Managing Evolving Code Sets and Integration of Multiple Data Sources in Health Care Analytics

James P. McGlothlin
Fusion Consulting Inc
5215 North O'Connor Blvd Suite 650
Irving, TX 75039
jmcglothlin@fusionconsultinginc.com

Latifur Khan
The University of Texas at Dallas
800 W. Campbell Road
Richardson, Texas 75080
lkhan@utdallas.edu

ABSTRACT

This paper presents our industry experience related to developing data warehouses for healthcare analytics. With the rapid advancement of medical record digitization, there is a very large amount of information available for analysis. With the heavy focus on driving down health care costs, managing preventive care and improving patient outcomes and satisfaction, there is a growing emphasis on healthcare metrics and analytics. The information for a single patient's history is composed of data from every hospital, provider, lab, pharmacy and insurance company the patient has encountered. This information needs to be viewed as a whole to accurately analyze the patient's health. In turn, each patient's complete health information is needed to accurately evaluate the performance of his or her providers. This paper will address some challenges we have faced when merging and correlating these diverse data sources. We will provide our solutions and experience addressing key challenges including code set integration and migration and patient identification.

Categories and Subject Descriptors

[**Information Systems**] : Data Management Systems –
Information Integration: Data Warehouses

General Terms

Management, Design.

Keywords

Data Warehouses, Healthcare, ETL, Inference, Data Integration.

1. INTRODUCTION

There is a growing interest in using computer analytics to both improve the cost efficiency of health care and to improve patient quality and outcomes. In the United States, many health care initiatives arise from CMS (Centers for Medicare and Medicaid Services). In 2010, the Health Information Technology for Economic and Clinical Health Act (HITECH) began incentive programs for health care providers who utilize electronic health record software systems (EHRs) to improve delivery of care.[13][24] More recently, the Patient Protection and Affordable Care Act (ACA)[1] relies on information technology

as a key method to drive down health care costs. As defined by ACA, CMS is responsible for overseeing the testing of innovative payment and delivery models. Initiatives such as Medicare and Medicaid EHR Incentive Programs, Meaningful Use[24], Physician Quality Reporting System (PQRS)[25][38], Accountable Care Organizations (ACOs)[18][19], Patient-Centered Medical Home (PCMH)[12], Patient-Centered Outcomes Research Institute, and Pay for Performance are all designed to use information technology to provide better health care at a lower cost. There are over 40 pay for performance programs for inpatient care alone[2]. All of these incentives and organizations encourage and require increased analytical assessment of health care analytics. It is easy to see by the number of new regulations and incentives programs, that such metrics and analytics must be flexible and support rapid change.

Hospitals are required to be fully digitized by 2015 or face penalties. As a result, there is an increasing amount of data available for business intelligence analytics and reporting. Clearly, such analysis and reporting is required to adhere to new regulations and to qualify for new incentive programs. Moreover, quality analysis of this data can be used to improve best practices for patient care, find inefficiencies and anomalies and to identify opportunities to improve both outcomes and operational efficiency. Under the incentives programs, organizations that are successful in such analysis can receive financial rewards. As some examples, analysis can identify bottlenecks in emergency department responsiveness, inefficiencies in bed assignments or operating room scheduling, and at-risk patients who would benefit from certain medications or diagnostics.

For many reasons, including the need for increased IT capital expenditures, there is a recent surge in the United States in hospital mergers and collaborations.[26] There are also provider groups and networks where patient care is managed across providers and hospitals. The ACO model encourages and rewards hospital networks and integrated healthcare systems. [26] There are also community organizations designed to promote information exchange between health care organizations (health information exchanges or HIEs)[27], and there are growing networks of community hospitals.

While certainly not a simple or small task, building a data warehouse for business analytics for a single hospital on a single electronic medical record system (EMR) is relatively straightforward. The main challenge is the different ways of documenting similar or related patient events. This becomes especially cumbersome as documentation methods and coding practices change over time. This problem becomes more difficult when the hospital uses multiple software systems for clinical care, billing, supply chain management, diagnostic test management, etc. The difficulty is further magnified when multiple hospitals are involved, especially if these hospitals use different EMRs.

The challenge becomes much greater when the goal is to integrate patient data from many providers and hospitals using different software systems to manage the records of the same patients. To achieve a complete understanding of patient care, insurance claims records, pharmacy records, and lab records should also be included. Thus, the problem of integrating and merging patient information from multiple sources and documentation methods is the central challenge of building a health care data warehouse.

This paper will present a case study of our industry experience related to developing data warehouses for healthcare. We will focus primarily on the key challenges of dealing with multiple and evolving code sets and of integrating information from multiple data sources. In Section 2, we will briefly explain our approach and methodology and provide background related to the process and goals. In Section 3, we will identify a few key challenges that we have faced and will discuss in the paper. In Section 4, we will examine the key challenge of code set integration and migration. In Section 5, we will address the key challenge of patient identification. In Section 6, we will describe our results. In Section 7, we will identify some areas we have chosen for future improvement.

2. BACKGROUND AND APPROACH

Our industry experience revolves around development of healthcare data warehouses and datamarts for clinical, operational and financial analytics. We have worked with source systems including large hospital EMRs, smaller EMRs for providers, insurance claim records, lab diagnostic data directly from diagnostic lab organizations, and electronic prescription data. We have single hospitals that have used as many as three different EMRs and hospitals where different systems are used for billing, personnel management, supply chain and quality control. We have worked with hospital networks where different EMRs are used at each hospital. We have developed business analytics for provider networks where different offices use different software and some providers are still managing patient records on paper and thus the only available data is from insurance claims. In many cases, the same provider utilizes the hospital EHR for inpatient scheduling and patient records and a different EMR is his regular office.

Given our goal of providing efficient and flexible business analytics, we have designed all of our solutions using best practices for data warehouses including ETL (extract, transform and load), facts, dimensions, and star schema. Basically, the table structures follow the industry standards designed by Kimball[28] and others. One key practice is we always using surrogate keys for joining our facts and dimensions. This separates our keys and joins from the natural identification codes.

In order to support a diverse selection of data sources, we must first create a data model of the business process which is data source agnostic. We did not design the data model based on the data, rather we designed the data model based on the way health care is practiced. We found this information could be viewed from two different perspectives. There is the view of a patient's health which includes their medical and surgical history, allergies, immunizations, chronic diseases, active medications, lab tests and results. Then, there is the view of the operations of health care related to the patients which includes their office and hospital visits and providers. The level of information available is very detailed down to every bed the patient is assigned to, every time they call the doctor's office, every provider that is consulted on

their behalf, every vital sign recorded for them. It is easy to see that the two perspectives are closely coupled.

Since we are attempting a data source agnostic data warehouse, we would like to only source "client unique" data from the client's software systems. Client unique data includes department, employees and other organizational information. Client unique data also includes all events that happen at the client, such as patient encounters and procedures. However, there is a great deal of master data that is not client unique, such as the lists of all providers, diagnoses, medications, immunizations, allergies and procedures. Wherever possible, we want to source this information not from the client software but rather from the most universal data source. Some of the information sets we have utilized include RxNorm for medications[7], National Drug File Reference Terminology (NDF-RT)[8], CVX codes for immunizations [30], Healthcare Common Procedure Coding System (HCPCS) codes for procedures [34], Systematized Nomenclature Of Medicine Clinical Terms (SNOMED)[32], Logical Observation Identifiers Names and Codes (LOINC) codes for lab observations [3], NPI codes for providers [33], and Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) for mental disorders[5]. For diagnoses, the universal standard is the International Statistical Classification of Diseases and Related Health Problems (ICD) codes from the World Health Organization. The ICD codes sets are referred to as ICD-9 (the ninth revision from 1972) and ICD-10 (the 10th revision completed in 1992).[4] There are several additional code sets and terminology such as Current Procedural Terminology (CPT) and FDB (First Databank) for medications and allergies which are protected by copyright. We integrate and convert these terminologies as well; we just do not expose the terminology to the user outside of what was available in their source system.

In addition to all of these standardized code sets, there are several standards for data formats and information exchange. These include Health Level Seven International (HL7)[34], Continuity of Care Record (CCR)[14], and Continuity of Care Document (CCD)[36]. We have attempted to use these standards wherever possible. For example, we have utilized HL7 messages as a way of receiving lab diagnostic results from third party vendors. However, in our experience, these standards are insufficient for us to receive and process complete patient information from EMRs and insurance records. There is significant literature, including [35] and [14], which demonstrate the challenges with relying upon these standards. Therefore, we have utilized custom ETL and custom data formats (closely modeled after HL7) to exchange information. The advantages and disadvantages of each standard is outside of the scope of this paper. However, even if these standards were fully utilized, the challenges and solutions presented in this paper would still apply. None of these information exchange formats completely standardizes the code sets used to document patient events.

We have discussed the data model approach, now we will discuss the metrics and analytics. While a good data warehouse supports flexible ad hoc queries, we also want to enable standard CQMs (clinical quality measures). CMS and other government entities regularly define and update standardized metrics such as ACO Quality Performance Standards[37] and PQRS System Measures List[38] which identifies over 300 quantifiable clinical measures. One of our goals is to be able to support these measurements without requiring repeated software code. CMS defines these measures in terms of the codes and code sets which identified

positive patient results, negative patient results and exceptions. These code lists are provided in spreadsheets and CSV formats.[39] We have automated the download of these documents, some transformations, and the automated population of groupers and metrics based on the CMS definitions.

We could write a book on how to define a data source agnostic health care data model and how to use it to effectively improve health care management. This is well beyond the scope of this paper. In this paper, we will concentrate on key design challenges related to integrating heterogeneous health care data.

3. CHALLENGES

The key challenge discussed in this paper is that of merging data from different data sources, or from the same data source when the data differs either in documentation method or temporally. EMRs and insurance software systems utilize a wide variety of identification codes for patients, encounters, providers, and clinical events. Integrating data involves the very complex task of matching these codes.

Codes such as encounter record numbers and order identification numbers generally describe events which occur entirely in one hospital or provider office and are documented within one EMR or software system. Therefore, it is generally sufficient to store the source identification code in the tables. This id can be used to match information from the data source with corresponding rows in the data warehouse, which are identified with surrogate keys. By including a datasource dimension and a datasource key in every table, we have insured that these source identification codes, in conjunction with the datasource key, are valid natural keys that can be used for lookups. This allows us to correctly match and join source records during ETL.

Clinical codes such as those discussed in Section 2 are much more difficult to correlate. Converting between codes can require unit conversion of the values as well. A bidirectional crosswalk for code matching is often unattainable due to differing level of specificity and grain. In Sections 3.1 and 4, we shall discuss two different examples of this problem.

Another key identification problem is patient identification. Years of research and algorithms have addressed matching patient records, with varying degrees of success. In Sections 3.2 and 5 we will address the challenge of patient identification.

3.1 Code Set Mismatches

For this study we will look at two specific challenges. First we will look at the challenges stemming from migration from ICD-9 to ICD-10. Secondly we will look at the challenges of managing medication codes and documentation.

3.1.1 Diagnoses and ICD-9 to ICD-10

The ICD terminology and diagnosis codes are the universal standard for documenting patient diseases and diagnoses. One of the key challenges throughout the industry is moving from the ICD-9 standard to ICD-10. Currently the US Department of Health and Human Services (HHS) is requiring that this migration be achieved by October 1, 2014. Any ICD-9 codes used in transactions for services or discharges on or after October 1, 2014 will be rejected as non-compliant and the transactions will not be processed.[22]

The ICD-10 code sets for identifying diagnoses and procedures are far more detailed than the ICD-9 code sets. Because there is not a one-to-one mapping between these codes, the conversion from ICD-9 to ICD-10 is a significant challenge. ICD-10 includes 68,000 diagnosis codes, more than five times as many as ICD-9. There is an increased level of detail, plus laterality is now specified. The difference with ICD-10 procedure codes is even more pronounced. There were only 3,000 ICD-9 procedure codes, ICD-10 includes approximately 87,000. In every instance, the ICD-10 codes are more detailed and specific.

Due to the increased level of specificity, it is possible to map ICD-10 to ICD-9 codes (with only one exception, autopsy procedure) but it is not always possible to accurately map ICD-9 codes to ICD-10 codes without additional information.[23][6] This presents a variety of challenges. It is possible to have two data sources for the same patient, where one EMR is documenting the diagnoses using ICD-9 codes and the other is utilizing ICD-10 codes. It is also possible to have new diagnoses for a patient documented with ICD-10 codes while the same patient has historical diagnoses document with ICD-9 codes in the same EMR. Furthermore, it is possible to have reports and queries defined using ICD-9 codes which need to reference data coded with ICD-10 codes. Finally, it is possible to have reports and queries defined using ICD-10 codes which need to reference data coded with ICD-9 codes. Due to our goal of using measurements definitions directly from CMS and other third parties, this difficulty is magnified because we cannot even control how the measurement requirements are coded.

Due to the lack of specificity in ICD-9, many EMRs developed diagnosis tables that provided more information. Such tables are at a lower level of grain, i.e. there may be many diagnoses in the EMR diagnosis table which all have the same ICD-9 code. However, the grain still does not match ICD-10, so a third level of specificity appears when the EMR has its own diagnosis table.

The fact that ICD-10 was completed in 1992 and is still being implemented or awaiting implementation in many health care organization and EMRs demonstrates the enormity of the task of migrating clinical datasets. Work on ICD-11 is already underway and may be completed in 2015. This demonstrates that this challenge is not only significant, but that solutions will be needed again in the future even after ICD-10 compliance is achieved.

3.1.2 Medications

Coding of medications is in many ways more complex than other clinical terminologies. Medications change much more rapidly than diagnoses or procedures (RxNorm releases a database updates monthly). There are many different code sets which are widely used for medications. The same medication can have many different brand names, dosages, quantities and formats. Many medications are similar and medications often share some but not all ingredients. The terminology varies between code sets and EMRs, with a wide range of synonyms available for the same medications. Many clinical measurements require information about inpatient, prescription and over the counter medications, each of which may be documented differently.

Fortunately, RxNorm has already addressed many of these challenges. RxNorm includes terminology codes and synonyms from 11 different data sources. The challenge is that many EMRs do not include RxNorm identifiers. So conversion is required. Medication codes appear at a variety of levels of specificity, so one-to-one conversion is often not possible. This problem is magnified because of the rapid changing nature of medication information and the need to regularly update codes and datasets.

The challenge is further magnified because metrics usually address groups of medications such as statins, beta blockers or antibiotics. The medications belonging to such a group is constantly changing.

3.2 Patient Identification

One of the most important challenges in merging health care information from two or more sources is matching patients. In the United States, the most common and well-known unique identifier for a person is social security number (SSN). However, using a SSN as a patient identifier can cause problems. SSNs may be entered incorrectly. A person may use another's SSN fraudulently. While SSN is still widely used, there has been a significant push to reduce its use.

There are many algorithms and software programs for master patient record management. Such algorithms are difficult for many reasons. Patient's names change, they move, they can even change genders. Twins often share birth dates. No group of identifiers is consistently reliable.

Effective, generic patient matching is also beyond the scope of this paper. However, for our industry experience, we have utilized insurance information as discussed in Section 5.

It is better to have two records for the same patient, than to errantly match patient records when there are in fact two separate patients. Therefore, patient records are often not initially matched. One challenge is that we may realize, via new information or even human intervention, that two patients are the same. If this information is realized after encounters and other events are included in the data warehouse, combining patient records after the fact is a challenging problem. This challenge will also be addressed in Section 5.

4. CODE SET INTEGRATION AND MIGRATION

A data warehouse should be optimized for analytical queries and reporting. As such, we want to perform as many conversions and transformations as possible during ETL time, when populating the database tables.

As a first step, it is necessary we populate the dimensions for each of our code sets. We have populated dimensions for RxNorm, ICD9, ICD10, as well as several EMR specific medication and diagnosis tables. As previously discussed, these dimensions are populated via automated scripts by downloading the codesets from the universal source and applying transformations as needed. We do not use the EMR's data to populate the RxNorm, ICD9 and ICD10 dimensions.

The next step is to correlate the fact tables and events to the dimensions. As we support an arbitrary number of code sets, we have joined the fact with a generic code set bridge table. This table in turn joins to the code sets, creating a snowflake schema. This means that a single fact can be directly linked to multiple codes. For example, a patient diagnosis event may be directly linked to a ICD-9 diagnosis record, a ICD-10 diagnosis record and an internal EMR diagnosis record.

To achieve the code conversion at ETL time, we utilized materialized inference. This leverages our past research on materialized inference and aligning codesets[15][16][17]. When we assert a medication in one code set, we want to assert any

other possible code sets. Thus if we assert PATIENT1 received aspirin, we want to also assert, at ETL time, PATIENT1 received RxNorm medication 1191.

We set up the code conversion rules and crosswalks as ontologies using the Web Ontology Language (OWL)[20]. Fortunately, there has already been work completed to develop RxNorm's library in RDF format [21] and to create OWL ontologies for RxNorm. We programmatically created additional OWL ontology triples for medication crosswalks and we converted the CMS General Equivalency Mapping (GEM) for ICD-9 / ICD-10 conversion[23] to OWL format. We then utilized the Pellet reasoner[42] and RDFKB [15] to perform materialized inference for each code added. We used this knowledge to populate the diagnosis and medication facts. For each new fact row, we link the bridge for the source codeset. Then, we used materialized inference to determine any other code sets to link to this fact, and linked them as well. We did not store any ontology or mapping information in the data warehouse, thus performance of queries and reports was unaffected.

5. PATIENT IDENTIFICATION

Two of our clients were provider networks based on an insurance pool. Thus, the most accurate master patient list was provided by insurance. This led us to the unusual choice of using the combination of group number and policy number as a primary patient identifier. This is unusual because patient insurance changes over time, whereas SSN and birth date are static. However, because providers and hospitals rely on being paid, in our experience, the insurance information is more accurate. Most patients present insurance cards at registration while SSN cards are scarcely ever presented. Furthermore, as the ACA drives more patients in the US to insurance coverage, insurance information will become more prevalent. Therefore, we were able to use insurance information and SSN as the two patient identifiers for our patient identification.

All EMRs we have encountered have their own patient identifiers. These id's are unique for that installation of the EMR. Therefore we created a data source table that lists each installation of each EMR in our system. We created a table called PATIENT_XREF that includes SOURCE_PATIENT_IDENTIFIER, DATASOURCE and PATIENT_KEY. The PATIENT_KEY is the surrogate key and primary key for our DIM_PATIENT patient dimension and a foreign key in PATIENT_XREF. The goal is that each patient only appears once in DIM_PATIENT. However, the patient can appear many times in PATIENT_XREF. The combination of the DATASOURCE and SOURCE_PATIENT_IDENTIFIER is a natural key for PATIENT_XREF. Thus PATIENT_XREF is a translation table to identify patients from the source system identifiers.

One problem is that two rows in DIM_PATIENT may be for the same patient, and we may determine this after the fact. In our scenario, we encountered this situation when a patient had not yet presented their insurance card to a provider. So, when the provider added the current insurance information, we realized the patient was a match, but we had already brought in the patient records. To solve this problem we used a durable surrogate key technique from Kimball[29]. This solution was invented for the problem of changing credit card numbers for the same account. Our problem is essentially equivalent. We have multiple records for the same patient in our dimension and these records are already joined to multiple facts. We introduced a PATIENT_DURABLE_SURROGATE_KEY on to

DIM_PATIENT. For metric purposes, we always group patients by this key. One patient may have multiple records in DIM_PATIENT, but if they are known to represent the same patient, all of these rows will have the same PATIENT_DURABLE_SURROGATE_KEY. This allows metrics to aggregate across the common patient rows. There are two physical rows for the patient, but only one logical row.
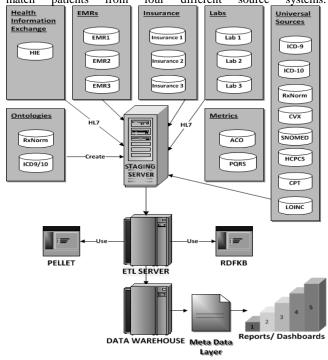
While we were trying to solve the problem using insurance information, this solution is applicable in many ways. Patient records may be matched after the fact using a Master Patient Index tool or even using human intervention. With the push for a universal, or at least national, patient identifier, eventually patient records may be matched because their universal patient ID or community patient ID is assigned. In all of these scenarios, patient matching comes after initial load. Our PATIENT_DURABLE_SURROGATE_KEY allows us to support such matches without reloading the database.

6. RESULTS

Using our code set migration we were able to merge records from two EMRs, one of which used RxNorm and one which did not. We were also able to match lab component results identified with LOINC codes to those identified with proprietary codes. We were successfully able to implement ACO and PQRS measures, using code lists from CMS in conjunction with source data from multiple EMRs.

Finally, we were able to load a master patient list from the insurance company who supported a provider network. Using our PATIENT_XREF table and PATIENT_DURABLE_SURROGATE_KEY we were able to match patients from four different source systems.

Figure 1. Business Intelligence Data Flow

Figure 1 shows our solution graphically. Data is received into the staging server through "listening for HL7 messages" (for labs ad HIE information), through receiving data files via FTP (the preferred method of insurance companies), through automated

download scripts (used for universal sources and metrics definitions) and through custom ETL queries (used for EMRs). We created the ontologies programmatically. We then used SQL and Java code plus the ontologies, Pellet and RDFKB to populate the data warehouse. All patient matching and code conversion was done during this ETL process. There were instances where user information was required, but we were able to provide such information (for example, LOINC conversion codes and unit conversion rules) in spreadsheets and ontologies which were developed by source system experts. This information becomes a source for the ETL, so the ETL operation is still automated and unsupervised. All of the metrics and data was stored in the data warehouse and available for the metadata layer and analytics without further transformation.

We also developed rules to detect possible mismatches or erroneous data. For example, we used birthdate as a validation test to make certain we did not match different patients. Such data was moved to an exception table and sent to a expert user for manual evaluation. Our experience is that all such errors resulted from invalid source data (typos, etc). The exception table provides information so that the source systems can be corrected.

7. FUTURE WORK

While it is possible to map ICD-10 codes to ICD-9 codes, it is often not possibly to accurately map ICD-9 codes to the more specific ICD-10 codes without more information. For future work, we would like to use probabilistic reasoning in conjunction with textual analysis to attempt this matching. For example, ICD-10 codes often specify laterality so we could look for the words "left" or "right" to improve code matching. We would also like to develop algorithms for predictive analysis. This would enable us to use patterns and historical data to predict the likelihood of events such as payment, patient compliance with treatment, and operational throughput. Also, as standards continue to evolve, we would like to work towards enabling information exchange directly through standardized documents such as CCDs. This would reduce the amount of custom ETL needed to support an EMR.

8. REFERENCES

[1] 111th Congress of the United States. 2010. H.R 3590: The Patient Protection and Affordable Care Act.

[2] Orzag, P., Emanuel, E. 2010. Health Care Reform and Cost Control. In *The New England Journal of Medicine*, 601-603.

[3] McDonald. C.J., Huff, M., Suico, J., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G. and Hook, J. 2003. LOINC, a universal standard for identifying laboratory observations: a 5-year update. In *Clinical Chemistry*, 49, 4, (April 2003), 624-633.

[4] World Health Organization. 2004. ICD-10: International Statistical Classification of Diseases and Related Health Problems.

[5] American Psychiatric Association. 2010. Diagnosis and Statistical Manual of Mental Disorders. 4th Edition.

[6] Schulz, S., Zaiss, A., Brunner, R. , Spinner, D., and Klar, R. 1998. Conversion problems concerning automated mapping from ICD-10 to ICD-9. In *Methods of Information in Medicine*, 37, 3, (1998), 254-259.

[7] Liu, S., Ma, S., Moore, R., Ganesan, V. and Nelson, S. 2005. RxNorm: Prescription for Electronic Drug Information Exchange. In *IT Pro*, 7, 5, (October 2005), 17-23.

[8] Fung, K.W., McDonald, C, and Bray, B. 2008. RxTerms – a drug interface terminology derived from RxNorm. In *Proceedings of AMIA Annual Symposium*, 227-231.

[9] Hernandez, P., Podchiyska, T., Weber, S., Ferris, T and Lowe, H. 2009. Automated Mapping of Pharmacy Orders from Two Electronic Health Record Systems to RxNorm within the STRIDE Clinical Data Warehouse. In *Proceedings of AMIA Annual Symposium*, 244-248.

[10] Podchiysk, T., Hernandez, P., Ferris, T., Weber, S. and Lowe, H. 2010. Managing Medical Vocabulary Updates in a Clinical Data Warehouse: An RxNorm Case Study. . In *Proceedings of AMIA Annual Symposium*, 477-481.

[11] Marshall. M.S., Boyce, R., Deus, H., Zhao, J., Willighagen, E., Samwald, M., Pichler, E., Hajagos, J., Prudhommeaux, E, and Stephens, S. 2012. Emerging practices for mapping and linking life sciences data using RDF — A case series. In *Journal of Web Semantics,* 14 (July 2012), 2-13.

[12] Stange, K., Nutting, P., Miller, W., Jan, C and Gill, J. 2010. Defining and measuring the patient-centered medical home. In *Journal of General Internal Medicine*, 25, 6, 601-612.

[13] Moreno, L., Peikes, D, and Krilla, A. 2010. Necessary But Not Sufficient: The HITECH Act and Health Information Technology's Potential to Build Medical Homes. Agency for Healthcare Research and Quality.

[14] Ferranti, J., Musser, R. C., Kawamoto, K. and Hammond, W. 2006. The clinical document architecture and the continuity of care record: a critical analysis. In *Journal of the American Medical Informatics Association*, 13, 3 (2006), 245-252.

[15] McGlothlin, J. and Khan, L. 2009. RDFKB: efficient support for RDF inference queries and knowledge management. In *Proceedings of IDEAS*. 259-266.

[16] McGlothlin, J. and Khan, L. 2010. Materializing and Persisting Inferred and Uncertain Knowledge in RDF Datasets. In *Proceedings of AAAI*.

[17] McGlothlin, J. and Khan, L. 2010. A Semantic Web Repository for Managing and Querying Aligned Knowledge. In *Proceedings of ISWC*.

[18] Berwick, D. 2011. Launching accountable care organizations—the proposed rule for the Medicare Shared Savings Program. In *New England J. of Medicine*, 364, 16.

[19] Singer, S and Shortell, S. 2011. Implementing accountable care organizations. In *The Journal of the AMA*. 306, 7.

[20] Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P. and Stein, L. 2004. OWL web ontology language reference.

[21] A script to convert RxNorm to RDF. 2013. https://github.com/jhajagos/RxNormRDF. Accessed May 2013.

[22] CMS. 2013. ICD-10 Transition Basics.

[23] Centers for Medicare & Medicaid Services. 2011. General Equivalence Mappings. http://www.cms.gov/Medicare/Coding/ICD10/downloads/GEMs-CrosswalksBasicFAQ.pdf. Accessed March 2013.

[24] Blumenthal, D. and Tavenner, M. 2010. The "meaningful use" regulation for electronic health records. In *New England Journal of Medicine*, 363, 6, 501-504.

[25] Anumula, N and Sanelli, P. 2011. Physician quality reporting system. In *American Journal of Neuroradiology*, 32, 11, 2000-2001.

[26] Brown, T., Werling, K., Walker, B., Burgdorfer, R. and Shields, J. Current trends in hospital mergers and acquisitions. In *Healthcare financial management: journal*, 66, 3, 114.

[27] Vest, J. and Gamm, L. 2010. Health information exchange: persistent challenges and new strategies. In *Journal of the American Medical Informatics Association*, 17, 3, 288-294.

[28] Kimball, R and Ross, M. 2002. The data warehouse toolkit: the complete guide to dimensional modeling.

[29] The Kimball Group. 2012. Design Tip #147 Durable "Super-Natural" Keys. http://www.kimballgroup.com/2012/07/10/design-tip-147-durable-super-natural-keys/. Accessed December 2012.

[30] Centers for Disease Control and Prevention. 2013. IIS: HL7 Standard Code Set CVX -- Vaccines Administered.

[31] Centers for Medicare and Medicaid Services. 2012. Healthcare Common Procedure Coding System (HCPCS). http://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html?redirect=/medhcpcsgeninfo. Accessed May 2013.

[32] Spackman, K. and Campbell, K. and Cote, R. 1997. SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium,* 640-644.

[33] National Plan and Provider Enumeration System. 2013. NPI Registry. https://npiregistry.cms.hhs.gov/NPPESRegistry.

[34] Dolin, R., Alschuler, L., Boyer, S., Beebe, C., Behlen, F., Biron, P. and Shvo, A. 2006. HL7 clinical document architecture, release 2. In *Journal of the American Medical Informatics Association*, 13, 1, 30-39.

[35] Smith, B and Cuesters, W. 2006. HL7 RIM: an incoherent standard. In *Ubiquity: Technologies for Better Health in Aging Societies*. 133-138.

[36] Health Level Seven International. 2013. HL7/ASTM Implementation Guide for CDA Release 2 -Continuity of Care Document Release 1.

[37] Centers for Medicare & Medicaid Services. 2012. ACO 2013 Program Analysis - Quality Performance Standards Narrative Measure Specifications.

[38] Centers for Medicare & Medicaid Services. 2012. 2013 Physician Quality Reporting System (PQRS) Measures List.

[39] Centers for Medicare & Medicaid Services. 2012. ACO GPRO Supporting Documents. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/ACO-GPRO-Supporting-Docs-2012.zip.

[40] World Health Organization. 2007. Patient Identification. In *Patient Safety Solutions*, 1, 2.

[41] Versel, Neil. 2013. National Patient ID System: Debate Stoked. In *Information Week HealthCare* (March 29, 2013).

[42] Parsia, B. and Sirin, E. 2004. Pellet: An owl dl reasoner. In *Proceedings of ISWC*, 18.