9:202.8 | ICD-9, NK/T-cell lymphoma |
| mNHT: NK/T-cell lymphoma | ICD10:C84.5 | ICD-10, NK/T-cell lymphoma |
| mNHT: NK/T-cell lymphoma | ICDO3:9719/3 | ICD-O-3, NK/T-cell lymphoma |
| mNHT | mNHT: large granular lymphocytic lymphoma | T-cell large granular lymphocytic leukemia |
| mNHT: large granular lymphocytic lymphoma | ICD9:204.8 | ICD-9, T-cell large granular lymphocytic leukemia |
| mNHT: large granular lymphocytic lymphoma | ICD10:C91.7 | ICD-10, T-cell large granular lymphocytic leukemia |
| mNHT: large granular lymphocytic lymphoma | ICDO3:9831/3 | ICD-O-3, T-cell large granular lymphocytic leukemia |
| mNHT | mNHT: T-PLL | Prolymphocytic leukemia, T-cell |
| mNHT: T-PLL | ICD9:204.8 | ICD-9, Prolymphocytic leukemia, T-cell |
| mNHT: T-PLL | ICD10:C91.3 | ICD-10, Prolymphocytic leukemia, T-cell |
| mNHT: T-PLL | ICDO3:9834/3 | ICD-O-3, Prolymphocytic leukemia, T-cell |
| Lymphoid neoplasm Non-hogkins:T Cell | NHL_NOS_T-cell | non hogkins lymphoma T cell , unspecified |
| NHL_NOS_T-cell | ICD9: 202.8 | ICD-9, non hogkins lymphoma T cell , unspecified |
| NHL_NOS_T-cell | ICD10: C85.9 | ICD-10, non hogkins lymphoma T cell , unspecified |
| NHL_NOS_T-cell | ICDO3: 9591/3 | ICD-O-3, non hogkins lymphoma T cell , unspecified |
| Lymphoid neoplasm Non-hogkins:T Cell | precursor | T Lymphoblastic Leukemia/ lymphoma |
| precursor | ICD9: 204.8 | ICD-9, |
| precursor | ICD10: C91.7 | ICD-10, |
| precursor | ICDO3: 9837/3 | ICD-O-3, |
| Lymphoid Neoplasm | Lymphoid neoplasm: Composite hogkins_Non-hogkins | Composite hogkins_Non-hogkins lymphoma |
| Lymphoid neoplasm: Composite hogkins_Non-hogkins | Composite Hodgkin/NHL unknown lineage | Composite Hodgkin/NHL unknown lineage |
| | ICD9: 202.8 | ICD-9, Composite Hodgkin/NHL unknown lineage |
| | ICD10: C85.9 | ICD-10, Composite Hodgkin/NHL unknown lineage |
| | ICDO3: 9596/3 | ICD-O-3, Composite Hodgkin/NHL unknown lineage |

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

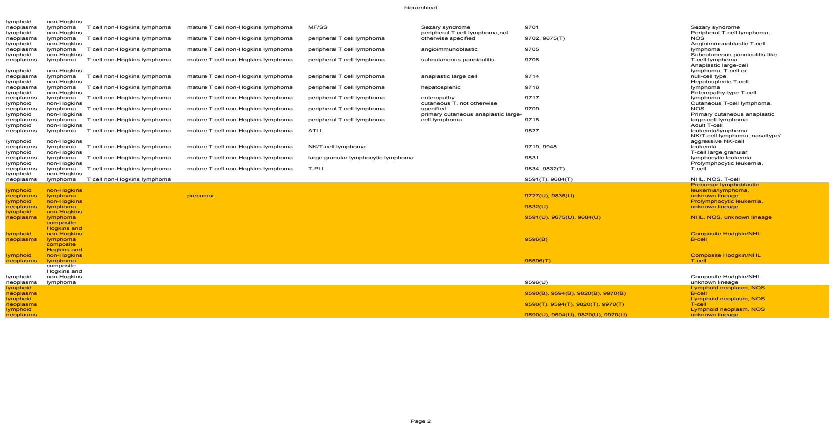
# Appendix II – Morton Classification Schema



hierarchical

| 1 | 2 | 3 | 4 | 5 | 6 | ICD-O-3 | I | WHO CAT | ICD9 CAT |
|---|---|---|---|---|---|---|---|---|---|
| lymphoid neoplasms | Hogkins lymphoma | Classical Hodgkins Lymphoma | Classical Hodgkins Lymphoma_Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical Hodgkins Lymphoma_Lymphocyte Rich | | 9651 | | classic Hogkin lymphoma_Lymphocyte-rich | |
| lymphoid neoplasms | Hogkins lymphoma | Classical Hodgkins Lymphoma | Classical Hodgkins Lymphoma_Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical Hodgkins Lymphoma_Mixed Cellularity | | 9652 | | classic Hogkin lymphoma_Mixed cellularity | |
| lymphoid neoplasms | Hogkins lymphoma | Classical Hodgkins Lymphoma | Classical Hodgkins Lymphoma_Lymphocyte Rich_Mixed Cellularity_Lymphocyte Depleted | Classical Hodgkins Lymphoma_Lymphocyte Depleted | | 9653-9655 | | classic Hogkin lymphoma_Lymphocyte-depleted | |
| lymphoid neoplasms | Hogkins lymphoma | Classical Hodgkins Lymphoma | Classical Hodgkins Lymphoma_Nodular Sclerosis | | | 9663-9667 | | classic Hogkin lymphoma_Nodular sclerosis | |
| lymphoid neoplasms | Hogkins lymphoma | Classical Hodgkins Lymphoma | | | | 9650,9661,9662 | | classic Hogkin lymphoma_not otherwise specified | |
| lymphoid neoplasms | Hogkins lymphoma | Hodgkins Lymphoma_Nodular Lymphocyte Predominant | | | | 9659 | | nodular lymphocyte predominant Hogkin lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | precursor | precursor B cell non-Hogkins lymphoma | | 9811, 9727(B), 9728, 9835(B), 9836 | | precursor lymphoblastic leukemia/lymphoma, B-cell | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | chronic lymphocytic lymphoma/small lymphocytic lymphoma/prolymphocytic leukemia/mantle cell lymphoma | chronic lymphocytic lymphoma/small lymphocytic lymphoma | 9670 | | Small lymphocytic lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | chronic lymphocytic lymphoma/small lymphocytic lymphoma/prolymphocytic leukemia/mantle cell lymphoma | chronic lymphocytic lymphoma/small lymphocytic lymphoma | 9823 | | Chronic lymphocytic leukemia | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | chronic lymphocytic lymphoma/small lymphocytic lymphoma/prolymphocytic leukemia/mantle cell lymphoma | prolymphocytic leukemia, B cell | 9833, 9832(B) | | Prolymphocytic leukemia, B cell | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | chronic lymphocytic lymphoma/small lymphocytic lymphoma/prolymphocytic leukemia/mantle cell lymphoma | mantle cell lymphoma | 9673 | | Mantle-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | Lymphoplasmacytic lymphoma/Waldenstrom | Lymphoplasmacytic lymphoma | 9671 | | Lymphoplasmacytic lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | Lymphoplasmacytic lymphoma/Waldenstrom | Waldenstrom macroglobulinemia | 9761 | | Waldenstrom macroglobulinemia† | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | diffuse large B cell lymphoma | diffuse large B cell lymphoma,not otherwise specified | 9680 (excl site C49.9), 9684(B) | | Diffuse large B-cell lymphoma, NOS | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | diffuse large B cell lymphoma | intravascular large B cell lymphoma | 9680 (site C49.9) | | Intravascular large B-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | diffuse large B cell lymphoma | primary effusion lymphoma | 9678 | | Primary effusion lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | diffuse large B cell lymphoma | mediastinal large B cell lymphoma | 9679 | | Mediastinal large B-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | Burkitt lymphoma/leukemia | | 9687,9826 | | Burkitt lymphoma/leukemia | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | marginal-zone lymphoma | splenic marginal-zone lymphoma | 9689 | | Splenic marginal zone lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | marginal-zone lymphoma | extranodal marginal-zone lymphoma, MALT type,MALT | 9699 (excl. site C77.0-77.9), 9760, 9764 | | Extranodal marginal zone lymphoma, MALT type | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | marginal-zone lymphoma | nodal marginal-zone lymphoma | 9699 ( site C77.0-77.9) | | Nodal marginal zone lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | follicular lymphoma | | 9690, 9691, 9695, 9698 | | Follicular lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | hairy cell lymphoma | | 9940 | | Hairy-cell leukemia | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | plasma-cell neoplasm | plasmacytoma | 9731, 9734 | | Plasmacytoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | plasma-cell neoplasm | multiple myeloma | 9732, 9733 | | Multiple myeloma/plasma-cell leukemia | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | mature B cell non-Hogkins lymphoma | | | 9762 | | Heavy chain disease | |
| lymphoid neoplasms | non-Hogkins lymphoma | B cell non-Hogkins lymphoma | | | | 9591(B), 9675(B) | | NHL, NOS, B-cell | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | precursor | precursor T cell non-Hogkins lymphoma | | 9727(T), 9729, 9835(T), 9837 | | Precursor lymphoblastic leukemia/lymphoma, T-cell | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | MF/SS | Mycosis fungoides | 9700 | | Mycosis fungoides | |

Page 1



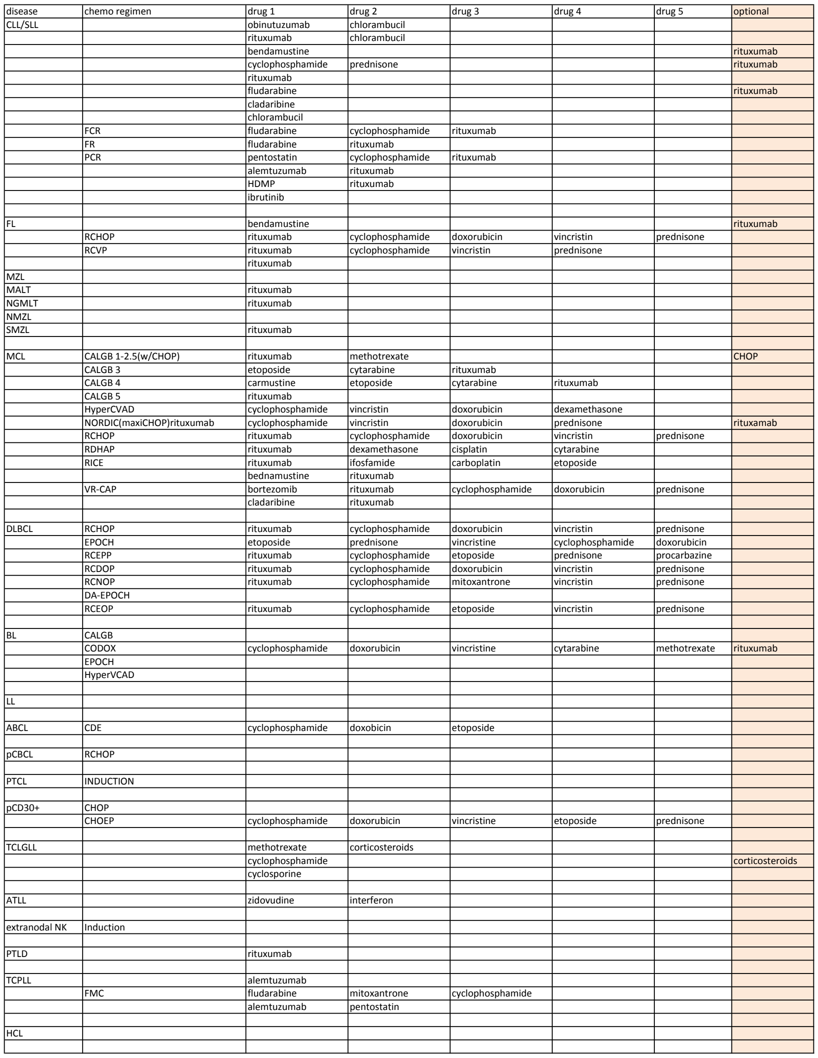
hierarchical

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | MF/SS | Sezary syndrome | 9701 | | Sezary syndrome | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | peripheral T cell lymphoma,not otherwise specified | 9702, 9675(T) | | Peripheral T-cell lymphoma, NOS | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | angioimmunoblastic | 9705 | | Angioimmunoblastic T-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | subcutaneous panniculitis | 9708 | | Subcutaneous panniculitis-like T-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | anaplastic large cell | 9714 | | Anaplastic large-cell lymphoma, T-cell or null-cell type | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | hepatosplenic | 9716 | | Hepatosplenic T-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | enteropathy | 9717 | | Enteropathy-type T-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | cutaneous T, not otherwise specified | 9709 | | Cutaneous T-cell lymphoma, NOS | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | peripheral T cell lymphoma | primary cutaneous anaplastic large-cell lymphoma | 9718 | | Primary cutaneous anaplastic large-cell lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | ATLL | | 9827 | | Adult T-cell leukemia/lymphoma | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | NK/T-cell lymphoma | | 9719, 9948 | | NK/T-cell lymphoma, nasaltype/ aggressive NK-cell leukemia | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | large granular lymphocytic lymphoma | | 9831 | | T-cell large granular lymphocytic leukemia | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | mature T cell non-Hogkins lymphoma | T-PLL | | 9834, 9832(T) | | Prolymphocytic leukemia, T-cell | |
| lymphoid neoplasms | non-Hogkins lymphoma | T cell non-Hogkins lymphoma | | | | 9591(T), 9684(T) | | NHL, NOS, T-cell | |
| lymphoid neoplasms | non-Hogkins lymphoma | | precursor | | | 9727(U), 9835(U) | | Precursor lymphoblastic leukemia/lymphoma, unknown lineage | |
| lymphoid neoplasms | non-Hogkins lymphoma | | | | | 9832(U) | | Prolymphocytic leukemia, unknown lineage | |
| lymphoid neoplasms | non-Hogkins lymphoma | | | | | 9591(U), 9675(U), 9684(U) | | NHL, NOS, unknown lineage | |
| lymphoid neoplasms | composite Hogkins and non-Hogkins lymphoma | | | | | 9596(B) | | Composite Hodgkin/NHL B-cell | |
| lymphoid neoplasms | composite Hogkins and non-Hogkins lymphoma | | | | | 96596(T) | | Composite Hodgkin/NHL T-cell | |
| lymphoid neoplasms | composite Hogkins and non-Hogkins lymphoma | | | | | 9596(U) | | Composite Hodgkin/NHL unknown lineage | |
| lymphoid neoplasms | | | | | | 9590(B), 9594(B), 9820(B), 9970(B) | | Lymphoid neoplasm, NOS B-cell | |
| lymphoid neoplasms | | | | | | 9590(T), 9594(T), 9820(T), 9970(T) | | Lymphoid neoplasm, NOS T-cell | |
| lymphoid neoplasms | | | | | | 9590(U), 9594(U), 9820(U), 9970(U) | | Lymphoid neoplasm, NOS unknown lineage | |

Page 2

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right: 0.25"

Appendix III – Listing of Chemotherapy Regimens for Non-Hodgkin's Lymphoma (based on the NCCN guidelines)



| disease | chemo regimen | drug 1 | drug 2 | drug 3 | drug 4 | drug 5 | optional |
|---|---|---|---|---|---|---|---|
| CLL/SLL | | obinutuzumab | chlorambucil | | | | |
| | | rituxumab | chlorambucil | | | | |
| | | bendamustine | | | | | rituxumab |
| | | cyclophosphamide | prednisone | | | | rituxumab |
| | | rituxumab | | | | | |
| | | fludarabine | | | | | rituxumab |
| | | cladaribine | | | | | |
| | | chlorambucil | | | | | |
| | FCR | fludarabine | cyclophosphamide | rituxumab | | | |
| | FR | fludarabine | rituxumab | | | | |
| | PCR | pentostatin | cyclophosphamide | rituxumab | | | |
| | | alemtuzumab | rituxumab | | | | |
| | | HDMP | rituxumab | | | | |
| | | ibrutinib | | | | | |
| | | | | | | | |
| FL | | bendamustine | | | | | rituxumab |
| | RCHOP | rituxumab | cyclophosphamide | doxorubicin | vincristin | prednisone | |
| | RCVP | rituxumab | cyclophosphamide | vincristin | prednisone | | |
| | | rituxumab | | | | | |
| MZL | | | | | | | |
| MALT | | rituxumab | | | | | |
| NGMLT | | rituxumab | | | | | |
| NMZL | | | | | | | |
| SMZL | | rituxumab | | | | | |
| | | | | | | | |
| MCL | CALGB 1-2.5(w/CHOP) | rituxumab | methotrexate | | | | CHOP |
| | CALGB 3 | etoposide | cytarabine | rituxumab | | | |
| | CALGB 4 | carmustine | etoposide | cytarabine | rituxumab | | |
| | CALGB 5 | rituxumab | | | | | |
| | HyperCVAD | cyclophosphamide | vincristin | doxorubicin | dexamethasone | | |
| | NORDIC(maxiCHOP)rituxumab | cyclophosphamide | vincristin | doxorubicin | prednisone | | rituxamab |
| | RCHOP | rituxumab | cyclophosphamide | doxorubicin | vincristin | prednisone | |
| | RDHAP | rituxumab | dexamethasone | cisplatin | cytarabine | | |
| | RICE | rituxumab | ifosfamide | carboplatin | etoposide | | |
| | | bednamustine | rituxumab | | | | |
| | VR-CAP | bortezomib | rituxumab | cyclophosphamide | doxorubicin | prednisone | |
| | | cladaribine | rituxumab | | | | |
| | | | | | | | |
| DLBCL | RCHOP | rituxumab | cyclophosphamide | doxorubicin | vincristin | prednisone | |
| | EPOCH | etoposide | prednisone | vincristine | cyclophosphamide | doxorubicin | |
| | RCEPP | rituxumab | cyclophosphamide | etoposide | prednisone | procarbazine | |
| | RCDOP | rituxumab | cyclophosphamide | doxorubicin | vincristin | prednisone | |
| | RCNOP | rituxumab | cyclophosphamide | mitoxantrone | vincristin | prednisone | |
| | DA-EPOCH | | | | | | |
| | RCEOP | rituxumab | cyclophosphamide | etoposide | vincristin | prednisone | |
| | | | | | | | |
| BL | CALGB | | | | | | |
| | CODOX | cyclophosphamide | doxorubicin | vincristine | cytarabine | methotrexate | rituxumab |
| | EPOCH | | | | | | |
| | HyperVCAD | | | | | | |
| | | | | | | | |
| LL | | | | | | | |
| | | | | | | | |
| ABCL | CDE | cyclophosphamide | doxobicin | etoposide | | | |
| | | | | | | | |
| pCBCL | RCHOP | | | | | | |
| | | | | | | | |
| PTCL | INDUCTION | | | | | | |
| | | | | | | | |
| pCD30+ | CHOP | | | | | | |
| | CHOEP | cyclophosphamide | doxorubicin | vincristine | etoposide | prednisone | |
| | | | | | | | |
| TCLGLL | | methotrexate | corticosteroids | | | | |
| | | cyclophosphamide | | | | | corticosteroids |
| | | cyclosporine | | | | | |
| | | | | | | | |
| ATLL | | zidovudine | interferon | | | | |
| | | | | | | | |
| extranodal NK | Induction | | | | | | |
| | | | | | | | |
| PTLD | | rituxumab | | | | | |
| | | | | | | | |
| TCPLL | | alemtuzumab | | | | | |
| | FMC | fludarabine | mitoxantrone | cyclophosphamide | | | |
| | | alemtuzumab | pentostatin | | | | |
| | | | | | | | |
| HCL | | | | | | | |
| | | | | | | | |

Wayne Harris 12/10/14 5:35 PM
**Formatted:** Right:  0.25"