Medication Recommendations vs. Peer Practice in Pediatric Levothyroxine Dosing - A Study of Collective Intelligence from a Clinical Data Warehouse as a Potential Model for Clinical Decision Support

by

Elisabeth Lee Scheufele

B.S. Biochemistry
Georgetown University, 1998

Doctorate of Medicine
Georgetown University School of Medicine, 2002

SUBMITTED TO THE HARVARD-M.I.T. DIVISION OF HEALTH SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN BIOMEDICAL INFORMATICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2009

Signature of Author _____
Harvard-MIT Division of Health Science and Technology
January 30, 2009

Certified by: _____
Anil K. Dubey, M.D., M.S.
Instructor of Medicine, Harvard Medical School
Thesis Supervisor

Approved by: _____
Ram Sasisekharan, Ph.D.
Director, HST
Edward Hood Taplin Professor of Health Science & Technology and Biological Engineering

Medication Recommendations vs. Peer Practice in Pediatric Levothyroxine Dosing – A Study of Collective Intelligence from a Clinical Data Warehouse as a Potential Model for Clinical Decision Support

by

Elisabeth Lee Scheufele

Submitted to the Harvard-MIT Division of Health Science and Technology
on January 30, 2009, in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Biomedical Informatics

ABSTRACT

Clinical decision support systems (CDSS) are developed primarily from knowledge gleaned from evidence-based research, guidelines, trusted resources and domain experts. While these resources generally represent information that is research proven, time-tested and consistent with current medical knowledge, they lack some qualities that would be desirable in a CDSS. For instance, the information is presented as generalized recommendations that are not specific to particular patients and may not consider certain subpopulations. In addition, the knowledge base that produces the guidelines may be outdated and may not reflect real-world practice. Ideally, resources for decision support should be timely, patient-specific, and represent current practice. Patient-oriented clinical decision support is particularly important in the practice of pediatrics because it addresses a population in constant flux. Every age represents a different set of physiological and developmental concerns and considerations, especially in medication dosing patterns. Patient clinical data warehouses (CDW) may be able to bridge the knowledge gap. CDWs contain the collective intelligence of various contributors (i.e. clinicians, administrators, etc.) where each data entry provides information regarding medical care for a patient in the real world. CDWs have the potential to provide information as current as the latest upload, be focused to specific subpopulations and reflect current clinical practice. In this paper, I study the potential of a well-known patient clinical data warehouse to provide information regarding pediatric levothyroxine dosing as a form of clinical decision support. I study the state of the stored data, the necessary data transformations and options for representing the data to effectively summarize and communicate the findings. I also compare the resulting transformed data, representing actual practice within this population, against established dosing recommendations. Of the transformed records, 728 of the 854 (85.2%, [95% confidence interval 82.7:87.6]) medication records contained doses that were under the published recommended range for levothyroxine. As demonstrated by these results, real world practice can diverge from established recommendations. Delivering this information on real-world peer practice medication dosing to clinicians in real-time offers the potential to provide a valuable supplement to established dosing guidelines, enhancing the general and sometimes static dosing recommendations.

Thesis Supervisor: Anil K. Dubey, M.D.
Title: Instructor of Medicine, Harvard Medical School

TABLE OF CONTENTS Page

ABSTRACT 2
INTRODUCTION 5
BACKGROUND 6
Clinical decision support system 6
Clinical data warehouses 7
Collective intelligence 8
Pediatric medication dosing 9
Levothyroxine 10
GOALS 11
MATERIALS & METHODS, OBSERVATIONS & RESULTS 12
Scope 12
Phase I: Retrieving, Transforming and Presenting Data 13
Clinical patient data warehouse 13
The IRB 13
Data acquisition 14
Unanticipated issues discovered in the initial data mining exercises ... 14
Query burden 14
Dates 15
Transferring tables 15
The MS Sql Server queries 15
Medication table 15
Weight table 16
Data transformation 16
Linking the weight and medication tables 16
Addressing initial redundancies 17
Determining intended dosage 17
Removal of redundancies 20
Examples of data representation 21

TABLE OF CONTENTS – cont'd Page

Consolidating the demographic data 27

The time interval between the weight and the medication event 27

Phase II: Analysis of data..... 29

General 30

As per age group 30

DISCUSSION 38

Data Acquisition and Transformation 39

Data structure of the medication records 39

The age attribute 39

Query size 41

Issues when combining the weight and medication table 41

Interval between dates of weight and medication records 42

Consolidating the data 42

Peer practice data representation 44

Analysis of Data 44

LIMITATIONS 45

FUTURE DIRECTIONS 47

CONCLUSION 48

ACKNOWLEDGEMENTS 50

APPENDIX 51

INTRODUCTION

The practice of medicine is catching up to the age of computers with the implementation of electronic medical records (EMR) and expansion of EMR functionalities. The purpose of the EMR has evolved from a computerized information recording and retrieval system to an interactive ordering system and is currently delving into the realm of becoming an intelligent system that provides decision support to clinicians at various points in patient care. For the most part, clinical decision support systems (CDSS) are based on knowledge from evidence-based research, guidelines, trusted resources and domain experts. Basically, they are static predetermined logical algorithms generalized to particular medical issues. These types of decision support can fall victim to particular drawbacks: they require timely review and modifications, are not specific to individual patients, may not represent specific subpopulations, and do not always represent real world practices.

An ideal CDSS would provide recommendations based on the most up to date information available and focus it to the patient at hand. EMRs are eliciting, daily, an incredible amount of information from clinicians regarding patient attributes and disease findings as well as tests ordered and treatments prescribed. Most of this information is stored in clinical databases. The model of collective intelligence (CI) has significant potential to provide decision support that is both timely and patient specific through point of care (POC) database querying and data analysis. In this paper, I study the potential of CI using data from a well-established clinical data warehouse on the clinical question of pediatric levothyroxine dosing. I conduct the study by evaluating the state of the data in the data warehouse and the necessary steps towards optimal data transformation, comparing the findings to established dosing recommendations, and demonstrating a few methods of data representation. As a result of this study, I discuss the steps necessary to develop a clinical decision support system to harness the collective intelligence of peer practice medication dosing patterns present in a clinical data warehouse.

BACKGROUND

Clinical decision support systems

Clinical decision support systems (CDSS) supplement the electronic medical record with the goal of providing the clinical user appropriate medical decision making knowledge specific to the patient. CDSS have many definitions[1,2,3,4], but in particular Osheroff, et. al. says, "Clinical decision support (CDS) refers broadly to providing clinicians or patients with clinical knowledge and patient-related information, intelligently filtered or presented at appropriate times, to enhance patient care. Clinical knowledge of interest could include simple facts and relationships, established best practices for managing patients with specific disease states, new medical knowledge from clinical research and many other types of information." CDSS can take on many forms, including: alerts and reminders, diagnostic assistance, therapy critiquing and planning, prescribing decision support systems, information retrieval, and image recognition and interpretation.[5] For the most part, CDSS have had a positive influence on patient medical management. Coiera states that "there is now a body of research which provides good evidence of the effectiveness of CDSS, specifically computerized medication order entry systems, in increasing the safety of patients by reducing errors, adverse events, and by increasing the proportion of appropriate and safe prescribing decisions."[5] While CDSS is delivering on the expectation of improvements in patient safety and effectiveness in patient medical management, there is still significant potential for more complex, and timely, patient focused decision support.

---

[1] http://www.himss.org/ASP/topics_clinicalDecision.asp - Accessed 01-28-2009

[2] Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., et al. (2001). Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc*, 8(6), 527-534.

[3] *Crossing the quality chasm: A new health system for the 21st century*(2001). Washington, D.C. ; National Academy Press, c2001.

[4] Osheroff,MD, FACP, FA, Jerry, Pifer, M., Eric, Teich,FACMI, FHIMSS,, Jonathan, Sittig,PhD, FACMI, Dean, & MD,MS, FACP, Robert Jenders. (2005). *Improving outcomes with clinical decision support: An implementer's guide* (1st ed.) Productivity Press.

[5] Coiera, E. (2003). *Guide to health informatics* London : Arnold ; 2003.

Clinical data warehouses

Clinical data repositories are databases that collect and store patient care information from a variety of data sources, optimized for storage and retrieval of data on particular patients.[6] Clinical data warehouses (CDW) are similar to repositories except, in CDWs, the data is optimized for long-term storage and retrieval. CDWs are also a primary resource for data mining, benchmarking and other forms of safety- and quality-related analysis. CDWs may contain data specific to a single institution, or can encompass a much larger group, with data acquired at the regional, national or international level.[6] The data in CDWs hold significant potential to elaborate on the medical practices of clinicians[7], through the practice of data mining and data modeling.

Data mining is "a method for obtaining useful information from large databases and includes data collection, extraction, manipulation and summarization, as well as analysis."[6] Through data mining and subsequent transformation of the data, patterns and models can be derived from the collective intelligence of peer practice stored in the CDW. These models may in turn provide a means of decision support for clinician users. Usually, models are built on a case by case basis. The steps to go from the data collection phase to the transformation phase to the modeling phase and then to the presentation phase require significant effort and time. However, in the case of medicine, certain concepts have similar overall structure. For example, although no two medications are exactly alike, they share similar prescribing components. In general, medications have indications as well as name, dose or dosing range, route of administration, units of measure, frequency and duration. The structured nature of medications can be translated into a model which can be generalized into algorithms that allow users to query the database for peer practice prescribing behavior.

---

[6] Aspden, P. (2004). *Patient safety: Achieving a new standard for care* Washington, D.C. : National Academies Press, c2004.

[7] Cimino, J. D., & Shortliffe, E. H. (2006). *Biomedical informatics: Computer applications in health care and biomedicine*. New York : Springer.

Collective intelligence

The MIT Center for Collective Intelligence defines collective intelligence as "groups of individuals doing things collectively that seem intelligent."[8] An example of collective intelligence resides in nature with bees whose individual actions summatively impact the survival of the hive as a whole.[9] It is not so much the individual effort that affords survival, but rather the collective actions of relative successes and insignificant failures that affect the survival of the colony. MIT CCI lists a more human centric example of case studies demonstrating that users of a product are better sources of innovation for a company than their own researchers.[8] Basically, the laypeople who use the product in the real world are better at using it than the researchers who are supposed to be the experts.

Another example of collective intelligence is the Internet, where the contributions of users afford the survival of the World Wide Web (WWW). As millions of users contribute, knowledge is added to the WWW and a collective intelligence is emerging. With the evolution to Web 2.0, various applications are harnessing this knowledge, and bringing to surface the collective intelligence stored within.[10] The website Amazon.com is a thriving on-line shopping site in part because it employs a recommendation engine based on the collective intelligence of the users. The engine successfully assimilates the buying history of all their customers, and makes recommendations to particular users based on not only their buying habits, but also on users with similar buying habits.[11] Cloudmark, a collaborative spam filtering application, builds its knowledge base from the input of individual email users as to what they mark as spam. It works so well, that it outperforms other spam filtering applications that evaluate the actual message content.[10] When knowledge from many individuals is added into a communal environment, there is incredible potential for collective intelligence to emerge. Clinical data warehouses can be thought of as a communal environment of medical management information by various clinicians and administrators. With the amount of data stored in CDWs, the potential for collective intelligence is immense.

---

[8] http://cci.mit.edu/about/MaloneLaunchRemarks.html - Accessed 01-28-2009
[9] Surowiecki, J. (2005). *The wisdom of crowds* Anchor.
[10] http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=2 -Accessed 01-28-2009
[11] Segaran, T. (2007). *Programming collective intelligence: Building smart web 2.0 applications* O'Reilly Media, Inc.

Pediatric medication dosing

A distinctly challenging issue in the practice of pediatrics is prescribing medications to patients whose physiology is constantly developing and maturing. Prescribing practices in adult and pediatric medicine have some similarities. They both depend on the indication, as well as pharmacokinetics like hepatic and renal function. However, almost all pediatric medication dosing take into account the physiological development of the child. This is done through the measures of age, weight and surface area as surrogate markers for physiological development.

The mental workaround regarding pediatric medication prescribing is somewhat complicated for the pediatric clinician. It requires knowledge of the appropriate range of dosing for a particular medication for particular age ranges and then calculating the dose by weight or surface area. Sometimes the range of dosing for a particular medication is wide, and the recommendations do not specify how to accommodate this range. Take for example the recommendations for cefazolin (see Figure 1).

Dosage recommendations for cefazolin

Neonate IM, IV:

Postnatal age ≤ 7 days: 40 mg/kg/24 hr ÷ Q12 hr
Postnatal age >7 days:
≤2000 g: 40 mg/kg/24 hr ÷ Q12 hr
>2000 g: 60 mg/kg/24 hr ÷ Q8 hr

Infant >1 mo/child: 50–100 mg/kg/24 hr ÷ Q8 hr IV/IM; max. dose: 6 g/24 hr
Adult: 2–6 g/24 hr ÷ Q6–8 hr IV/IM; max. dose: 12 g/24 hr

Figure 1. Cefazolin dosage recommendations, from the Harriet Lane[19].

There are several vague components to this medication guideline. One of the age ranges indicates "Infant >1 mo/child", with the next age range at "Adult". Adults are usually defined as 18 year old, so is a child up to 18 years in age? The max dosage for the "Infant > 1/child" age range is maxed out at 6g/24hours, while for the adult, the max dosage is 12g/24hours. If the patient, such as an adolescent, is 60kg or more, the child range max dose is already reached. Also, the dose range is quite wide, where the upper bound is 2x the lower bound. Such a large range can be concerning, because it may lead to uneven dosing practices if certain clinicians lean

toward the low side, and others lean toward the high side. Also, medications for children are often prescribed off label for which dosing recommendations not available[12], which could lead to even more erratic dosing practices. A major component of pediatric training is to learn the subtlety of dosing for a core set of medications. But once training is complete and the pediatric clinician works in the real world, they may not have colleagues or supervising attendings to help figure out the usual method of dosing unfamiliar medications. For such situations, having a resource that provides insight into real world usage or peer practice would be very helpful.

Through the knowledge stored in clinical data warehouses, the user can query the actions of others to determine the real world practices of prescribing medications, particularly those that are unfamiliar to them. This can also be extended for use in adult medication prescribing; however, the added complexity of calculating the dosage from a range makes this concept particularly attractive in the practice of pediatrics. The concept of collective intelligence can also be further extended for use by helping users determine patterns of diagnosis, monitoring and treatment with certain conditions. The collective intelligence of patient management as recorded in clinical data warehouses has significant potential to provide insight into the real world application of various treatment and diagnostic techniques not formally addressed in the usual reference tools.

Levothyroxine

Levothyroxine is a synthetic hormone whose primary indication is the treatment of hypothyroidism. In the pediatric population, hypothyroidism is addressed in two particular scenarios: congenital and acquired.[13]

Congenital hypothyroidism affects newborns, at a rate of 1/4000 worldwide, with a female/male ratio of 2:1.[13] For the most part, it is identified through newborn screening, primarily because it presents with non-specific signs and any delay in treatment can result in catastrophic neurological deficits. The treatment is life-long thyroid hormone replacement with levothyroxine and is monitored through testing the blood for thyroid stimulating hormone (TSH) and thyroxine (T4) levels. Failure to diagnose quickly and to attain a euthyroid state can result in

[12] Kozer, E., Berkovitch, M., & Koren, G. (2006). Medication errors in children. *Pediatric Clinics of North America, 53*(6), 1155-1168.
[13] Kliegman, R., Behrman, R. E., Jenson, H. B., & Stanton, B. F. (2007). *Nelson textbook of pediatrics* Saunders.

devastating consequences; however early diagnosing from newborn screening programs have led to decreased rates of mental and growth retardation.

Acquired hypothyroidism affects children beyond the newborn stage, at a rate of 0.3% with a female/male ratio of 2:1.[13] It is diagnosed in children who present with signs and symptoms such as failure to thrive, deceleration in growth, goiter, constipation, cold intolerance and decreased energy. The treatment is life-long thyroid hormone replacement with levothyroxine and is also monitored through testing the blood for appropriate TSH and T4 levels.

The recommendations of the Pediatric Dosage Handbook[14] can be found in Figure 2.

Levothyroxine dosing for children

Age Range	[mcg/kg/dose]
0-3 months	10-15
>3-6 months	8-10
>6-12 months	6-8
>1-5years	5-6
>6-12	4-5
>12	2-3
after growth/puberty	1.7

Figure 2. Pediatric levothyroxine dosing, as per the Pediatric Dosage Handbook.[14]

Levothyroxine was an ideal medication for this study because the recommendations specify distinct age groups and distinct medication dosage ranges, allowing for unambiguous compliance analysis of the data from the CDW with the recommendations.

GOALS

The first goal in this paper is to perform a proof of concept study to simulate a clinical decisions support system using the collective intelligence of a well known clinical data warehouse to address the question of real world practices regarding pediatric levothyroxine dosing. I accomplish this goal by doing the following:

- Study the quality of the data in the CDW.
- Study the transformation necessary to present the intended dosing by using only the data retrieved through querying the CDW.

[14] Taketomo, C. K., Hodding, J. H., & Kraus, D. M. (2008). *Pediatric dosage handbook: Including neonatal dosing, drug administration, & extemporaneous preparations (pediatric dosage handbook)* Lexi-Comp.

- Suggest methods of data representation.

The second goal in this paper is to study the data to determine whether there is any significant difference between the initial data and the transformed data, and whether the data demonstrates compliance with medication dosing recommendations.

The third goal is to discuss the barriers to implementations uncovered in this study and make recommendations on how to address them in order to develop a clinical decision support system of medication dosing recommendations based on the collective intelligence of real world clinical data.

MATERIALS & METHODS, OBSERVATIONS & RESULTS

Scope

This study is conducted in two phases: 1) data retrieval, transformation and presentation and 2) analysis of the initial and transformed dosing information against the dosing recommendations. In the first phase, the data warehouse is queried to obtain the necessary data. Records for weights and medication prescriptions are retrieved and transformed into doses by weight that the clinician intended using only the information retrieved from the CDW. Once the data is transformed, methods of data representation are described. In the second phase, the initially joined (raw) data is compared with the consolidated data to evaluate the utility of transforming the original data. The data is also studied for compliance with pediatric levothyroxine dosing recommendations because substantial compliance or non-compliance can impact the utility of presenting such information to the user as a form of clinical decision support.

Phase I: Retrieving, Transforming and Presenting Data

Clinical patient data warehouse

The RPDR is a well known and well utilized clinical patient data warehouse. It collects data from 6 Boston area hospitals.[15] The RPDR data warehouse is built on a modified star schema, with a central "fact table", linked to other tables that store records for medical "concepts", "diagnoses", "encounters", "demographics", etc. The RPDR has well over 1 billion unique patient "facts." The database is built on the 2005 MS Sql Server framework, and requires institutional approval for direct querying and data retrieval.

The RPDR is particularly unique because it provides two methods through which the data can be queried. Approved users knowledgeable in the Sql query language can perform queries directly with the database. On the other hand, novice users, without any knowledge of the query language, can query the database through the user friendly interface, the QueryTool.[16] It allows the user to build queries through a drag and drop style interface which returns the number of people in the query cohort as well as some general characteristics of the group. If the user is satisfied by the results, they can then fill out an IRB and request the deidentified data for use in their research. The QueryTool is mentioned here because an analogous program can be developed to provide decision support based on modified query findings. Instead of querying to return the size of the cohort, the query can return data that can be programmatically transformed into decision support recommendations.

The IRB

The IRB approves access to deidentified patient data in the RPDR clinical data warehouse for the purpose of clinical research. IRB #:2002P000381.

---

[15] http://www.partners.org/rescomputing/template.asp?pageid=99&ArticleTitle=RPDR&level1ID=9&tocID=9&articleSubPage=true - Accessed 01-28-2009

[16] Murphy, S. N., Gainer, V., & Chueh, H. C. (2003). A visual interface designed for novice users to find research patient cohorts in a large biomedical database. *AMIA ...Annual Symposium Proceedings / AMIA Symposium.AMIA Symposium,* , 489-493.

Data acquisition

Initial data mining exercises with the RPDR were performed from 9/2008 through 11/2008 in order to investigate whether the appropriate data for patient weight and medication details were available to perform the study and to optimize the query request in terms of run-time and quality of data retrieved. After initial exercises and data manipulation established that there was sufficient data to pursue this study, I built the optimal queries, and performed them on 12-03-2008.

Unanticipated issues discovered in the initial data mining exercises

Age

The RPDR records the patient's age in a demographic table as static value according to the age of the patient at the time of the data upload into the CDW. Basically, the stored age is equivalent to the "current" age of the patient. However, the age of interest is the patient's age at the time of the event. Therefore, the stored age is inappropriate for this study. In order to obtain the appropriate value, the age was calculated from the "date of birth" in the "demographics table" and the "date of the event" from the "fact table". There is no established function in the MS Sql Server application to calculate age in years from two dates, so sql code was necessary to perform this function, (see Appendix - Query A for the code).

Query burden

The optimal query would obtain all the levothyroxine prescriptions, and join them with the patient's most recent weight value relative to the date of the medication prescription. Unfortunately, the query developed for the MS Sql Server was quite unwieldy, and did not complete in less than several hours. Thus, it was necessary to deconstruct the query into smaller components and then manipulate the resultant tables in a separate database application (MS Access). This decreased the query run-time from hours to minutes (see "Linking the weight and medication tables" section for further details.).

Dates

There were some issues with interoperability of the data between MS Sql Server and MS Access applications. In particular, the date attribute did not transfer well from the server to the database. Thus, as a surrogate for time, the patient's age in days at the time of the event was used. Since time according to each patient's own timeline was significant, the dates of the events were not as important as the time between certain events, such as number of days between the weight and the medication event. When querying the RPDR, sql code was incorporated to add a column to calculate the patient's age in days for each event by calculating the difference between the patient's "date of birth" and the "start date" of the event. However, if it was necessary to study yearly prescribing patterns, then representing dates in an alternate way would be necessary. For example, numerical components of the dates can be parsed out into separate columns, or the limitations of transferring dates between the applications can be further investigated.

Transferring tables

There was a delimiter issue when exporting tables from MS Sql Server to MS Access. The query results from the MS Sql Server were saved as comma-separated values (CSV) files. However, there were many instances where commas were part of the text in the record. Importing a CSV file with these unaccounted commas into MS Access caused import errors by inappropriately shifting data across columns. To deal with this issue, the commas were replaced in all the fields with a semi-colon during the query to the CDW. This action satisfactorily resolved the issue.

The MS Sql Server queries

Medication table

The RPDR was queried (see Appendix - Query A) to obtain levothyroxine medication records for patients between the ages of 0 and 17 years at the time of the event. This resulted in 1299 records, from 438 individual patients. There are a couple issues regarding the query. First, the database does not store all the individual components of the medication prescription as structured fields in the table. Only the name, the dose, and the dose unit were available. However, each medication record included the attribute of HL7_text which contained the

expected medication details in XML (extensible mark-up language) format. From this attribute, the necessary components of the medication prescriptions were parsed out for use in this study.

Weight table

The RPDR was queried (see Appendix - Query B) was performed to obtain weights for patients between the ages of 0 and 17 years at the time of the event. This resulted in 602771 records, from 100,005 individual patients. Since there were cases where there was a record for weight, but no value, the query was limited to records where the values were not "NULL".

Data transformation

The goal was to transform the data into a format that is analogous to the recommendations from the PDH[14] (dosage in mcg per weight per day) in order to compare the results of the data warehouse to the recommendations. This transformation also allowed the data to be presented in a manner that could supplement the established recommendations. Several steps were required. The first step was to join the medication prescription table to the weight table to determine which medication records could be linked to appropriate weight records. The next step was to transform the medication details into the intended 'daily dose' for each record, using only the data retrieved from the data warehouse. The final step was to calculate the medication dose per weight per day with the consolidated data. Several of these steps had some interesting and unexpected issues.

Linking the weight and medication tables

Linking the weight and medication tables required more than simply joining them by patient id. Since there could be several weight and medication values for any particular patient, it was important to isolate the appropriate weight value for each medication record. In particular, it would be optimal to simulate the usual pediatric practice of using the most recent weight, either on the same date or the most recent prior to the date of the medication record.

Linking the medication record with the most recent weight relative to the medication prescription date required four simple queries (see Appendix - Queries D1-4). The first query determined all the combinations of medication and weight encounters, joined on the attribute of

patient id. It also calculated the number of days between the dates, with the weight date as earlier or on the same day as the medication date (leaving off any negative days). The next query singled out distinct medication encounters with the least number of days between the med and weight encounter. The third query joined all of the medication records and the weight records by patient id and included a column for the number of days between the medication and the weight date. The last query joined the third query results with the second query results, based on medication encounter, and the least number of days between the weight and medication encounter. The resulting table contained 1088 records, representing 352 patients. Next, this table of data required transformation to address redundancies and determine intended dosage.

Addressing initial redundancies

Of these records, 24 unique medication encounters represented 48 redundant medication records. The cause for the redundancy was that for some medication records, the date of the most recent weight record had two weight records for the same date. Of the 24, 16 encounters had the same weight for the same day recorded twice. For this situation, one of the two encounters was randomly selected for expulsion. Of the remaining, 6 of them had two different but plausible weights. For these, the higher value was chosen because it seemed logical that the higher value was a simple remeasurement of the weight. And for the final 2 encounters, the two values were quite extreme, so a judgment call was necessary on what was more appropriate. One child was 79 days old, whose weight values were 11 and 22 lbs. For this age, the weight of 11 lbs was more plausible. The other child was a 14 year old whose weight values were either 112 or 148 lbs. Since this child was an adolescent, and patients with hypothyroidism can present with significant weight, the higher value was chosen as the appropriate weight.

Determining intended dosage

These records were then reviewed using MS Excel to decipher the intended "daily dose" of levothyroxine from the medication details. In order to optimally represent the "daily dose", the intent of the clinician had to be determined from the data retrieved. This required reviewing each prescription and interpreting the data stored in each attribute of the medication detail.

There are some assumptions to the type of information stored in the details of a prescription. The attributes of "name", "dose", "units", "route" and "frequency" can be somewhat structured, while the attribute of "directions" is less structured, and usually free-text. Sometimes the combination of the name, dose, units, route and frequency is all that is necessary to determine the intended dose because either there are no directions (Consolidated Data (CD): 69.7%, Raw Data (RD): 68.9 %; see Figure 3 for summary statistics), or the text in the directions do not indicate modification to the structured portion of the prescription (ie. "take as directed", "no substitutions", "3 month supply"; CD: 21.4%, RD: 21.4%; see Figure 3 for summary statistics, and Appendix - Figures D1-4 for examples). This would be an ideal situation for developing algorithms to query for dosing information.

Summary Table of Medication Details

Medication name	Raw	Consolidated
Generic name as part of Medication name of record	1084 (99.6)	850 (99.5)
No Generic name	4 (0.4)	4 (0.5)

Units	Raw	Consolidated
MCG only	1049 (96.4)	822 (96.3)
MCG as part of free text	35 (3.2)	28 (3.3)
MG or Other	4 (0.4)	4 (0.5)

Frequency	Raw	Consolidated
QD	940 (86.4)	743 (87.0)
Structured other (ie. QOD, QAM, BID, etc.)	61 (5.6)	32 (3.7)
Free text	87 (8.0)	79 (9.3)

Directions	Raw	Consolidated
No Directions	750 (68.9)	595 (69.7)
Free text- No affect on structured Rx	233 (21.4)	183 (21.4)
Free text- Potential for effect on structured Rx	105 (9.7)	76 (8.9)

Simple Structured	Raw	Consolidated
Total with combination of 'PO', 'MCG, 'QD', and 'No Directions'	667 (61.3)	532 (62.3)

Figure 3. Summary of type of text found in HL7 medication details. Number found (percentage of total). Total in Raw Data is 1088. Total in Consolidated Data is 854. See Appendix- Figures A, B, C, D1-4 for specific examples for the listed medication details.

In other cases, the directions further elaborate on how to take the medication, and may completely change the intended dose as interpreted only from the structured form (CD: 8.9%, RD: 9.7%; see Figure 3 for summary statistics, and Appendix - Figures D1-4 for examples). For instance, a prescription can be written as "levothyroxine 150 mcg PO QD, take half a pill a day". While the structured format indicates 150mcg daily, further directions indicate that the actual dosing is 75mcg daily. In many cases, the text doubled the dose, halved the dose, or specified much more complicated dosing regimes that would be difficult to capture by conventionally structured means. While this is somewhat expected with the attribute of "directions", similar but unexpected issues arose with some of the other medication components.

The other attributes parsed out from the HL7 text did not always contain information in the expected structured form. The "name" attribute had 4 instances that did not include the generic name of levothyroxine (see Figure 3 for summary statistics and Appendix - Figure A for examples). The frequency field mostly had "QD" (CD: 87.0%, RD 86.4%; see Figure 3 for summary statistics, and Appendix - Figure C for examples). But there were instances of other types of structured frequencies, such as BID, QOD, etc. (CD: 3.7%, RD: 5.6%), and even free-text (CD: 9.3%, RD: 8.0%). While the overwhelming majority of dose_units only had "mcg" (CD: 96.3%, RD: 96.4%, see Figure 3 for summary statistics and Appendix - Figure B for examples), there were instances of free-text in the dose_units field as well (CD: 3.3%, RD, 3.2%). In the raw data group, there were 61% of medication records had the expected structural data in the form of "mcg PO QD" without any further directions. In order to consolidate the remaining data, I worked through each medication record, and interpreted to the best of my clinical abilities the intended daily dose.

For the most part, I followed the instructions and calculated what would amount to a daily dose of the medication. If the structured dose was to be given twice a day, then the daily dose would be twice the structured dose. If the structured dose was give once every other day, then the daily dose was half the structured dose. If there were dosing clarifications in any of the fields, then I followed them as closely as possible to come up with an average daily dose. There were two records where "MG" instead of "MCG" was listed, so the dose required multiplication by 1000. Of the consolidated data, there was one record with the sig of 112mcg IM, with "IM"

(intramuscular route of administration) as the medication route rather than "PO". The record was kept because the dose of 112mcg appeared to be appropriate, and the "IM" was most likely in error.

To evaluate for outliers, the data from the initial join was graphed as the dose per weight (mcg/kg/day) vs. age (in days; see Figures 16 and 17). Points that appeared to be visually separated from the main group were investigated. There were three such points. In one case, the dose recorded was 625 mcg for a 16 years old patient who weighed 101 lbs. The most likely error is the dosing value. The other two points are for the same patient, who was a 12 year old with a prescription for 75mcg, but had weight values of 9.68 lbs. The error was most likely the recorded weight. There is not enough information in the data to infer actual dose or weight without either looking to other records, or performing a chart review, so it was decided to remove these three points from the cohort.

Removal of redundancies

The final step in consolidating the data involved removing duplicates once the data was cleaned. The simplest way to accomplish this was to utilize a table creation functionality in MS Access (see Appendix - Figure E for instructions). The records with redundant daily dose values for the same patient with the same date were removed. In order to accomplish this, a multifield key was built on the attributes of: patient age in days for the medication record, the derived intended dose, the patient's weight value. The primary issue with this maneuver is that medication encounter id values of the removed records would be lost. However, since such information was not useful to the question at hand, this is a reasonable action to take. Overall, the data transformation resulted in decreasing the number of levothyroxine medication records available from 1299 to 1088 (83.8%) after the initial join to 854 (75.8%) after the data transformation (see Figure 4 for the effects of the manipulation on the data in the various age groups). The final transformation still left records with some ambiguity. Five records have 3 different medication daily dosing records for a unique patient-date combination, and 42 records have 2 medication records. Unfortunately, the appropriate dose is difficult to determine with the information available in the data retrieved. The primary options are either to remove them as a whole, or simply keep them. For this study, these records were kept.

Age ranges	Original	Raw	Consolidated
0-3 months	61	58 (95.1%)	40 (65.6%)
4-6 months	18	17 (94.4%)	15 (83.3%)
7-12 months	31	29 (93.5%)	18 (58.1%)
1-5 years	210	178 (84.8%)	138 (65.7%)
6-11 years	272	254 (93.4%)	211 (77.6%)
> 11 years	707	552 (78.1%)	432 (61.1%)
Total	1299	1088 (83.8%)	854 (75.8%)

Figure 4. Number of levothyroxine medication records present in the original data warehouse query, the initial joining (Raw) of the medication table with the weight table and the transformed data (Consolidated) (percentage of the original).

Examples of data representation

A simple concept for presenting data would be to present the frequency of doses. One of the current mainstream applications of collective intelligence is in the form of word clouds, which is basically a method of presenting the frequency of the same or similar words. The size of the facts represents the frequency so that the higher number of facts, the larger the text for that fact. One idea would be to put together "dosage clouds" of the doses used for a particular medication for a particular patient cohort. For this study, this was accomplished through the use of an online application from the website tagcrowd.com.

First, the consolidated medication records for children of a particular age range were isolated using MS Access. The data was modified to round the dose per kg per day to the nearest tenth. The value was concatenated with the term "mcg/kg/day". Unfortunately, the application recognizes periods and forward slashes as separaters, so the decimal points were replaced with a letter (in this case a 'p') and the forward slash marks were excluded. The results were entered into the Tagcrowd application to develop a "dosage cloud". The application also provides HTML code of the cloud, which was copied and pasted into an HTML editor. The text was edited so the dosages adequately presented decimal points and forward slashes. Also, to contrast

the usage of clouds, a more structured representation of the frequency was developed as horizontal bar charts using the R statistical program. The bar charts simply represented the rounded doses by weight per day vs. the frequency of the dose in the age range of interest.

Figure 5 is a dosage cloud of patients who are 12 to 17 years old. The most frequent dose is 1.1 mcg/kg/dose (41 records).

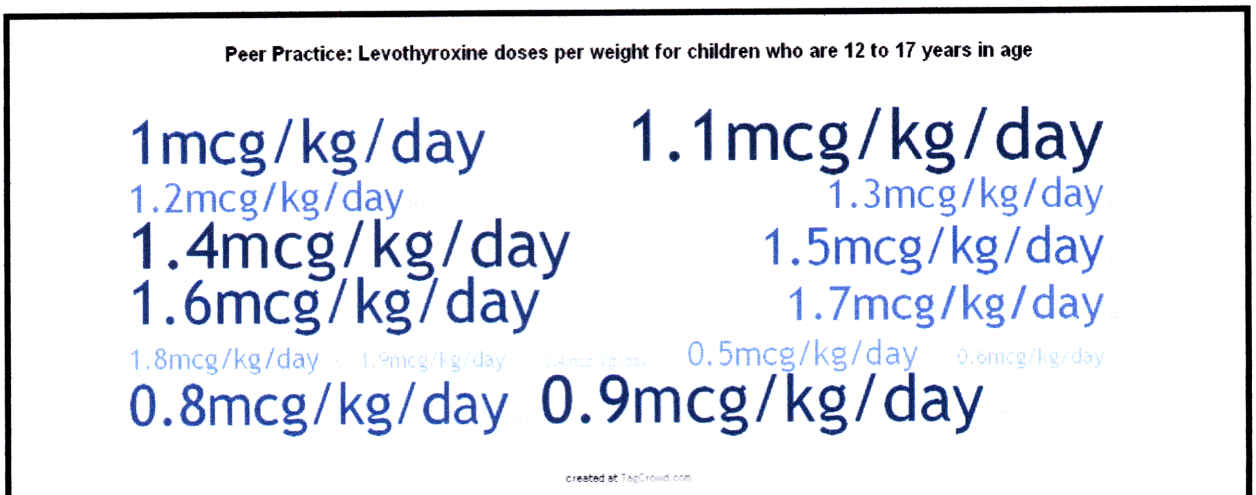

Figure 5. Visualization of the top 15 most frequent Levothyroxine doses per weight from the consolidated data for children who are 12 to 17 years in age using the TagCrowd application. Note that the most frequent dose is the largest and the darkest text.

If a more organized graphic is desired, Figure 6 demonstrates the same information but in the form of a horizontal bar chart. Either way, the most frequent dose recorded is below the dosage range of the recommendations. However, if the clinician user wanted to see the levothyroxine dosing practices for 12 year old patients, they can query for a "dosage cloud" for a cohort of 12 year olds, (see Figure 7 for the dosage cloud and Figure 8 for the bar chart). Note that the most frequent dose for a cohort of 12 year old is 1.4 mcg/kg/dose (6 records), which is a higher dose than the most frequent dose for the 12-17 year age group. Also note that the most frequent doses for a 15 year old is a three way tie of 1.1, 1.4 and 1.5 mcg/kg/dose (8 records each) (Figure 9 for the dosage cloud and Figure 10 for the bar chart), while the most frequent dose for 17 year old is a tie between 0.9 and 1.6 mcg/kg/dose (Figure 11 for the dosage cloud and Figure 12 for the bar

chart). The variability of most frequent dosing for each age demonstrates the utility of presenting peer practice information. While the established recommendations suggest dosages based on age ranges, the peer practice information from the CDW can present data according to the needs of the user.

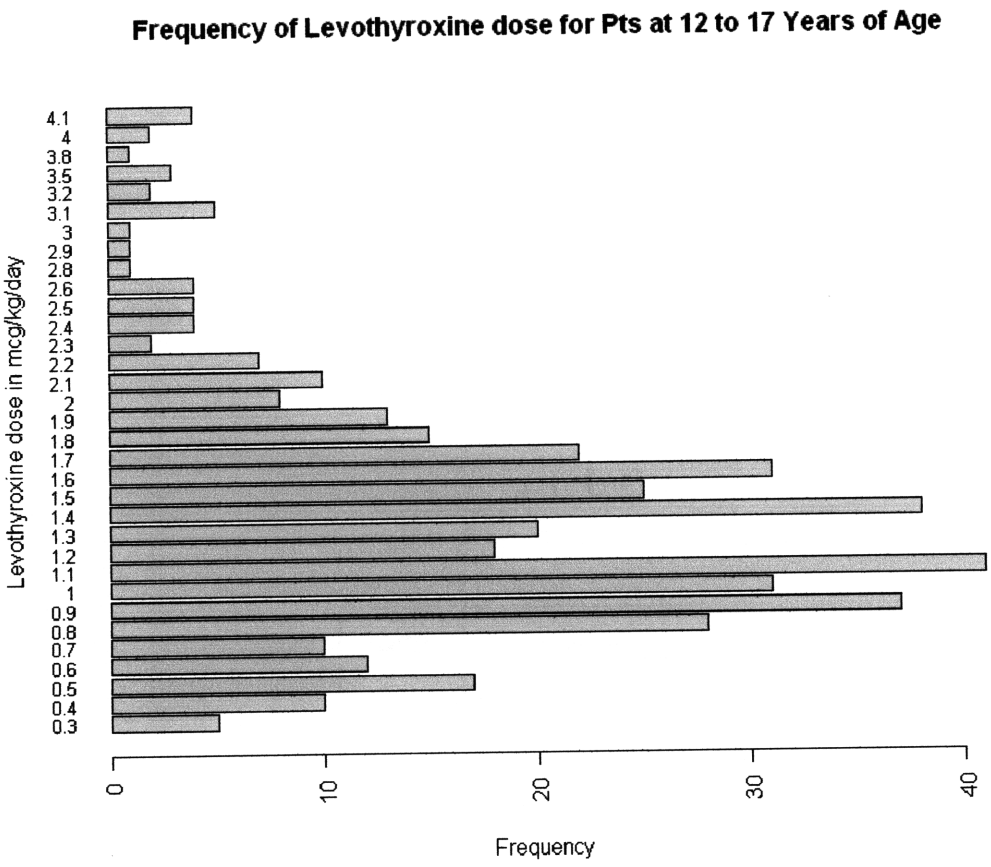

Figure 6. Visualization of all the Levothyroxine doses per weight from the consolidated data for children who are 12 to 17 years old using the barplot function in R. The bar plot is a more structured version of the information in Figure 5.

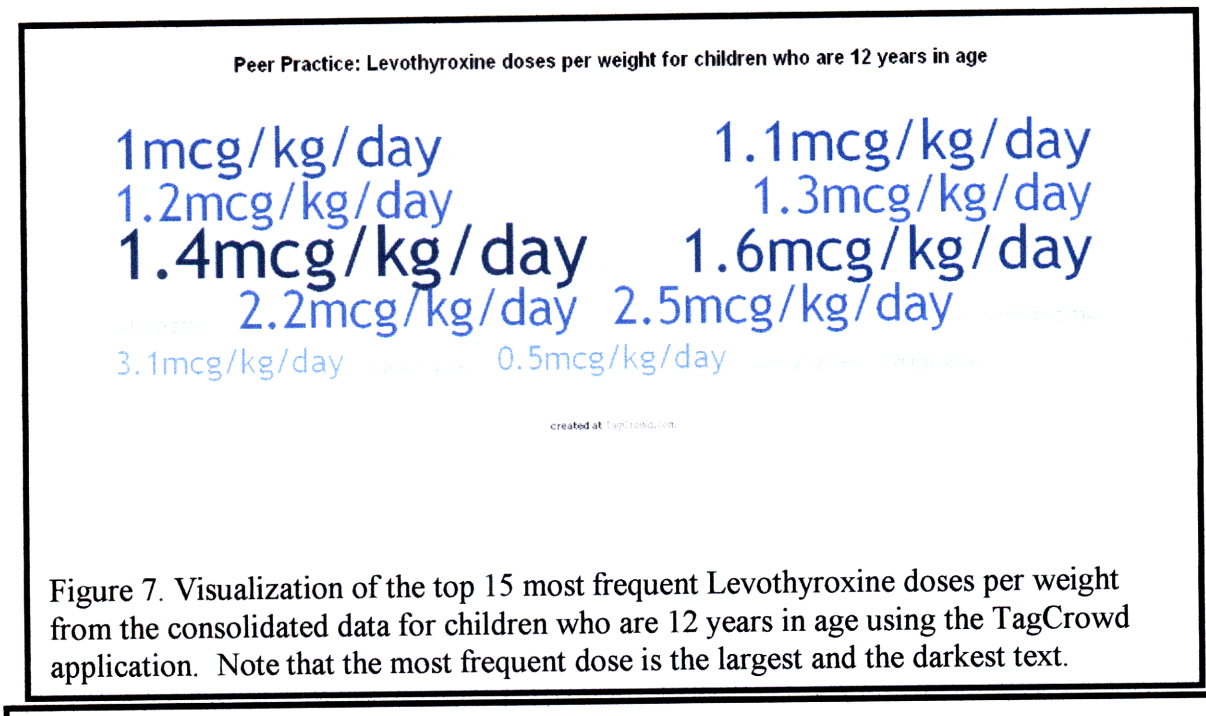

Figure 7. Visualization of the top 15 most frequent Levothyroxine doses per weight from the consolidated data for children who are 12 years in age using the TagCrowd application. Note that the most frequent dose is the largest and the darkest text.

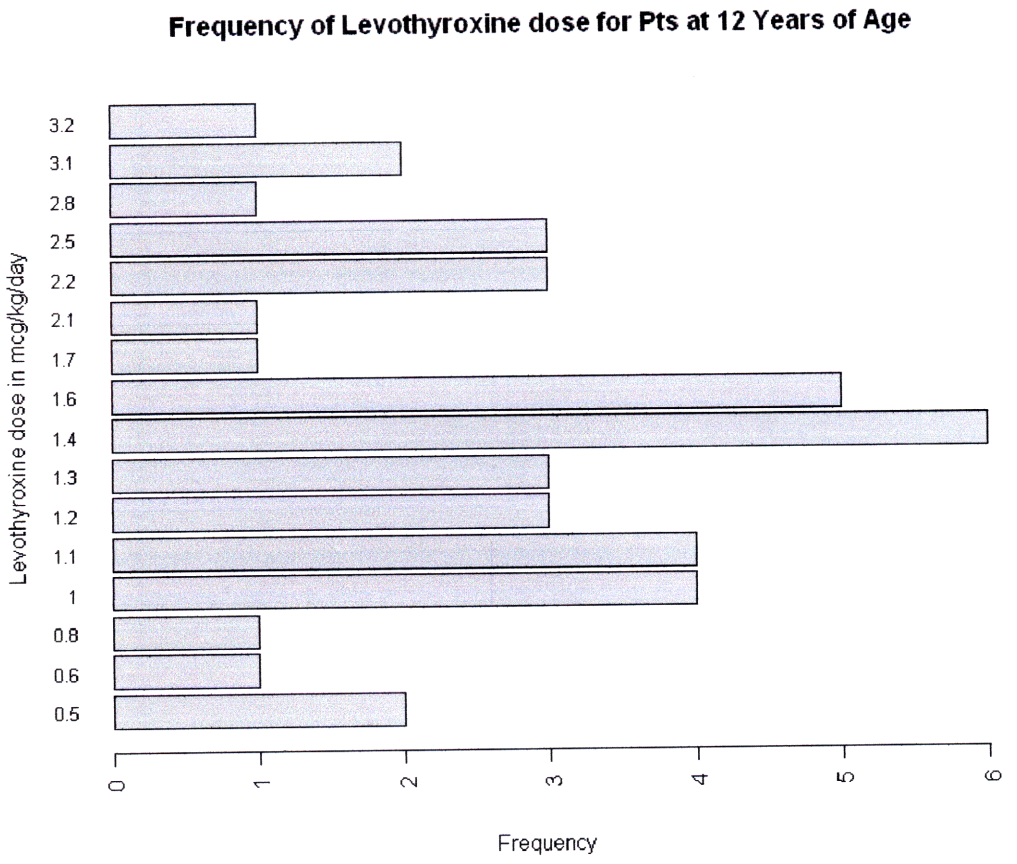

Figure 8. Visualization of all the Levothyroxine doses per weight from the consolidated data for children who are 12 years old using the barplot function in R. The bar plot is a more structured version of the information in Figure 7.

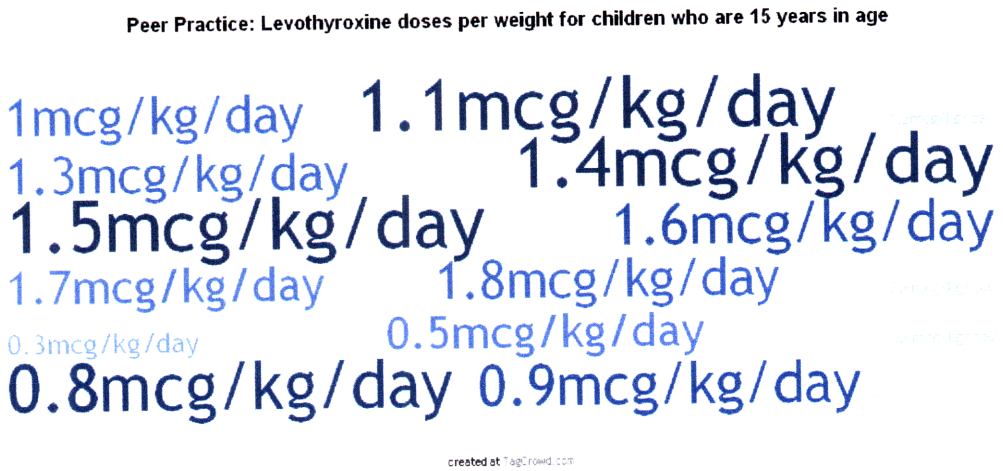

Figure 9. Visualization of the top 15 most frequent Levothyroxine doses per weight from the consolidated data for children who are 15 years in age using the TagCrowd application. Note that the most frequent dose is the largest and the darkest text.

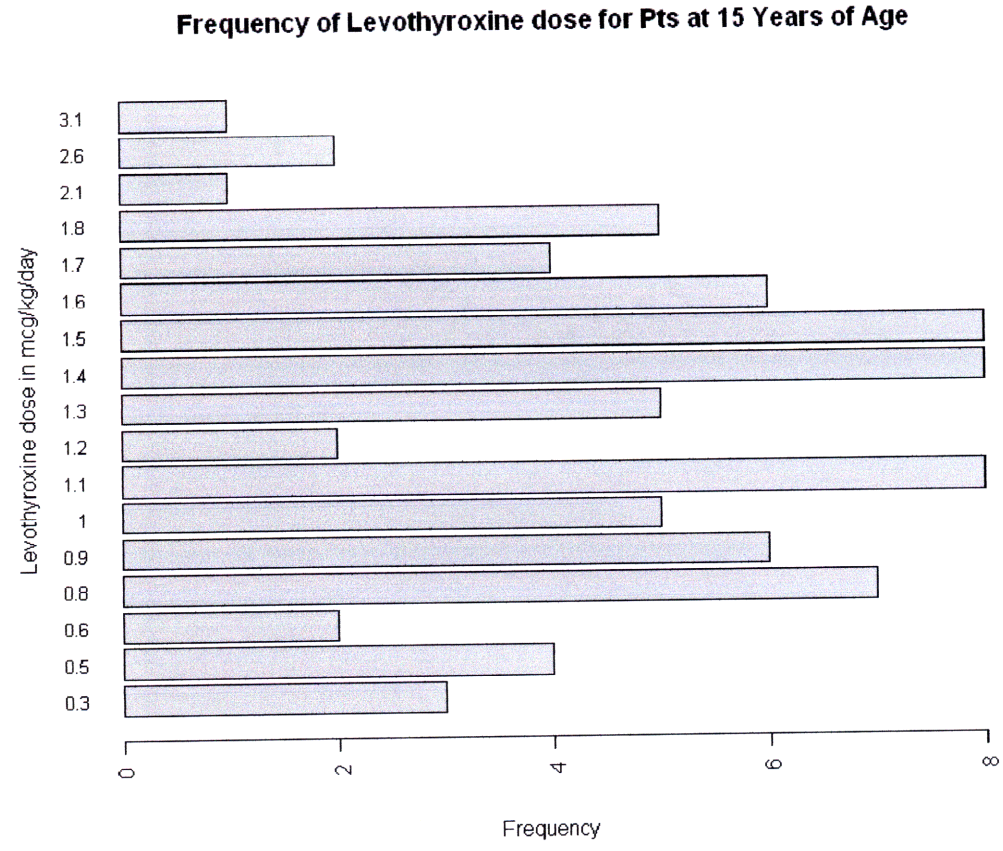

Figure 10. Visualization of all the Levothyroxine doses per weight from the consolidated data for children who are 15 years old using the barplot function in R. The bar plot is a more structured version of the information in Figure 9.

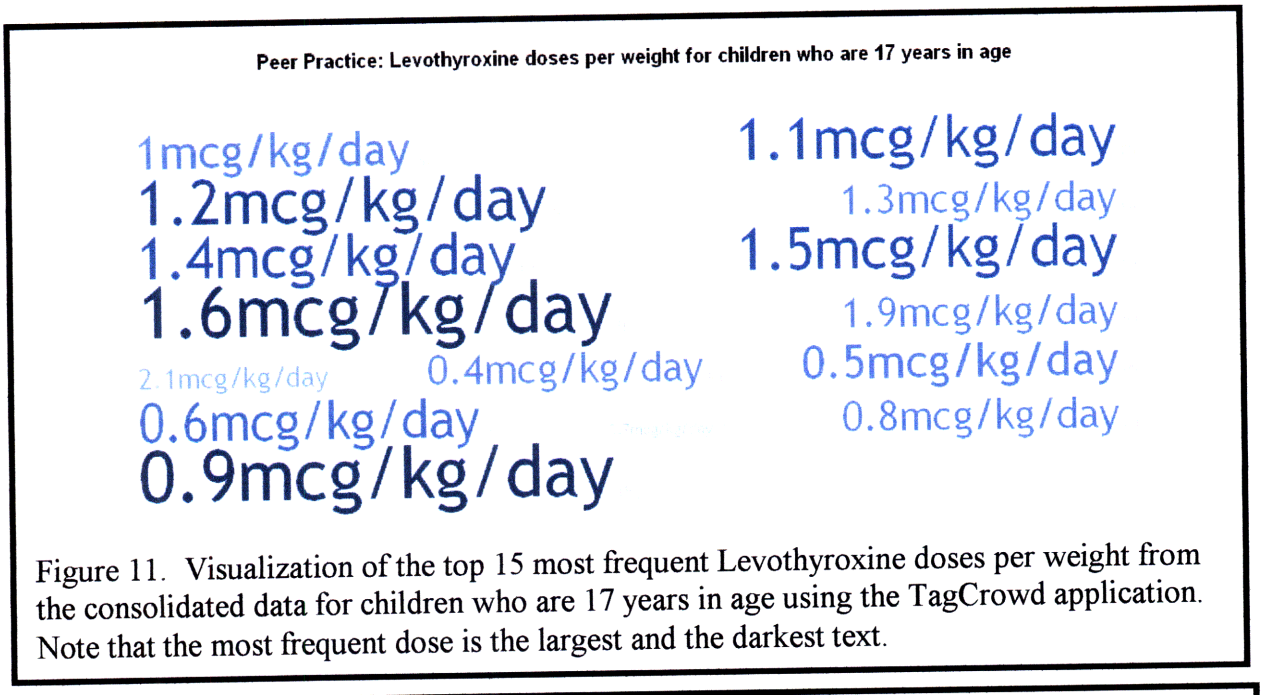

Figure 11. Visualization of the top 15 most frequent Levothyroxine doses per weight from the consolidated data for children who are 17 years in age using the TagCrowd application. Note that the most frequent dose is the largest and the darkest text.

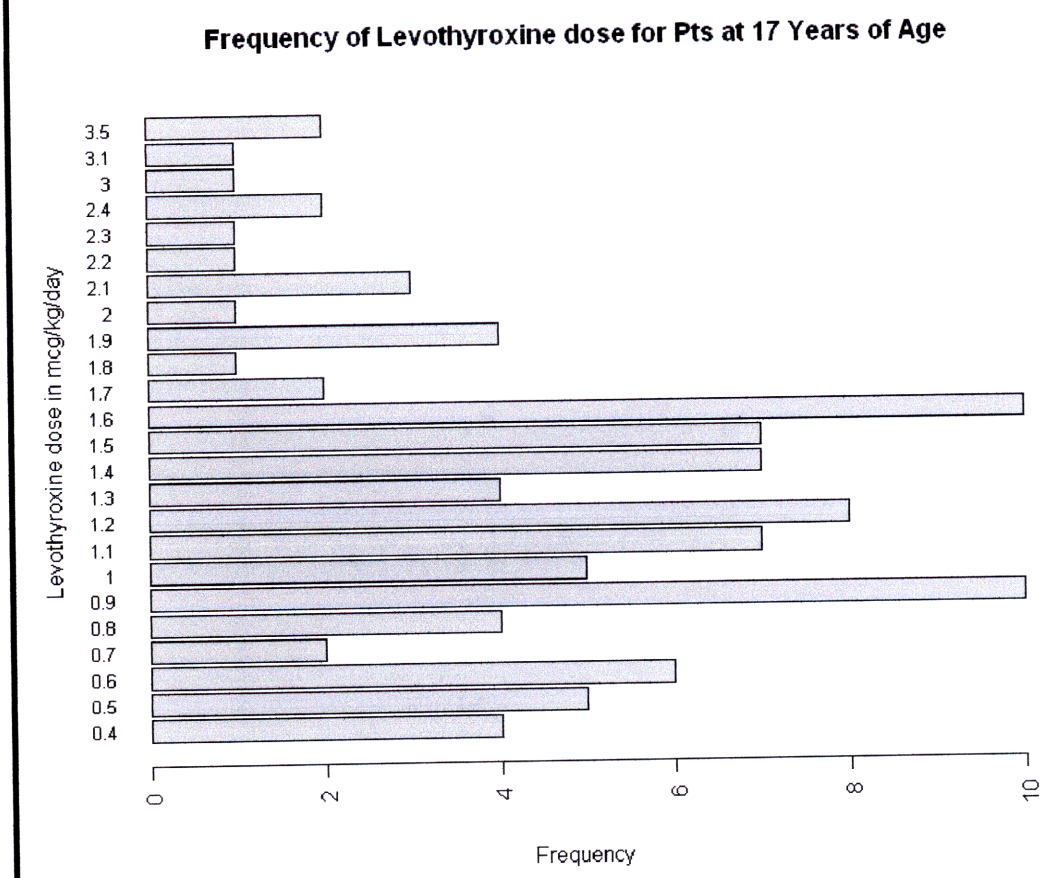

Figure 12. Visualization of all the Levothyroxine doses per weight from the consolidated data for children who are 17 years old using the barplot function in R. The box plot is a more structured version of the information in Figure 11.

Consolidating the demographic data

The database was queried for the demographic information of the patients in order to assess the gender and racial make up of the cohort (see Appendix - Query C). In terms of the gender, the field simply recorded "M" for male and "F" for female. There were not any "NULL" or "unknown" values for this attribute. In terms of race, there seems to be several different terms for certain races, which required consolidated. For the ambiguous terms "@", "UNK", "U", "UNKNOWN", "DECLINED", "OTHER", and "UNAVAILABLE", I grouped them as "UNKNOWN." For the terms "B", and "BLACK", I grouped them as "BLACK". For "W", "WHITE", "CAUCASIAN", I grouped them as "WHITE". For the remaining terms of "HISPANIC", "ASIAN" AND "AMER. INDIAN", I kept them as they were. In the end, I consolidated 15 separate terms into the 6 categories of "UNKNOWN", "WHITE", BLACK", "HISPANIC", "AMER.INDIAN" and "ASIAN."

The time interval between the weight and the medication event

In Figure 14, the distribution of patient weight vs. age from the consolidated data demonstrates that the weights of children when taken on the same day, within 10% of the age of the patient, or greater has a similar distribution. From 0 to 2000 days (about 5.5 years), the weights fall into a very tight space. Beyond 2000 days, the weight values fan out significantly, analogous to the weight distribution of pediatric growth charts.[17]

Age ranges	Same day	Within 10% of age*	More 10% of age	Total
0-3 months	28 (70.0%)	7 (17.5%)	5 (12.5%)	40
4-6 months	4 (26.7%)	4 (26.7%)	7 (46.7%)	15
7-12 months	6 (33.3%)	9 (50%)	3 (16.7%)	18
1-5 years	38 (27.5%)	62 (44.9%)	38 (27.5%)	138
6-11 years	65 (30.8%)	126 (59.7%)	20 (9.5%)	211
> 11 years	135 (31.3%)	277 (64.1%)	20 (4.6%)	432
Total	276 (32.3%)	485 (56.8%)	93 (10.9%)	854

Figure 13. The number of medication records where the difference between the date of the weight and medication record is zero (same day), within 10% of the patient's age or more than 10% of the patient's age (percentage of total for age group).

*not including the same day.

[17] http://www.cdc.gov/nchs/about/major/nhanes/growthcharts/clinical_charts.htm - Accessed on 01-28-2009

Of the whole group (see Figure 13 for the following data), 32.3% of the medication doses used weights from the same day, with almost 90% of all weight records with dates within 10% of the age of the patient. In particular, children in the 0-3 month range had the highest percentage of same day concordance (70.0%). Interestingly, the worst same day concordance was in the 4-6 month group (26.7%).

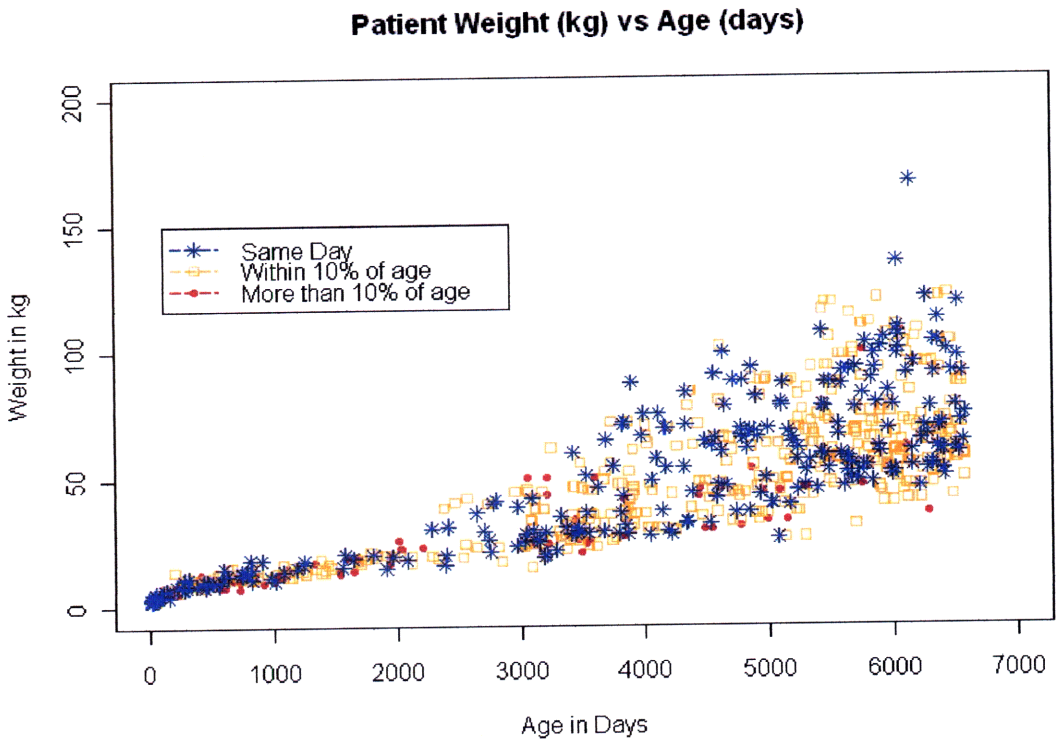

Figure 14. For Consolidated Data, the Patient Weight in kg vs. Age in days. Please see Figure 13 for numerical representation.

Phase II: Analysis of Data

The data was analyzed to evaluate several issues. Since transforming the data from the initial join is cumbersome, and may be difficult to automate, it is of interest to evaluate whether the transformation is worthwhile. The data from the initial join of the medication and the weight tables ("raw") and the final transformation ("consolidated") were compared. For each age group, the average dosages in each age group, the demographic distribution of the patients whose medication records are represented, and the rate of compliance to the medication recommendations were analyzed.

Of note, the definition of a cohort is an interesting issue for this study. While the records obtained were for particular patients, the medication records, themselves, were the cohort in this study. Only Figure 15 describes the patients as distinct individuals.

	Initial Join (Raw)	Cleaned data (Consolidated)
Total	352	348
Race		
White	250	248
Black	10	10
Hispanic	47	47
Asian	9	9
American Indian	2	2
Unknown	34	32
Gender		
Male	104	102
Female	248	246

Figure 15. Patient Demographics. The number of distinct patients in the initial join of the weight and medication tables (Raw), and subsequent transformed data (Consolidated) and the distribution of race and gender within each group.

For the remaining figures, the cohort is the medication records. For example, Figures 19 and 20 detail the gender distribution and summarize the number of males or females that are represented by all the medication records. Thus, it is possible that a patient is counted more than once if that patient has more than one viable medication record as part of the medication cohort. The same

patient can also span several age groups because the patient may have received many separate prescriptions as they were growing older.

The following is the summary of the analysis of the data by dosage, gender, race and dosing compliance for the whole group and the age ranges according to the PDH medication recommendations. The age ranges were defined as follows: 0-3 months: 0-90 days; 3-6 months: 91-181 days; 6-12 months: 182-364 days; 1-5 years: 365-2189 days; 6-11 years: 2190-4379 days; and older than 11 years: >= 4380 days. For the attribute of gender, the percentage of males and females represented by the medication records were calculated for each age range. Also for the attribute of race, the percentage of "White", "Black", "Hispanic", "Asian", "American Indian", and "unknown" were calculated for each age range. For the attribute of medication recommendation compliance, the percentage of medication doses that were "under", "in range' and "over" the dosing recommendations were calculated for each age range. The 95% confidence interval was calculated for each percentage based on the binomial distribution.

General

Consolidating the data resulted in a 21.5% decrease from 1088 records in the raw data group to 854 records in the consolidated group. Figures 16 and 17 illustrates the relationship of the dose per weight (mcg/kg/day) with the patient's age (in days) and the compliance with PDH recommendations for the raw and the consolidated data, respectively. Interestingly, the number of patients represented decreased from 352 to 348, only 4 fewer patients (see Figure 15). The patient cohort is represented by a majority of females and white patients (see Figure 15). The average number of prescription per patient dropped from 3.1 to 2.45. The actual number of prescriptions per patient ranged from 1 to 28 in the raw initial join, and 1 to 23 in the cleaned data group.

As per age group

For the following, the results will be reviewed in the age groups based on the levothyroxine dosing recommendations. Please refer to the following tables for the appropriate information and further details.

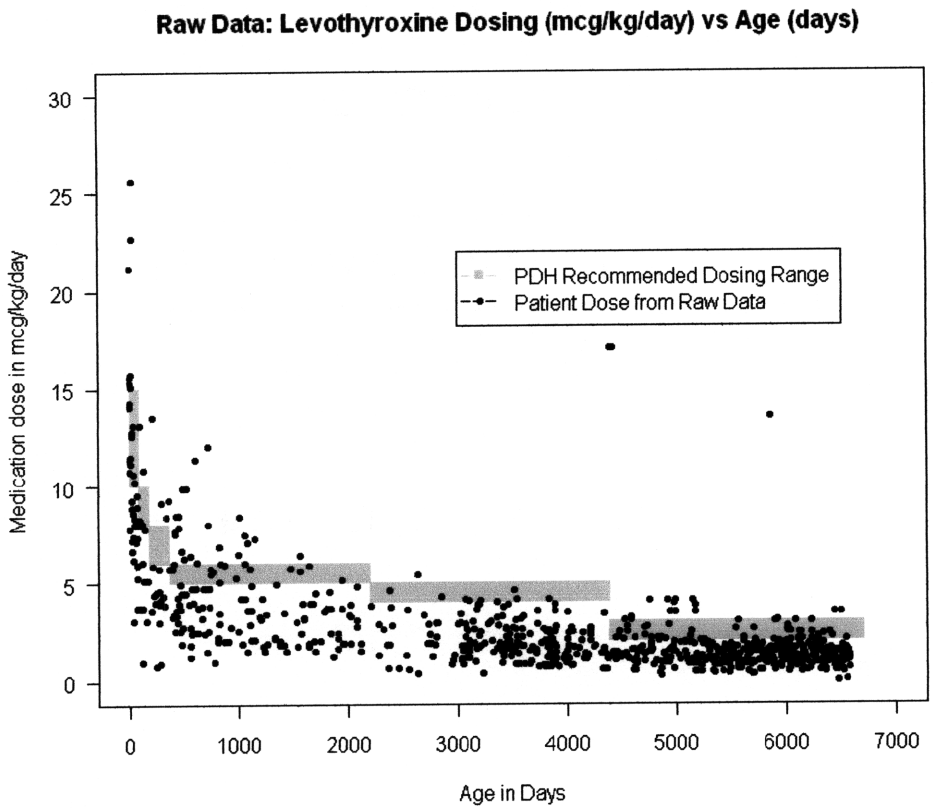

Figure 16. Raw data: Levothyroxine dosing in mcg/kg/day vs. Age in days. The gray bars indicate the recommended dosage ranges from the Pediatric Dosage Handbook.[14] Please see Figure 2 for PDH dosing recommendations and Figure 23 for categorical representation of the data compliance with the recommendations.

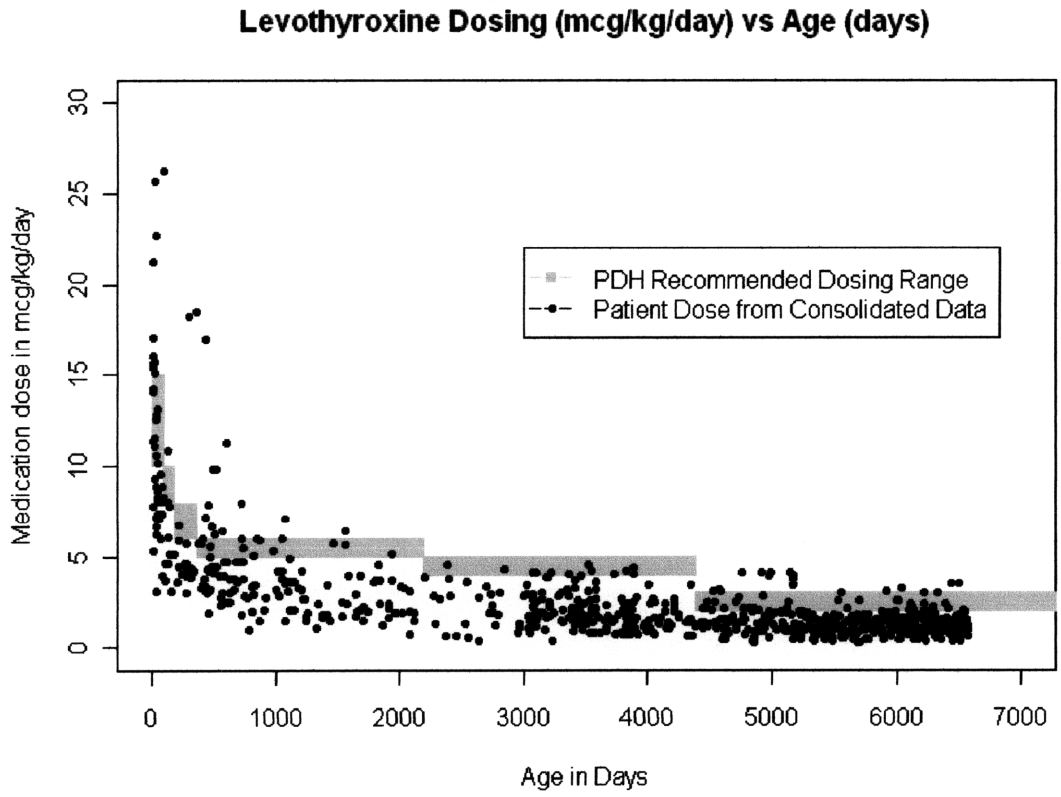

Figure 17. Consolidated data: Levothyroxine dosing in mcg/kg/day vs. Age in days. The gray bars represent the recommended dosage ranges from the Pediatric Dosage Handbook.[14] Please see Figure 2 PDH dosing recommendations and Figure 23 for categorical representation of the data compliance with the recommendations.

- For the average levothyroxine dosages for each age group, please see Figure 18 for the results of both datasets.

Ages	Raw	Consolidated
0-3 months	10.2 [8.94,11.4]	11.4 [9.80,12.9]
4-6 months	6.34 [4.87,7.82]	7.66 [4.83,10.5]
7-12months	4.86 [3.95,5.77]	6.07 [3.95, 8.18]
1-5 years	3.97 [3.66,4.29]	3.78 [3.42,4.14]
6-11 years	2.04 [1.92,2.16]	1.99 [1.85,2.12]
> 11 years	1.46 [1.36,1.57]	1.37 [1.31,1.44]
Total	2.64 [2.47,2.81]	2.59 [2.39,2.79]

Figure 18. The average levothyroxine dose per weight, for the raw data and the consolidated data. Mean [95% confidence interval, calculations based on t-test]

- For the gender distribution by age groups, please see Figure 19 for the consolidated data, and Figure 20 for the raw data.

Age ranges	Gender		Totals
	Female	Male	
0-3 months	17 (42.5) [27.0,59.1]	23 (57.5) [40.9,73.0]	40
4-6 months	3 (20.0) [4.33,48.1]	12 (80.0) [51.9,95.7]	15
7-12 months	9 (50.0) [26.0,74.0]	9 (50.0) [26.0,74.0]	18
1-5 years	76 (55.1) [46.4,64.5]	62 (44.9) [36.5,53.6]	138
6-11 years	150 (71.1) [64.5,77.1]	61 (28.9) [22.9,35.5]	211
> 11 years	323 (74.8) [70.4,78.8]	109 (25.2) [21.2,29.6]	432
Total	578 (67.7) [64.4,70.8]	276 (32.3) [29.2,35.6]	854

Figure 19. Distribution of gender among the medication records for the consolidated data. Number of prescriptions written for patients of the gender in the age group (percent of age group) [95% confidence interval, calculations based on the binomial distribution]

Age ranges	Gender		Age Totals
	Female	Male	
0-3 months	29 (50.0) [36.6,63.4]	29 (50.0) [36.6,63.4]	58
4-6 months	4 (23.5) [6.81,49.9]	13 (76.5) [50.1,93.2]	17
7-12 months	16 (55.2) [35.7,73.6]	13 (44.8) [26.4,64.3]	29
1-5 years	95 (53.4) [45.8,60.9]	83 (46.6) [39.1,54.2]	178
6-11 years	182 (71.6) [65.7,77.1]	72 (28.3) [22.9,34.3]	254
> 11 years	423 (76.6) [72.9,80.1]	129 (23.3) [19.9,27.1]	552
Total	749 (68.8) [66.0, 71.6]	339 (31.2) [28.4,34.0]	1088

Figure 20. Distribution of gender among the medication records for the raw data. Number of prescriptions written for patients of the gender in the age group (percent of age group) [95% confidence interval, calculations based on the binomial distribution]

- For the race distribution by age groups, please see Figure 21 for the consolidated data, and Figure 22 for the raw data.

Age ranges	Race						
	White	Black	Hispanic	Asian	Am. Ind.	Unknown	Total
0-3 months	22 (55.0) [38.5,70.7]	0 [0,8.81]	16 (40.0) [24.9,56.7]	0 [0,8.81]	0 [0,8.81]	2 (5.00) [0.61,16.9]	40
4-6 months	7 (46.7) [21.3, 73.4]	0 [0,21.8]	7 (46.7) [21.3,73.4]	1 (6.67) [0.17,31.9]	0 [0,21.8]	0 [0,21.8]	15
7-12 months	11 (61.1) [35.7,82.7]	0 [0, 18.5]	6 (33.3) [13.3,59.0]	0 [0, 18.5]	0 [0, 18.5]	1 (5.56) [0.14, 27.3]	18
1-5 years	79 (57.2) [48.5,65.6]	20 (14.5) [9.08,21.5]	32 (23.2) [16.4, 31.1]	4 (2.90) [0.80, 7.26]	0 [0,2.64]	3 (2.12) [0.45, 6.22]	138
6-11 years	147 (69.7) [63.0,75.8]	4 (1.90) [0.52, 4.78]	38 (18.0) [13.1,23.9]	7 (3.32) [1.34,6.72]	0 [0, 1.73]	15 (7.11) [4.03,11.5]	211
> 11 years	312 (72.2) [67.7,76.4]	5 (1.16) [0.38,2.68]	50 (11.6) [8.71,15.0]	11 (2.54) [1.28,4.51]	7 (1.62) [0.65,3.31]	47 (10.9) [8.10,14.2]	432
Total	578 (67.7) [64.4,70.8]	29 (3.40) [2.29,4.84]	149 (17.4) [15.0,20.2]	23 (2.69) [1.71,4.01]	7 (0.82) [0.33,1.68]	68 (7.9) [6.24,9.99]	854

Figure 21. Distribution of race among the medication records of the consolidated data. Number of prescriptions written for patients of a particular race in the age group (percent of age group) [95% confidence interval, calculations based on the binomial distribution]

Age ranges	Race						
	White	Black	Hispanic	Asian	Am. Ind.	Unknown	Total
0-3 months	34 (58.6) [44.9,71.4]	0 [0, 6.16]	22 (37.9) [25.5,51.6]	0 [0,6.16]	0 [0,6.16]	2 (3.4) [0.42,11.9]	58
4-6 months	7 (41.2) [18.4,67.1]	0 [0,19.5]	8 (47.1) [23.0,72.2]	1 (5.88) [0.15,28.7]	0 [0,19.5]	1 (5.88) [0.15,28.7]	17
7-12 months	13 (44.8) [26.4,64.3]	0 [0,11.9]	12 (41.4) [23.5,61.1]	0 [0,11.9]	0 [0,11.9]	4 (13.8) [3.89,31.7]	29
1-5 years	100 (56.2) [48.6,63.6]	28 (15.7) [10.7,21.9]	36 (20.2) [14.6,26.9]	6 (3.37) [1.25,7.19]	0 [0,2.05]	8 (4.49) [1.96,8.66]	178
6-11 years	174 (68.5) [62.4,74.2]	5 (1.97) [0.64,4.53]	42 (16.5) [12.2,21.7]	11 (4.33) [2.18,7.62]	0 [0,1.44]	22 (8.66) [5.51,12.8]	254
> 11 years	409 (74.1) [70.2,77.7]	5 (0.91) [0.29,2.10]	61 (11.1) [8.56,14.0]	12 (2.17) [1.13,3.77]	7 (1.27) [0.51,2.60]	58 (10.5) [8.08,13.4]	552
Total	737 (67.7) [64.9,70.5]	38 (3.49) [2.48,4.76]	181 (16.6) [14.5,19.0]	30 (2.76) [1.87,3.91]	7 (0.64) [0.26,1.32]	95 (8.73) [7.12,10.6]	1088

Figure 22. Distribution of race among the medication records of the raw data. Number of prescriptions written for patients of a particular race in the age group (percent of age group) [95% confidence interval, calculations based on the binomial distribution]

- For the distribution of appropriate dosing when compared with the recommendations by age groups, please see Figure 23 for the results for both datasets.

Ages	Raw data				Consolidated data			
	Under	In range	Over	Total	Under	In range	Over	Total
0-3 months	32 (55.2) [41.5,68.2]	15 (25.9) [15.3,39.0]	11 (19.0) [9.87,31.4]	58	19 (47.5) [31.5,63.9]	11 (27.5) [14.6,43.9]	10 (25.0) [12.7,41.2]	40
4-6 months	10 (58.8) [32.9,81.6]	5 (29.4) [10.3,56.0]	2 (11.8) [1.46,36.4]	17	9 (60.0) [32.3, 83.7]	4 (26.7) [7.79,55.1]	2 (13.3) [1.66,40.5]	15
7-12 months	25 (86.2) [68.3,96.1]	0 [0,11.9]	4 (13.8) [3.89,31.7]	29	15 (83.3) [58.6,96.4]	1 (5.56) [0.14,27.3]	2 (11.1) [1.38,34.7]	18
1-5 years	131 (73.6) [66.5,79.9]	18 (10.1) [6.10,15.5]	29 (16.3) [11.2,22.6]	178	110 (79.7) [72.0,86.1]	14 (10.1) [5.66,16.4]	14 (10.1) [5.66,16.4]	138
6-11 years	243 (95.7) [92.4,97.8]	9 (3.5) [1.63,6.62]	2 (0.8) [0.10,2.82]	254	199 (94.3) [90.3,97.0]	12 (5.7) [2.97,9.72]	0 [0,1.73]	211
> 11 years	472 (85.5) [82.3,88.3]	53 (9.6) [7.27,12.4]	27 (4.9) [3.25,7.04]	552	376 (87.0) [83.5,90.1]	38 (8.8) [6.30,11.9]	18 (4.2) [2.49,6.50]	432
Total	913 (83.9) [81.6,86.1]	100 (9.19) [7.54,11.1]	75 (6.89) [5.46,8.56]	1088	728 (85.2) [82.7,87.6]	80 (9.37) [7.50,11.5]	46 (5.39) [3.97, 7.12]	854

Figure 23. Dosage Compliance. The data from the raw and the consolidated data categorized by whether the dose is under, over or within the range of the medication recommendations from the Pediatric Dosage Handbook for levothyroxine for each of the specified age groups. First value is the number of prescriptions, followed by (percentage of age group) and the [95% confidence interval, calculations based on the binomial distribution].

Overall

The consolidated data group (CD) had 854 records, while the raw data (RD) group had 1088. The average dosing was 2.59 mcg/kg/day [95% confidence interval: 2.39:2.79] for the consolidated group and 2.64 mcg/kg/day [2.47:2.81] for the raw group. There was a majority of medications written for females (CD 67.7%, [64.4:70.8]; RD 68.8% [66.0:71.6]) and for white patients (CD 67.7% [64.4:70.8]; RD 67.7 [64.9:70.5]). In terms of recommendation compliance, the overwhelming majority was underdosed (CD 85.2 [82.7:87.6]; RD 83.9% [81.6:86.1]).

0-3 months

The CD had 40 records, while the RD had 58. The average dose for the CD was 11.4 mcg/kg/day [9.80:12.9] and for the RD was 10.2 mcg/kg/day [8.94:11.4]. In the consolidated group, the number of prescriptions written for males was higher (57.5%, [40.9:73.0]), whereas in the raw data group, the number for males and females were the same (50%, [36.6:63.4]). Regarding race, in both groups, the highest percentage of prescriptions was written for white patients (CD 55.0%, [38.5: 70.7]; RD 58.6%, [44.9, 71.4]). In terms of following recommendations, in both groups, the highest percentage of the prescriptions fell into the "underdosing" category (CD 47.5%, [31.5: 63.9]; RD 55.2%, [41.5:68.2]).

4-6 months

The CD group had 15 records, while the RD had 17. The average medication dose for the CD was 7.66 mcg/kg/day [4.83: 10:5], and for the RD was 6.34 mcg/kg/day [4.87:7.84]. In both groups, there were a large percentage of males represented, (CD 80.0% [51.9:95.7], RD 76.5% [50.1:93.2]). In the CD group, the highest percentage of records was presented equally by white and Hispanic patients (46.7%, [21.3, 73.4]), while in the RD group, the records for Hispanic patients (47.1%, [23.0, 72.2]) edged out the white patients (41.2%, [18.4, 67.1]). When it came to following recommendations, both groups had a majority of the records in the "underdosing" category (CD 60.0%, [32.3:83.7]; RD 58.8%, [32.9:81.6]).

7-12 months

The CD group had 18 records, while the RD group had 29. The average dose for the CD was 6.07 mcg/kg/day [3.95:8.18] and for the RD was 4.86 mcg/kg/day [3.95:5.77]. In the CD group, males and females were represented equally (50% [26.0:74.0]), whereas in the RD group there were slightly more females (55.2%, [35.7:73.6]). In both groups, the highest percentage of records was presented by white patients (CD 61.1%, [35.7, 82.7]; RD 44.8%, [26.4, 64.3]). In terms of recommendations, in both groups a large majority of the prescriptions fell into the underdosing category (CD 83.3%, [58.6:96.4]; RD 86.2%, [68.3:96.1]).

1-5 years

The CD group had 138 records while the RD group had 178. The average medication dose for CD was 3.78 mcg/kg/day [3.24:4.14], and for the RD was 3.97 mcg/kg/day [3.66:4.29]. In both groups more records were represented by females (CD 55.1%, [46.4:64.5]; RD 53.4%, [45.8,60.9]). Also in both group, records for white patients made up the majority (CD 57.2%, [48.5,65.6]; RD 56.2%, [48.6, 63.6]). This age range is also the first time there were any records for black patients (CD 14.5%, [9.08, 21.5]; RD 15.7%, [10.7:21.9]). In terms of recommendations, in both groups, a large majority of the records fell in the underdosing category (CD 79.7%, [72.0: 86.1]; RD 73.6%, [66.5: 79.9]).

6-11 years

The CD group had 211 records, while the RD group had 254. The average dosing for the CD was 1.99 mcg/kg/day [1.85:2.12] and for the RD was 2.04 mcg/kg/day [1.92:2.16]. In both groups, there were many more records for females (CD 71.1%, [64.5,77.1]; RD 71.6%, [65.7, 77.1]) than for males. Also in both groups, records for white patients had a much higher representation (CD 69.7%, [63.0: 75.8]; RD 68.5%, [62.4:74.2]). In terms of recommendations, in both groups, an overwhelming percentage were in the underdosing category (CD 94.3%, [90.3:97.0]; RD 95.7%, [92.4:97.8]).

> 11 years

The CD group had 432 records, while the RD group had 552. The average levothyroxine dose for the CD was 1.37 mcg/kg/day [1.31:1.44], and for the RD was 1.46 mcg/kg/day [1.36:1.57]. In both groups, records for females (CD 74.8%, [70.4:78.8]; RD 76.6%, [72.9,80.1]) had a much higher percentage. In both groups, records for white patients make up the majority (CD 72.2%, [67.7: 76.4]; RD 74.1%. [70.2:77.7]). In terms of recommendations, a very large percent of the records are in the underdosing category for both groups (CD 87.0%, [83.5:90.1]; RD 85.5%, [82.3:88.3]).

DISCUSSION

In this paper, I studied the potential of providing peer practice information from a clinical data warehouse as a form of clinical decision support for pediatric medication dosing practices. To start, I posed the question, "what is the weight based dosing practices of levothyroxine in the pediatric specialty?" First, the data warehouse was queried to obtain the necessary data. In particular, I collected a pediatric cohort with records for weights and medication prescriptions, calculating the patients' ages in days and in years at the time of the event. The data was transformed in order to present what the clinician intended using only the data retrieved from the database. Also, I demonstrated a few methods by which the peer practice data can be presented. Second, the initially joined, raw data was compared with the consolidated data to evaluate the utility of transforming the original data. The data was also compared with recommendations for pediatric levothyroxine dosing because findings of significant similarity or difference would be useful to the clinician. In the following, I will discuss the lessons learned and my recommendations for the major steps in developing a peer practice clinical decision support system from a clinical data warehouse for pediatric medication dosing.

Data Acquisition and Transformation

Data structure of the medication records

The structure of the facts in this particular data warehouse is somewhat limiting as it relates to the components of a medication prescription. The general details of a medication prescription include name, dose, route, frequency, duration and directions or comments. The only medication components available in a structured form in the RPDR were the name, dose value and the units. Unfortunately, this limits the granularity of information that can be queried and manipulated regarding medications prescribed to patients. However, along with each medication record was a field for HL7 text, which included all the details as listed above for medication prescriptions in XML syntax. HL7 is a standard in electronic medical information exchanged, used by 90% of U.S. medical institutions.[18] This can serve as an alternate source of medication information if the database tables do not store medications details in full.

The age attribute

There were some unexpected issues related to the attribute of age while querying this data warehouse relating to how age is represented, the limitations of the framework, and errors in the data.

First, the age recorded in the demographics table was age as the patient would be at the time of the last upload, instead of age at the time of the event or record. Since the upload occurs on a monthly basis, it stores what basically amounts to the current age. Unfortunately, this does not represent age in the manner appropriate for this study, where age at the time of the event, such as when the prescription was written or when the weight was recorded, is ideal. Two options by which this issue can be addressed are either to have the database store age or date of birth with each patient related record, or to join tables during the query containing the patient's date of birth and the event's date and then calculate the age at the time of the event.

Second, the framework lacked a method through which to calculate age in years. While MS Sql Server has a function to perform the date difference, "DateDiff()", it presents the difference between the numerical values for years. For instance, if performed on the dates of December 31, 2007 and January 1, 2008, the function would return the result of 1 year, even

[18] http://www.itl.nist.gov/div897/docs/Message_Maker.html - Accessed on 01-28-2009

though only a day had passed, and the age in years should be zero. Instead, I incorporated Sql query code to accomplish this (see Appendix - Query A). However, age in days functioned as expected using the DateDiff() function.

Also, there may have been some erroneous data in the database. While it may be obvious to indicate an upper bound for age when querying the database for pediatric data, indicating the lower bound of 0 may not be as obvious. Initially, when only the upper bound was indicated, several records with negative values for ages were retrieved. It is difficult to determine the cause for this error. Some possible reasons include that the patient ids may not be unique or may be recycled as patients expire, that there are errors in the recorded dates of birth or of events or that there are other unexpected inconsistencies in the data. Basically, by specifying upper and lower bounds for age, records with negative ages can be excluded.

In general, age at the time of the event is most useful when studying patient cohorts from querying databases for medical information. Several possible solutions are: 1) include age in an accepted granular form, of years, months, or days with each record, 2) incorporate a join of the demographic table with the fact table at the time of the query in order to facilitate calculation of age, or 3) include date of birth with each patient fact. Option one is useful because it leaves the burden of calculation at the time of the upload. However, this option adds more attributes to each record. If the user is interested in a time unit that is not represented, then the user would have to incorporate additional calculations during the query, which may significantly increase the query burden. Option two is useful, because no significant changes are necessary in the data warehouse, however, the act of joining during the time of the query can have significant effects on the run-time burden of the query. Option three may be optimal because only one attribute is added to the data warehouse, and while calculations will be necessary during the query, the lack of a joining step will significantly decrease the query burden. In all cases, if the age of the patient in years is to be calculated, especially in the MS Sql Server environment which has no function to calculate age in years, it will be necessary to include sql code to accomplish this task (see Appendix - Query A for the sql code used for this study).

Query size

The query I had initially attempted to run was large, complicated and took significant time to complete. The query attempted to obtain information about medication as dose per weight and to calculate the age of the patient in days and years. In particular, I was trying to join each medication record with the most recent weight whose date was the same as or prior to the date of the medication.

The primary recommendation is to retrieve the data in a more manageable way. Consider querying separate tables for weights and medications. Perform the transformation in a separate application, such as MS Access, using smaller, more manageable queries. This worked much faster in this study, and should be considered when developing the query application, especially if it is to function quickly or at the point of care.

Issues when combining the weight and medication table

The primary issue when the weight and the medication records tables were joined was the redundancy of some weight records for a particular medication record because more than one weight was recorded on a particular date. Overall, this data manipulation is programmatically possible through the following steps. For the simple case of multiple weights with the same value, randomly remove one of the records. Since the weight encounter information for the removed record is not significant to the clinical question, this is a valid solution. However, there may be an overarching issue of determining whether the value of the weight is appropriate for the age of the patient. This issue may be a bit more complicated. One method, not pursued in this study, is to calculate two functions based on the information in pediatric growth curves.[17] One function will represent the lower bound using lowest percentile of interest (1^{st} , 3^{rd} or 5th), while the other function will represent the upper bound with the highest percentile of interest (95th, 97^{th} or 99^{th}). Using these functions, the range of weights can be calculated for a particular age, and then the patient's weight can be evaluated against the range. If only one weight fits the range, select that one. If both weights are plausible, selected the higher one since it is most likely a result of reweighing and children usually gain weight with time. If neither weight fits, then either remove both, or consider performing a regression of all the patient's weight values and determine whether or not the particular value for the weight falls along the regression line. It

is possible that a patient can have their own growth curve that goes above the 99th or below the 1st percentile.

Interval between dates of weight and medication records

When calculating the dose per weight for a pediatric patient, the most recent weight is used in pediatric practice. A weight on the same day as the medication prescription would be optimal, but a weight within a reasonable time frame may be acceptable. For the most part, the younger the child, the closer to date of the weight should be to the date of the medication, however, within 10% of the age of the patient may be acceptable for the pediatric population. For a 5 day old, same day is optimal. For a 6 month old (about 181 days), 18 days is acceptable. For a 15 year old (5475 days), over a year may not be as acceptable, but usually these kids are probably already close to adult weight. When children are in the weight range of adults, the max dosing comes into effect, and usually these children receive adult medication doses. In these cases, doses calculated may go significantly over the adult doses, so instead adult dosing practices come in to play. While weights taken on the same date as the medication prescription is ideal, it appears from this data that the user can retain approximately 90% of the consolidated data if they are willing to request that the weights to be used be about 10% or less than the age of the patient in days. .

Consolidating the data

Consolidating the data in order to minimize redundancy and ambiguity included determining the purpose of the information for the user, manipulating the data to reflect that purpose, addressing values that appeared to be outliers, and handling redundant records.

First, it was necessary to figure out the purpose of the data to be presented to the user. One method would be to compare medication prescriptions in their entirety. Instead of simply presenting only the dose, the full prescription may look like "levothyroxine 75 mcg PO QD". The user may find it interesting to know that while levothyroxine is recommended as a single daily dose, there were many instances where the medication was dosed twice a day or once every day, or different doses on alternating days. However, I wanted to compare the results with the recommendations. In order to accomplish that, the data needed to be consolidated to reflect daily

dosing because the recommendations call for once a day dosing. This meant that for medications with frequencies other than daily, doses were doubled for twice a day, halved for every other day, and averaged if different amounts were given on alternating days. Also, it was necessary to consolidate the units. Levothyroxine is usually dosed in micrograms as opposed to milligrams. However, there were some instances where milligrams were used, so that was addressed by multiplying the structured dose of MG by a 1000. If the medication details stored the information in a structured manner, it should not be difficult to perform this programmatically.

Another, more difficult, issue to contend with were prescription clarifications that modify the information of the intended dosing as represented in the structured fields. This is somewhat expected in fields such as "directions", where free text was standard. However, for the fields of "units" and "frequency", structured data was expected, but free text was allowed. In some cases, the free text in these fields was not consistent with what was written in the actual direction detail or with the prescription as a whole. This would be difficult to address programmatically. The primary solutions are either to limit the free text option during the data entry stage, offer more options for structured input at the data entry stage, or develop natural language processing techniques for the data manipulation stage. Unfortunately none of these options are simple to execute.

Outliers are also somewhat difficult to address programmatically. The medication doses and the weight values were graphed against the patient's age in days (see Figures 16 and17). Those that looked visually out of the general cohort were removed. Although not pursued in this study, one method of addressing this would be to perform a linear regression of the data, and then look for values that are significantly removed from the line, and systematically remove these points.

Redundant data were addressed via two means. The redundant weight data was addressed in a manual transformation. The redundant daily dose values were addressed in an automated format, by taking advantage of a functionality in MS Access. Redundancies for exactly the same values, for either weight or dosage, can be automated. However, the subtle differences of multiple similar but slightly different values for weight or dosages will be difficult to address. A solution for addressing the redundant weights is detailed above. However, for medications, it would be difficult to determine the intended prescription if multiple medications

are prescribed on the same date, and no further clarifications are present in any of the records. The primary solution would be to encourage improved data quality at the order entry level.

Peer practice data representation

There are several possibilities for presenting the data obtained in this study. As demonstrated in this paper, the data can be presented as frequencies of dose per weight per day as a "dosage cloud" or a horizontal bar chart. The data can also be represented as an average dose per weight per day of the cohort in question (see Figure 18 for mean doses). Examples of other options not pursued in this paper would be to take the patient demographic attributes, such as race and gender, and isolate a more specific cohort. For instance, medication dosing patterns can be queried for a cohort of female patients who are 7 to 9 years old. Race may be an interesting attribute to query by in certain circumstances. However, in this study, there is a very high percentage of white patients. Thus, grouping by any other race may result in a very small cohort. However, if data from many CDWs are collected and then queried, grouping by race will be a viable option. Also, another option for data representation would be to link an outcome measure to the medication records, to determine doses associated with optimal disease control. Effective real world prescription practices can be presented to the user through the use of more complex modeling of data from the CDW.

Analysis of the data

For the group as a whole, the cohort seemed to appropriately represent the population diagnosed with hypothyroidism. Most of the patients are white and female. Once the medication records were accounted for in each age group, most of the prescriptions were written for white patients and female patients. Only in the age group of 4-6 months for both the raw and the consolidated group, and the consolidated data's age group of 0 to 3 months were there more males than females represented. Also, it seems that the younger age groups were represented by less records than the older age groups. In reality, this is most likely reflective of the fact that the younger age groups represent shorter age ranges. Three of the age groups encompass age ranges of several months, while the other three encompass age ranges of several years, hence the higher number of records represented in the larger age ranges.

Regarding the concordance of levothyroxine dosing of the retrieved data with the recommendations, in every age range for both groups, the highest percentage of prescriptions fell into the "underdosing" category. The consistency of this finding across the age groups is very interesting because the data would indicate that clinicians in real world practices prescribe at a much lower dose per weight than what the recommendations specify. There may be some concern regarding the data because some of the weight values used for the patients are many months earlier than the date of the prescription. However, since children usually gain weight with time, the value of the overall weight based dose may be higher than expected, with the dose per weight being lower with a more recent value for weight. It would be interesting to pursue a clinical study of the real world medication dosing practices that maintain the euthyroid state. If the findings in that study are consistent with what is inferred from this study, the current recommendations may need to be reconsidered.

Overall, the raw data and the consolidated data were very similar in the distribution of race, gender and dosing compliance. Also, the two groups were very similar in the calculated average medication dose. The significance of these findings is that the transformation of the data from the initial join may not be necessary if the outcomes are categorical measurements (of compliance with recommendations) or the average values (of medication dosing). The similarity may be a byproduct of the inherent fact that the raw data and the consolidated data have many records that overlap. However, the analysis shows that two groups are still very similar, which calls into question the need for complicated data transformation to obtain 'cleaner' data, when the original data is adequate in conveying outcomes of categorical measurement or average values.

LIMITATIONS

The RPDR data warehouse, while a large, well-known and well-used repository of patient clinical information, has some shortcomings. The data are only as recent as the last upload. However, the data are uploaded at least every month, and some types of data are uploaded every couple days. The structure of the stored records can be limiting in terms of medication details, but the HL7 text stores much of the prescription details so data transformation at the dosing level

can be performed. The data are also reflective of what has been put into the original database, and in some cases it is difficult to infer what the clinician was intending, and whether any records were made in error. This is not necessarily due to the setup of the RPDR, but rather could be a limitation of the data entry system of the original database from which the data are acquired. However, as patient clinical data warehouses are becoming more popular as a repository for observational research, steps need to be taken to improve the quality and efficacy of how data are recorded in the original data entry database and how information is transferred into the data warehouse.

The consolidated data have some issues. General manipulation, including joining the medication and weight tables, and removing redundant and ambiguous data, decreased the initial group of 1299 medication records to 854. Even the 854 still contained some concerns regarding ambiguity, but were difficult to address with the data available without a chart review. Also, dosing for a single medication depends on not only indication, pharmacokinetics, age, weight and or surface area, but also in some cases whether or not it is a "starting" dose or a "maintenance" dose. Levothyroxine is an example of a drug used for life-long disease control, where determining the difference between the start dose and the maintenance dose may be significant. Unfortunately, the simple answer is not associating the "start" of a medication with the record with the "earliest" date. The patient may have been started on a medication at an institution whose data do not feed into this particular CDW. A potential solution is to use relationships that would indicate a "start" point of the disease, and medication that was prescribed during that point in time. For example, levothyroxine is initiated when the diagnosis of hypothyroidism is made through very elevated TSH levels. It may be possible to associate the earliest levothyroxine prescription with a timely highly elevated TSH value when looking for a "start dose."

Inherent with using patient data from a clinical data warehouse to derive peer practice patterns is the caveat that the user will accept some weakness in validity for the chance to see the information. The main purpose of this study is not to demonstrate that this collective intelligence approach can supplant or replace observational or randomized clinical trials, but rather that it is possible to present peer practice dosing practices from such data, akin and analogous to asking a group of clinicians how do they prescribe a particular medication and summarizing the results.

Although the data is not perfect, the study demonstrates that it can be done, and offers the potential for decision support based on peer practice data.

This brings us to the main point in the discussion of limitations - whether peer practice data can be used as clinical decision support. Initial resistance will most likely arise from concerns over validity of data and whether presenting such data is appropriate. The counter to these concerns are that as long as the user is aware of the limitations of the database and is willing to take responsibility for using such information, some users may find the information interesting if not useful. Many reference manuals often include disclaimers, declining responsibility for the accuracy of the information and the use of it.[14,19] Thus with even trusted resources, clinicians are using them at their own risk, and are held responsible for whatever medical decisions are made with or with out such references. Whether clinicians would be interested in such information is a question that would be best considered in the future direction section of this paper.

Finally, the quality of this work is limited by my level of expertise and experience in the sql query language and in the use of the applications utilized in this paper (MS Sql Server, MS Access, Excel and the R statistical program). I used knowledge gleaned from various books, websites and experienced people. Thus, better, more efficient methods of querying and data transformation may be possible.

FUTURE DIRECTIONS

There are several interesting questions raised by this study. One of the primary questions is whether the collective intelligence of peer practice would be used by clinicians as a form of clinical decision support. A way to study this is through a survey of clinicians to see whether they would be open to the concept, what kinds of information they would be looking for, and what are the issues keeping them from utilizing this potential form of CDSS. Another question is whether this type of information could have an impact on clinician practice. This question can be pursued through a study where users are tested on medication dosing for clinical case scenarios when provided data from CDW. Dosing recommendations can be provided from

[19] Hospital, J. H., Custer, J. W., Rau, R. E., & Lee, C. K. (2009). *The harriet lane handbook: A manual for pediatric house officers* Philadelphia, PA : Mosby/Elsevier, c2009.

reputed resources vs. from peer practice as recorded in the data warehouse, and the dosing tendencies can be measured for both groups to evaluate whether there are any significant differences. Another question is whether medication recommendations, particularly to the dosing level, can be linked with outcomes, such as optimal disease control. Since most data warehouses used in observational studies store data from billing information which often does not include prescription detail information, modeling medication dosing to outcome is an area of novel research. Finally, this concept can be taken to a higher level by using the data in the data warehouse to link medications prescribed with indications. Take for instance a clinician who is familiar with the diagnosis of ADHD, and the medications used to treat, but has limited experience prescribing for it. They can query the data warehouse to elicit medication practice patterns, and make medication management decisions based on the results.

CONCLUSION

The role of clinical decision support systems has been to provide the clinician user information that should aid in the medical management of their patients. For the most part, the CDSS have been developed based on established recommendations or the knowledge bank of domain experts. Hence, they can fall victim to a lack of timeliness and overgeneralization. Clinical databases hold a wealth of information regarding patient practice management by various clinicians and health care professionals for a multitude of medical conditions and clinical situations. This data not only represents practitioner collective intelligence but also real world practices, ranging from adequate and optimal medical management to suboptimal management to erroneous order or record entry. If the erroneous data and the suboptimal management can be filtered out, the optimal management can be harnessed, with the potential to provide more patient focused and timely information to the user, complementary to the knowledge from guidelines and other resources.

In this paper, I studied the barriers involved in posing a clinical question, then utilizing a clinical data warehouse to gather data which was then transformed to obtain an answer. While there were some limitations to the data, for the most part, the study demonstrated that it is possible to obtain peer practice information, regarding medication prescription to the dosing

level, and that the difference between peer practice and established recommendations can be considerable. Presenting such information to users may modify practices, and offers the potential of improved medical management based on utilizing the collective intelligence of real world peer practice. Overall, clinical decision support based on peer practice data from clinical data warehouses has the potential to fill in the knowledge gap that exists for clinical questions that do not have established recommendations or evidence-based answers. Even when such answers do exist, peer practice data may further complement and even challenge them and provide alternative, real-world management options.

ACKNOWLEDGEMENTS

I would like to acknowledge and thank the following individuals and institutions:

- Anil Dubey, my thesis advisor, for his support, help, and guidance throughout the execution of this thesis.

- James Signorovitch, for his assistance and guidance with the statistical methods used in this thesis.

- Lucila Ohno-Machado, M.D., PhD, MHA, for her guidance and support in my master's thesis work and postdoctoral fellowship studies.

- Andrew A. Scheufele, my husband, for his support and insight throughout the execution of this thesis.

- Shawn N. Murphy, M.D., PhD, and the RPDR application group, for making possible and allowing me to use the RPDR in my study.

- The National Library of Medicine, for funding my postdoctoral research fellowship in medical informatics.

- Henry Chueh, M.D., M.S. and the Massachusetts General Hospital, Laboratory of Computer Science, for hosting my postdoctoral research fellowship.

- The other postdoctoral research fellows, the faculty and developers at the LCS, for their fellowship as well as their help and feedback regarding my work on this thesis.

APPENDIX

Query A- Levothyroxine records for pediatric cohort, for MS Sql Server

```
/*the levothyroxine query performed for the CI project on 12/3/08*/
select
/*there were some commas in certain fields, which would affect the import to
MS Access so the 'replace' function just changes it to semi-colons*/
main.AgeAtMedDays as AgeatMedDays,
main.AgeAtMedYr as AgeAtMedYr,
replace(main.c_name, ',', ';') as shortname,
replace(main.c_fullname, ',', ';') as longname,
replace(main.patient_id_e, ',', ';') as patmedid,
replace(main.nval, ',', ';') as medval,
main.start_date as start_date,
replace(main.units_cd, ',', ';') as units,
replace(main.quantity_num, ',', ';') as quantity_num,
replace(main.encounter_id_e, ',', ';') as encounter_id,
main.date_of_birth as DOB,
replace(main.medname, ',', ';') as medname,
replace(main.dose, ',', ';') as dose,
replace(main.dose_unit, ',', ';') as dose_unit,
replace(main.freq, ',', ';') as freq,
replace(main.medroute, ',', ';') as medroute,
replace(main.directions, ',', ';') as directions
from
(select
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
       end as AgeAtMedYr,
DATEDIFF(day, pat.date_of_birth, meds.start_date) as AgeAtMedDays,
meds.c_name,
meds.c_fullname,
meds.patient_id_e,
meds.nval,
meds.start_date,
meds.units_cd,
meds.quantity_num,
meds.encounter_id_e,
pat.date_of_birth,
meds.hl7_text,
/*The following is to break up the information found in the HL7_text
regarding the medication.  This if the data that is used to evaluate
the daily dosage of medication.*/
/*Name*/
medname = case
when charindex('<LMR|Medications|MED_NAME>', meds.hl7_text) > 0 then
substring(meds.hl7_text,
charindex('<LMR|Medications|MED_NAME>', meds.hl7_text) + 26,
charindex('</LMR|Medications|MED_NAME>', meds.hl7_text) - 26 -
```

```
charindex('<LMR|Medications|MED_NAME>', meds.hl7_text))
else
'None'
End,
/*Dose*/
dose = case
when charindex('<LMR|Medications|DOSE>', meds.hl7_text) > 0 then
substring(meds.hl7_text,
charindex('<LMR|Medications|DOSE>', meds.hl7_text) + 22,
charindex('</LMR|Medications|DOSE>', meds.hl7_text) - 22 -
charindex('<LMR|Medications|DOSE>', meds.hl7_text))
else
'None'
End,
/*Dose unit*/
dose_unit = case
when charindex('<LMR|Medications|DOSE_UNITS>', meds.hl7_text) > 0 then
substring(meds.hl7_text,
charindex('<LMR|Medications|DOSE_UNITS>', meds.hl7_text) + 28,
charindex('</LMR|Medications|DOSE_UNITS>', meds.hl7_text) - 28 -
charindex('<LMR|Medications|DOSE_UNITS>', meds.hl7_text))
else
'None'
End,
/*frequency*/
freq = case
when charindex('<LMR|Medications|FREQUENCY_MNEMONIC>', meds.hl7_text) > 0
then
substring(meds.hl7_text,
charindex('<LMR|Medications|FREQUENCY_MNEMONIC>', meds.hl7_text) + 36,
charindex('</LMR|Medications|FREQUENCY_MNEMONIC>', meds.hl7_text) - 36 -
charindex('<LMR|Medications|FREQUENCY_MNEMONIC>', meds.hl7_text))
else
'None'
End,
/*route*/
medroute = case
when charindex('<LMR|Medications|ROUTE>', meds.hl7_text) > 0 then
substring(meds.hl7_text,
charindex('<LMR|Medications|ROUTE>', meds.hl7_text) + 23,
charindex('</LMR|Medications|ROUTE>', meds.hl7_text) - 23 -
charindex('<LMR|Medications|ROUTE>', meds.hl7_text))
else
'None'
End,
/*direction*/
directions = case
when charindex('<LMR|Medications|DIRECTIONS>', meds.hl7_text) > 0 then
substring(meds.hl7_text,
charindex('<LMR|Medications|DIRECTIONS>', meds.hl7_text) + 28,
charindex('</LMR|Medications|DIRECTIONS>', meds.hl7_text) - 28 -
charindex('<LMR|Medications|DIRECTIONS>', meds.hl7_text))
else
'NoDirections'
End
```

```
from dbo.dw_dim_patient as pat
inner join
(select con.c_name, con.c_fullname, fact.patient_id_e,
fact.nval, fact.start_date, fact.units_cd, fact.quantity_num,
fact.encounter_id_e, fact.hl7_text
from dbo.dw_dim_concept as con
inner join dbo.dw_f_conc_noval as fact
on con.c_basecode = fact.concept_id
/*the best way to find find all the levothyroxines is to look for them under
the 'c_name' as 'levothy'.  Also, because the dosage value is important,
entries that were null were selected out.*/
where con.c_name like '%levothy%'
and fact.nval not like '%NULL%') as meds
on pat.patient_id_e = meds.patient_id_e
where
/*The following is the only way to get the number of years in age of the
patient.  Performing a 'DATEDIFF' with 'year' only give the subtraction of
the year for one date with the year of the next date, which means that
in most cases, the date is rounding up.*/
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
       end <= 17
and DATEDIFF(day, pat.date_of_birth, meds.start_date) >= 0) as main
```

Query B- Weight records for pediatric cohort, for MS Sql Server

```
/*The following is the query performed to obtain the demographics data of the
patients in the medication query. Query performed 12-09-2008*/

select
/*there were some commas in certain fields, which would affect the import to
MS Access so the 'replace' function just changes it to semi-colons*/
main.AgeAtMedDays as AgeatMedDays,
main.AgeAtMedYr as AgeAtMedYr,
replace(main.patient_id_e, ',', ';') as patmedid,
replace(main.sex_cd, ',', ';') as gender,
replace(main.langauge_cd, ',', ';') as Ptlanguage,
replace(main.race_cd, ',', ';') as race,
replace(main.religion_cd, ',', ';') as religion,
replace(main.zip_cd, ',', ';') as zip,
replace(main.date_of_birth, ',', ';') as DOB,
replace(main.date_of_death, ',', ';') as date_of_death
from
(select
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
       end as AgeAtMedYr,
DATEDIFF(day, pat.date_of_birth, meds.start_date) as AgeAtMedDays,
meds.patient_id_e,
pat.sex_cd,
pat.langauge_cd,
pat.race_cd,
pat.religion_cd,
pat.zip_cd,
pat.date_of_birth,
pat.date_of_death
from dbo.dw_dim_patient as pat
inner join
(select con.c_name, con.c_fullname, fact.patient_id_e,
fact.nval, fact.start_date, fact.units_cd, fact.quantity_num,
fact.encounter_id_e, fact.hl7_text
from dbo.dw_dim_concept as con
inner join dbo.dw_f_conc_noval as fact
on con.c_basecode = fact.concept_id
/*the best way to find find all the levothyroxines is to look for them under
the
'c_name' as 'levothy'.  Also, because the dosage value is important, entries
that were null were selected out.*/
where con.c_name like '%levothy%'
and fact.nval not like '%NULL%') as meds
on pat.patient_id_e = meds.patient_id_e
where
/*The following is the only way to get the number of years in age of the
patient.  Performing a 'DATEDIFF' with 'year' only give the subtraction of
the year for one date with the year of the next date, which means that
in most cases, the date is rounding up.*/
```

```
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
        end <= 17
and DATEDIFF(day, pat.date_of_birth, meds.start_date) >= 0) as main
```

Query C: Demographic records for pediatric cohort, for MS Sql Server

```
/*The following is the query performed to obtain the demographics data of the
patients in the medication query. Query performed 12-09-2008*/

select
/*there were some commas in certain fields, which would affect the import to
MS Access so the 'replace' function just changes it to semi-colons*/
main.AgeAtMedDays as AgeatMedDays,
main.AgeAtMedYr as AgeAtMedYr,
replace(main.patient_id_e, ',', ';') as patmedid,
replace(main.sex_cd, ',', ';') as gender,
replace(main.langauge_cd, ',', ';') as Ptlanguage,
replace(main.race_cd, ',', ';') as race,
replace(main.religion_cd, ',', ';') as religion,
replace(main.zip_cd, ',', ';') as zip,
replace(main.date_of_birth, ',', ';') as DOB,
replace(main.date_of_death, ',', ';') as date_of_death
from
(select
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
       end as AgeAtMedYr,
DATEDIFF(day, pat.date_of_birth, meds.start_date) as AgeAtMedDays,
meds.patient_id_e,
pat.sex_cd,
pat.langauge_cd,
pat.race_cd,
pat.religion_cd,
pat.zip_cd,
pat.date_of_birth,
pat.date_of_death
from dbo.dw_dim_patient as pat
inner join
(select con.c_name, con.c_fullname, fact.patient_id_e,
fact.nval, fact.start_date, fact.units_cd, fact.quantity_num,
fact.encounter_id_e, fact.hl7_text
from dbo.dw_dim_concept as con
inner join dbo.dw_f_conc_noval as fact
on con.c_basecode = fact.concept_id
/*the best way to find find all the levothyroxines is to look for them under
the
'c_name' as 'levothy'.  Also, because the dosage value is important, entries
that were null were selected out.*/
where con.c_name like '%levothy%'
and fact.nval not like '%NULL%') as meds
on pat.patient_id_e = meds.patient_id_e
where
/*The following is the only way to get the number of years in age of the
patient.  Performing a 'DATEDIFF' with 'year' only give the subtraction of
the year for one date with the year of the next date, which means that
```

```
in most cases, the date is rounding up.*/
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
        end <= 17
and DATEDIFF(day, pat.date_of_birth, meds.start_date) >= 0) as main

/*This is the query to get the demographics of the patients who are in the
medication query above.*/
select
/*there were some commas in certain fields, which would affect the import to
MS Access so the 'replace' function just changes it to semi-colons*/
main.AgeAtMedDays as AgeatMedDays,
main.AgeAtMedYr as AgeAtMedYr,
replace(main.patient_id_e, ',', ';') as patmedid,
replace(main.sex_cd, ',', ';') as gender,
replace(main.langauge_cd, ',', ';') as Ptlanguage,
replace(main.race_cd, ',', ';') as race,
replace(main.religion_cd, ',', ';') as religion,
replace(main.zip_cd, ',', ';') as zip,
replace(main.date_of_birth, ',', ';') as DOB
from
(select
year(meds.start_date)
     - year(pat.date_of_birth)
     - case when month(pat.date_of_birth) > month(meds.start_date) then 1
            when month(pat.date_of_birth) < month(meds.start_date) then 0
            when day(pat.date_of_birth) > day(meds.start_date) then 1
            else 0
        end as AgeAtMedYr,
DATEDIFF(day, pat.date_of_birth, meds.start_date) as AgeAtMedDays,
meds.patient_id_e,
pat.sex_cd,
pat.langauge_cd,
pat.race_cd,
pat.religion_cd,
pat.zip_cd,
pat.date_of_birth
from dbo.dw_dim_patient as pat
inner join
(select con.c_name, con.c_fullname, fact.patient_id_e,
fact.nval, fact.start_date, fact.units_cd, fact.quantity_num,
fact.encounter_id_e, fact.hl7_text
from dbo.dw_dim_concept as con
inner join dbo.dw_f_conc_noval as fact
on con.c_basecode = fact.concept_id
/*the best way to find find all the levothyroxines is to look for them under
the
'c_name' as 'levothy'.  Also, because the dosage value is important, entries
that were null were selected out.*/
where con.c_name like '%levothy%'
and fact.nval not like '%NULL%') as meds
```

```
on pat.patient_id_e = meds.patient_id_e
where
/*The following is the only way to get the number of years in age of the
patient.  Performing a 'DATEDIFF' with 'year' only give the subtraction of
the year for one date with the year of the next date, which means that
in most cases, the date is rounding up.*/
year(meds.start_date)
      - year(pat.date_of_birth)
      - case when month(pat.date_of_birth) > month(meds.start_date) then 1
             when month(pat.date_of_birth) < month(meds.start_date) then 0
             when day(pat.date_of_birth) > day(meds.start_date) then 1
             else 0
         end <= 17
and DATEDIFF(day, pat.date_of_birth, meds.start_date) >= 0) as main
```

Query D1 of 4. This query returns a table with the medication encounter, the weight encounter and the number of days from weight encounter to the medication encounter, allowing only positive values or zero for the difference, for MS Access.

```
SELECT m.encounter_id AS medenct, w.encounter_id AS wtenct, m.ageatmeddays-
w.patwtagedays AS medwtdaydiff
FROM CI_levothy_data AS m INNER JOIN CI_wt_data AS w ON w.patient_id =
m.patmedid
WHERE (m.ageatmeddays-w.patwtagedays) >= 0;
```

Query D2 of 4. This query returns a table of medication encounters, with the lowest days between the dates of the weight and medication encounters, for MS Access. (There is the assumption that the table that the query is accessing only has non negative numbers for the value of 'medwtdaydiff').

```
SELECT medenct, min(medwtdaydiff) AS minmedwtdays
FROM q1_medwtenctsonlyunionall
GROUP BY medenct;
```

Query D3 of 4. This query is an inner join of the data in the medication and the weight tables, with the addition of the attribute for the number of days between the weight and the medication encounters, for MS Access.

```
SELECT m.ageatmedyr AS medageyr, m.ageatmeddays AS medagedays, m.ageatmeddays
- w.patwtagedays AS medwtdaydiff, m.medval AS medval, m.medval/wtkgval AS
medkgval, m.shortname AS medshortname, m.patmedid AS ptid, m.units AS
medunits, m.encounter_id AS medenct, w.patwtageyr AS wtageyr, w.patwtagedays
AS wtagedays, w.wtval AS wtval, w.encounter_id AS wtenct, w.wtval/2.2 AS
wtkgval, w.units AS wtunits, m.medname AS medname, m.dose AS dose,
m.dose_unit AS dose_unit, m.freq AS freq, m.medroute AS medroute,
m.directions AS directions
FROM CI_levothy_data AS m INNER JOIN CI_wt_data AS w ON m.patmedid =
w.patient_id;
```

Query D4 of 4. This query returns the final table of attributes, for MS Access.

```
SELECT q3.medageyr, q3.medagedays, q3.medwtdaydiff, q3.medval, q3.medkgval,
q3.medshortname, q3.ptid, q3.medunits, q3.medenct, q3.wtageyr, q3.wtagedays,
q3.wtval, q3.wtenct, q3.wtkgval, q3.wtunits, q3.medname, q3.dose,
q3.dose_unit, q3.freq, q3.medroute, q3.directions
FROM q2_leasttimebwmedandwt AS q2 INNER JOIN q3_MedWtUnionall AS q3 ON
q2.medenct = q3.medenct
WHERE q2.minmedwtdays = q3.medwtdaydiff
ORDER BY q3.medenct;
```

Frequency of text in medication "Name" detail			
medname	raw	clean	c/r%
LEVOTHYROXINE SODIUM	427 (39.25)	344 (40.28)	80.56
LEVOXYL (LEVOTHYROXINE SODIUM)	294 (27.02)	233 (27.28)	79.25
SYNTHROID (LEVOTHYROXINE SODIUM)	273 (25.09)	205 (24)	75.09
LEVOTHYROXINE SODIUM (LEVOTHYROXINE SODIUM)	39 (3.58)	32 (3.75)	82.05
LEVOTHROID (LEVOTHYROXINE SODIUM)	23 (2.11)	21 (2.46)	91.3
THYROXINE (LEVOTHYROXINE SODIUM)	13 (1.19)	6 (0.7)	46.15
L-THYROXINE (LEVOTHYROXINE SODIUM)	10 (0.92)	7 (0.82)	70
SYNTHROID	4 (0.37)	4 (0.47)	100
LEVOTHYROXINE SODIUM INJ	3 (0.28)	0 (0)	0
LEVOXINE (LEVOTHYROXINE SODIUM)	2 (0.18)	2 (0.23)	100

Figure A. Frequency of text for medication detail of name, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation.

Frequency of text in "Unit" medication detail			
dose_unit	raw	consolidated	c/r%
MCG	1049 (96.42)	822 (96.25)	78.36
MCG TAKE 1/2 TAB	11 (1.01)	10 (1.17)	90.91
MCG TAKE 1 1/2 TABLETS	4 (0.37)	4 (0.47)	100
MCG TAKE 1 1/2 TABLET	4 (0.37)	2 (0.23)	50
MCG (1/2 OF A 137 MCG TABLET)	4 (0.37)	2 (0.23)	50
MCG (1/2 OF A 125 MCG TABLET)	3 (0.28)	3 (0.35)	100
None	2 (0.18)	2 (0.23)	100
MG	2 (0.18)	2 (0.23)	100
MCG/ML SUSPENSION	2 (0.18)	2 (0.23)	100
MCG PILL	2 (0.18)	1 (0.12)	50
MCG (1/2 OF 125MCG TABLET)	1 (0.09)	1 (0.12)	100
MCG (1/2 OF 125MCG TAB	1 (0.09)	1 (0.12)	100
MCG (1/2 OF 125 MCG TABLET)	1 (0.09)	0 (0)	0
MCG (1/2 OF 125 MCG TA	1 (0.09)	1 (0.12)	100
MCG TAKE 1/2 TAB	1 (0.09)	1 (0.12)	100

Figure B. Frequency of text for medication detail of unit of measure, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation.

Frequency of text in "Frequency" medication detail			
freq	raw	consolidated	c/r%
QD	940 (86.4)	743 (87)	79.04
as directed	26 (2.39)	20 (2.34)	76.92
QOD	26 (2.39)	15 (1.76)	57.69
QAM	14 (1.29)	7 (0.82)	50
take 1/2 tablet daily	13 (1.19)	11 (1.29)	84.62
daily	8 (0.74)	7 (0.82)	87.5
BID	6 (0.55)	5 (0.59)	83.33
Take 1/2 tablet QD	5 (0.46)	4 (0.47)	80
alternate daily with 2 tabs	4 (0.37)	4 (0.47)	100
4 days a week	3 (0.28)	3 (0.35)	100
once daily	3 (0.28)	1 (0.12)	33.33
1/2 tablet daily	3 (0.28)	3 (0.35)	100
Q24H	3 (0.28)	3 (0.35)	100
Take one tablet 25 mcg on Monday Wednesday and Friday and 1/2 tablet (12.5 mcg) rest of the week.	3 (0.28)	2 (0.23)	66.67
take 1/2 tablet (56 mcg) daily	3 (0.28)	3 (0.35)	100
give 25mcg on Tuesday Thursday and Sunday and 37.5mcg on Monday Wednesday and Saturday	2 (0.18)	2 (0.23)	100
alternate daily with 2	2 (0.18)	1 (0.12)	50
take 44mcg daily (1/2 tablet)	2 (0.18)	2 (0.23)	100
every other day	2 (0.18)	1 (0.12)	50
QHS	2 (0.18)	2 (0.23)	100
on Mondays Wednesday and Fridays	2 (0.18)	1 (0.12)	50
pill daily	2 (0.18)	1 (0.12)	50
Take 1/2 tablet (37.5mcg) daily	2 (0.18)	2 (0.23)	100
on Mondays Wednesday a	1 (0.09)	1 (0.12)	100
take 1/2 tablet (37.5 mcg) daily	1 (0.09)	0 (0)	0
q mon tu weds thurs	1 (0.09)	1 (0.12)	100
daily as directed	1 (0.09)	1 (0.12)	100
Take one tablet alternating with 37.5mcg (1 1/2 tablets) daily	1 (0.09)	1 (0.12)	100
Take 62.5 mcg daily QD	1 (0.09)	1 (0.12)	100
take 62.5mcg daily (1/	1 (0.09)	1 (0.12)	100
take 62.5mcg daily (1/2 tablet)	1 (0.09)	1 (0.12)	100
QAM on an empty stomac	1 (0.09)	1 (0.12)	100
5 days weekly	1 (0.09)	1 (0.12)	100
take 1/2 tablet (37.5	1 (0.09)	1 (0.12)	100
every tues and thurs	1 (0.09)	1 (0.12)	100

Figure C. Frequency of text for medication detail of frequency, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation.

Frequency of text in medication "Direction" detail			
directions	raw	clean	c/r%
NoDirections	750 (68.93)	595 (69.67)	79.33
1 tablet daily for hypothyroidism	19 (1.75)	16 (1.87)	84.21
take 1 tablet daily	17 (1.56)	16 (1.87)	94.12
One tablet daily for hypothyroidism	12 (1.1)	12 (1.41)	100
per endocrine	10 (0.92)	9 (1.05)	90
Crush pill and mix with liquid and give via syringe/dropper once a day	8 (0.74)	3 (0.35)	37.5
One tablet daily	7 (0.64)	7 (0.82)	100
Take 1 tablet by mouth daily	7 (0.64)	5 (0.59)	71.43
1 tablet alternating with 1/2 tablet (25 mcg/12.5 mcg) daily for hypothyroidism	6 (0.55)	1 (0.12)	16.67
3 month suppley	6 (0.55)	2 (0.23)	33.33
One pill daily Mon thru Sat; one and one half a pill on Sun. Take in am; on an empty stomach; without other medications; 30-60 minutes prior to breakfast.	6 (0.55)	4 (0.47)	66.67
One tablet daily for congenital hypothyroidism	6 (0.55)	4 (0.47)	66.67
1 tablet daily	5 (0.46)	4 (0.47)	80
75 mcg twice a week; 50 mcg other days	5 (0.46)	2 (0.23)	40
Repeat TSH and free T4 in a month	5 (0.46)	3 (0.35)	60
Take 1/2 tablet (37.5 mcg) orally each day	5 (0.46)	3 (0.35)	60
90 day supply	4 (0.37)	3 (0.35)	75
DAW	4 (0.37)	2 (0.23)	50
may crush and put in 2 tsp of water	4 (0.37)	3 (0.35)	75
Needs to attend next appointment.	4 (0.37)	2 (0.23)	50
NO SUBSTITUTION	4 (0.37)	3 (0.35)	75
No substitutions	4 (0.37)	3 (0.35)	75
per endocrine- once daily one hour before or after food	4 (0.37)	2 (0.23)	50
take 1 tablet of 100 mcg PO daily	4 (0.37)	3 (0.35)	75
take 1 tablet twice a day; May crush up and mix with formula	4 (0.37)	4 (0.47)	100
take 1/2 tablet daily	4 (0.37)	4 (0.47)	100
take as crushed tablet	4 (0.37)	3 (0.35)	75
take half of a tablet(62.5 mcg) daily	4 (0.37)	4 (0.47)	100
Take one pill of 75 mcg Synthroid with 10 mcg of Cytomel daily. No substitution	4 (0.37)	2 (0.23)	50
75 mcg on the other days.	3 (0.28)	2 (0.23)	66.67
90=3 months supply	3 (0.28)	3 (0.35)	100
can crush tablet; don't take at same time of day with multivitamin or soy products	3 (0.28)	2 (0.23)	66.67
Give one half tablet every day	3 (0.28)	2 (0.23)	66.67

Figure D1. Frequency of text for medication detail of directions, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation. (Personal identifiers have been removed.)

Frequency of text in medication "Direction" detail			
directions	raw	clean	c/r%
NO SUBSTITUTION: XXX is allergic to Levoxyl	3 (0.28)	2 (0.23)	66.67
one tablet by mouth daily	3 (0.28)	3 (0.35)	100
One tablet daily by mouth for hypothyroidism	3 (0.28)	3 (0.35)	100
Please take 1/2 of the 75mcg tablet. Crush and place in breastmilk or formula.	3 (0.28)	2 (0.23)	66.67
Please take 25mcg (1 tablets) and then 37.5 mcg (1.5 tablets) alternating.	3 (0.28)	3 (0.35)	100
SI se olvida; por favor tomar el doble el dia siguiente	3 (0.28)	3 (0.35)	100
Take 1 tablet PO daily 1 hour before eating.	3 (0.28)	2 (0.23)	66.67
Take 37.5 mcg every other day alternating with 25 mcg	3 (0.28)	2 (0.23)	66.67
Take half a tablet (i.e. 37 mcg) PO daily	3 (0.28)	2 (0.23)	66.67
1 pill po qd	2 (0.18)	2 (0.23)	100
3 month supply = 135 tablets	2 (0.18)	2 (0.23)	100
90 day supply # XXX	2 (0.18)	1 (0.12)	50
do not prescribe Levoxyl	2 (0.18)	2 (0.23)	100
Generic	2 (0.18)	1 (0.12)	50
If you forget to take your medication; please take it as soon as you remember or double your dose the next day. Do not take more than a double dose.	2 (0.18)	2 (0.23)	100
Need to make f/u appointment with MD	2 (0.18)	2 (0.23)	100
no substitutions.	2 (0.18)	0 (0)	0
Non generic please	2 (0.18)	1 (0.12)	50
One tablet every other day daily for hypothyroidism	2 (0.18)	2 (0.23)	100
Please give one tablet a day. If you forgot one day; please double up on the next day's dose. Do not adminsiter with soy or iron containing medications/foods.	2 (0.18)	2 (0.23)	100
Please take 1 tablet. Crush and place in breastmilk or formula.	2 (0.18)	2 (0.23)	100
Reference# XXX	2 (0.18)	2 (0.23)	100
Start with one tab daily; increase to two tabs daily after one week	2 (0.18)	0 (0)	0
take 1 tab every other day	2 (0.18)	2 (0.23)	100
take 1 tablet of 112 mcg alternating with 100 mcg PO daily	2 (0.18)	1 (0.12)	50
Take 1 tablet PO daily	2 (0.18)	2 (0.23)	100
Take 1/2 of a 75 mcg tablet (37.5 mcg) every other day alternating with 44 mcg levothyroxine (1/2 of 88 mcg pill)	2 (0.18)	0 (0)	0
Take 1/2 tab or 37.5 mcg everyday	2 (0.18)	2 (0.23)	100
take 1/2 tablet by mouth daily	2 (0.18)	2 (0.23)	100
take 1/2 tablet PO QD x 1 week; then increase to 1 tablet	2 (0.18)	2 (0.23)	100
take half a tablet 37.5 mcg per day	2 (0.18)	1 (0.12)	50
take half a tablet 68.5 mcg every	2 (0.18)	1 (0.12)	50

Figure D2. Frequency of text for medication detail of directions, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation. (Personal identifiers have been removed.)

Frequency of text in medication "Direction" detail			
directions	raw	clean	c/r%
Take one 88mcg tablet every other day alternating with 100mcg tablet on the other days.	2 (0.18)	1 (0.12)	50
Take one tablet daily	2 (0.18)	2 (0.23)	100
take one tablet daily for hypothyroidism	2 (0.18)	1 (0.12)	50
Take one tablet of the 100mcg tablet every other day alternating with 88mcg tablet on the other days.	2 (0.18)	0 (0)	0
this is a dose increase	2 (0.18)	2 (0.23)	100
Will need TSH test 4-6 weeks after starting this medication. Call office.	2 (0.18)	2 (0.23)	100
You can use up the old dose of 50 mcg/day by taking 8 pills per week and then switch to the new prescription.	2 (0.18)	1 (0.12)	50
1 tablet daily for hypothyroidism. NEED TO MAKE FOLLOW_UP APPOINTMENT ASAP	1 (0.09)	1 (0.12)	100
1 tablet daily for hypothyroidism; 60d supply	1 (0.09)	0 (0)	0
1 tablet PO daily	1 (0.09)	1 (0.12)	100
1/2 tab = 68.5mcg daily	1 (0.09)	1 (0.12)	100
162.5 mcg PO daily (150 mcg tablet + 1/2 of a 25 mcg tablet)	1 (0.09)	1 (0.12)	100
44 mcg twice a week	1 (0.09)	1 (0.12)	100
60 tabs = 2 months supply	1 (0.09)	1 (0.12)	100
6X/WK	1 (0.09)	1 (0.12)	100
90 d supply	1 (0.09)	1 (0.12)	100
90d supply	1 (0.09)	0 (0)	0
alternate with 37.5 mcg (1 1/2 tabs) qod	1 (0.09)	1 (0.12)	100
alternate with Levoxyl 75 mcg	1 (0.09)	0 (0)	0
alternate with levoxyl 88mcg	1 (0.09)	1 (0.12)	100
XXX is allergic to Levoxyl	1 (0.09)	1 (0.12)	100
May crush pill and mix in applesauce or liquid	1 (0.09)	1 (0.12)	100
Mother called and request to have a rx on file for next month	1 (0.09)	1 (0.12)	100
NEEDS TO GO TO UPCOMING APP WITH DR. XXX	1 (0.09)	1 (0.12)	100
no further refills until seen by MD	1 (0.09)	1 (0.12)	100
NO SUBSTITUTIONS PER CHILDREN'S HOSPITAL ENDOCRINOLOGY	1 (0.09)	1 (0.12)	100
Non generic please but if generic used please use same manufacturer for all	1 (0.09)	1 (0.12)	100
Note to pharmacist: Please dispense Sandoz manufactured levothyroxine.	1 (0.09)	1 (0.12)	100
On 2/8/08 please start 1.5 tablet daily for hypothyroidism.	1 (0.09)	1 (0.12)	100
On 2/8/08 please start 1/2 tablet daily for hypothyroidism.	1 (0.09)	1 (0.12)	100
One laf tablet (68.5 mcg) daily by mouth for hypothyroidism	1 (0.09)	1 (0.12)	100
One tablet daily - no substitution	1 (0.09)	1 (0.12)	100
One tablet orally daily	1 (0.09)	1 (0.12)	100
One-half tablet (68.5 mcg) daily by mouth for hypothyroidism	1 (0.09)	0 (0)	0
Please alternate this dose with Levoxyl 25 mcg every other day	1 (0.09)	1 (0.12)	100

Figure D3. Frequency of text for medication detail of directions, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation. (Personal identifiers have been removed.)

Frequency of text in medication "Direction" detail			
directions	raw	clean	c/r%
Please alternate with 125 mcg tablet every other day	1 (0.09)	1 (0.12)	100
PLEASE CALL OFFICE	1 (0.09)	1 (0.12)	100
please dispense 137 mcg tablets; patient to take 1/2 tablet (68.5 mcg) daily	1 (0.09)	1 (0.12)	100
please dispense 2 month supply on this one occassion since family will be out of the country July - August	1 (0.09)	1 (0.12)	100
Please dispense Levoxyl (no substitutions). Take 1 tablet PO daily.	1 (0.09)	1 (0.12)	100
Please give 1 and 1/2 tablets a day beginning on 3/1/08.	1 (0.09)	1 (0.12)	100
Please provide 1 years worth of refills.	1 (0.09)	1 (0.12)	100
Please take 1/2 tablet once a day. Crush it and mix with water. Can administer with syringe.	1 (0.09)	1 (0.12)	100
Please take Levothyroxine 225 mcg po daily. Take 1 tablet of 200mcg and 1 tablet of the 25mcg daily.	1 (0.09)	1 (0.12)	100
Repeat labs in a month	1 (0.09)	0 (0)	0
Repeat TSH and free T4 in six months	1 (0.09)	1 (0.12)	100
take 1 tablwt daily	1 (0.09)	1 (0.12)	100
take 1.5 tablets daily by mouth	1 (0.09)	1 (0.12)	100
Take 1/2 of a 75 MCG Tablet	1 (0.09)	1 (0.12)	100
Take 1/2 tab or 37.5 mcg on odd days and 25 mcg on even days	1 (0.09)	1 (0.12)	100
Take 1/2 tab or 37.5 mcg PO QD	1 (0.09)	1 (0.12)	100
Take 1/2 tablet (62.5mg) tablet daily for hypothyroidism	1 (0.09)	1 (0.12)	100
take 1/2 tablet equal to 62.5 mcg daily	1 (0.09)	1 (0.12)	100
Take 1/2 tablet every other day alternating with 44 mcg levothyroxine (1/2 of 88 mcg pill)	1 (0.09)	1 (0.12)	100
Take 2 pills on Sundays (8 pills total per week)	1 (0.09)	1 (0.12)	100
Take 25 mcg on even days and 37.5 mcg on odd days	1 (0.09)	0 (0)	0
take daily with levoxyl 125mcg 1/2 tab to equal 162.5 mcg	1 (0.09)	1 (0.12)	100
take daily with levoxyl 125mcg 1/2 tab to equal 162.5 mcg #XXX	1 (0.09)	1 (0.12)	100
Take half of 137mcg (68.5 MCG) tablet PO Daily	1 (0.09)	1 (0.12)	100
Take HALF of a tablet (i.e. 37.5 mcg) PO daily.	1 (0.09)	1 (0.12)	100
Take one pill daily	1 (0.09)	1 (0.12)	100
Take one pill of 75 mcg Synthroid with 10 mcg of Cytomel daily	1 (0.09)	1 (0.12)	100
Take one to two tablets as directed by clinic.	1 (0.09)	1 (0.12)	100
Take with meals.	1 (0.09)	1 (0.12)	100
This is a change in dose	1 (0.09)	1 (0.12)	100
This is a dose decrease.	1 (0.09)	1 (0.12)	100
This is a dose decrease. Recheck labs in 4-6 weeks.	1 (0.09)	1 (0.12)	100
TO take 100 mcg po daily	1 (0.09)	1 (0.12)	100
TO take 100 mcg po daily No Substitutions	1 (0.09)	0 (0)	0
will check labs in a month on new dose; if levels good; will provide 90d Rx then	1 (0.09)	1 (0.12)	100

Figure D4. Frequency of text for medication detail of directions, from the HL7 text. First column and second column show the number (and percent of the data group) of the text found in the raw and the consolidated data, respectively. The third column shows the percentage (c/r%) of the text that remained after the data manipulation. (Personal identifiers have been removed.)

To remove duplicate records from a table, follow these steps:

1. Make a copy of the structure of the table from which you want to remove the duplicate records.

   To copy a table:
   1. Select the table in the Database window
   2. On the Edit menu, click Copy.
   3. On the Edit menu, click Paste.
   4. Enter a name for the new table.
   5. Select Structure Only
   6. Click OK.
2. Open the new table in Design view.
3. Select the field(s) that contain the duplicate values.
4. To make your selection the primary key for the table, click the Primary Key button on the toolbar.
5. Save and close the table.
6. Create an append query based on the original table containing duplicates.
7. In the query Design View, on the Query menu, click Append Query.
8. In the Append dialog box, click the name of the new table from the Table Name list, and then click OK.
9. Include all the fields from the original table by dragging the asterick (*) to the query design grid.
10. On the Query menu, click Run.
11. Click Yes in the dialog box advising you that you are about to append records.
12. Because the Primary Key field(s) in the new table will not accept duplicate values, the following error message will be displayed:

> Microsoft Access can't append all the records in the append query.
>
> Microsoft Access set 0 field(s) to Null due to a type conversion failure, and it didn't add <number> record(s) to the table due to key violations, 0 record(s) due to lock violations, and 0 record(s) due to validation rule violations.
> Do you want to run the action query anyway?
> To ignore the error(s) and run the query, click Yes. For an explanation of the causes of the violations, click Help.

13. Click Yes.
14. View the contents of the new table. When you're sure the new table has the correct unique records, you can delete the original table, and then rename the new table using the name of the original table.

Figure E. Instructions on how to duplicate a table and also remove duplicate records using MS Access. Obtained from http://support.microsoft.com/kb/209183 on 01-20-08.