206-478-8227
www.healthdataconsulting.com

Data Aggregation by Health Condition:

The Ontological Model

Joseph C Nichols MD
Principal

A Health Data Consulting White Paper

2/8/2017

TABLE OF CONTENTS

Executive Summary

The shift in policy and payment to a "value-based purchasing" approach is changing the way healthcare is evaluated and potentially paid. There is a growing focus on the nature of the patient condition and the degree to which that condition may be improved. The volume of services is considered a cost to be dealt with, rather than a focus for policy and payment. In this new environment, the ability to define and aggregate data for the analysis of the precise nature, risk, severity and complexity of the patient state is increasingly important. There are substantial challenges in the definition and aggregation of data to meaningful categories of health conditions. Ontologies provide an effective solution to define and maintain clinical categories so they accurately and consistently represent the realities of the patient health state as patients move through the healthcare system.

Data Aggregation: Background

Aggregation of data is the key to all healthcare analytics. While structured (coded) data is required to accurately and consistently capture the facts of any instance of an event or encounter, the compilation of data across many events and encounters requires that these data points are aggregated in a way that supports meaningful information. Historically, aggregation or categorization of data has generally been assumed by the end user to be accurate and meaningful. Many information consumers believe that the meanings of categories such as "Myocardial Infarction" or "Hip Fractures" are commonly understood and comparable across enterprises. Those intimately involved in the analysis of large cross-enterprise healthcare data sets however are fully aware that nothing could be further from the truth. Without consistent, standardized, clearly defined data aggregation, healthcare information is unreliable and may lead to misdirected actions.

CHANGING FOCUS FROM SERVICE TO CONDITION

Healthcare is rapidly evolving to a system focused on the nature of the patient's health condition and the degree to which healthcare services maintain or improve that condition. As payment has been traditionally focused on healthcare services, data and analytics have likewise been primarily service-based. With the advent of a "value-based" purchasing approach to healthcare policy and payment, there is a renewed interest in the patient's health condition as a key focal point for data analytics. This topic was covered in some detail in a previous HDC white paper.[1] Unfortunately historical data analysis of claims has been so "service" driven that there has been little attention paid to the quality of diagnosis-based data capture or diagnosis-based analysis. The aggregation of data to defined clinical concepts that represent the nature of patient's condition is complex and requires an understanding of the clinical parameters of diseases as well as how those disease concepts are represented in codified data.

[1] http://healthdataconsulting.com/_wp/wp-content/uploads/2016/10/ValueBasedPurchasing_Risk.pdf

THE ROLE OF THE INTERNATIONAL CLASSIFICATION OF DISEASES

The International Classification of Diseases (ICD) is the only nationally mandated and universally standardized definition of the patient health condition for healthcare transactional data in the United States. While there are other defined "standards", only ICD codes are used across all healthcare enterprises. ICD-10-CM updated this classification to include 71,831 diagnostic codes to represent the nature of the healthcare condition in a much more robust way. Used properly, these codes have a much greater opportunity to capture significant differences in risk, severity and complexity of the patient's health state. The operative word, however, is "opportunity". The coding classification, no matter how robust, is only as good as the data captured, preferably by someone with clinical knowledge at the point of care. If observation or documentation of the facts is incomplete or inaccurate, the coding system will only reflect what is observed and documented. While incentives for improving documentation and coding are becoming more common place, there is still a long way to go to improve the level of completeness, detail and accuracy of data around patient events and encounters. Clinicians have traditionally not seen the value in capturing high quality diagnostic data since payment has been focused on the service rather than the patient's underlying health state. Some of the challenges with data quality have been discussed in a three part series in HIMSS business news.[2]

Data Aggregation: Challenges

Even with the most accurate, detailed and comprehensive coded data however; accurate aggregation of this data is required to arrive at reliable information that includes all of parameters that differentiate the risk, severity and complexity of health conditions. There are wide ranges in the nature of various disease classifications such as diabetes, malignant neoplasm, cardiac ischemic disorders, intracranial injuries, fractures and a host of other conditions. Depending on these parameters the clinical and financial risk for patients can vary greatly within the same general clinical domain. Comparing outcomes and cost requires aggregation of similar conditions at a clearly defined and reasonably homogeneous level. Adjusting for variations in risk, severity and complexity is essential to properly understanding and managing the disease burden of health conditions.

CATEGORICAL DEFINITIONS

Clear, consistent and shared definitions of any classification are essential to meaningful, comparable information that can lead to reliable interpretation. Often information users assume that certain categories of health conditions are globally defined and standardized. When information about categories of conditions like "cardiac ischemic disorders" is presented, few users critically question how this category was defined. Did it include myocardial infarction? Are acute or chronic conditions included? Does the category include angina or other symptoms of ischemia? Are patients with a family or personal history of cardiac ischemic disease included?

---

[2] http://www.himss.org/news/value-quality-healthcare-data
http://www.himss.org/news/data-quality-understanding-past-improve-future
http://www.himss.org/news/data-quality-strategies-improving-healthcare-data

Does the analysis include asymptomatic patients with findings of occlusion? The answer to these questions can make a profound difference in the analytic values measured. Different definitions of what appears to be the same category in different databases result in non-comparable analytic results. It is surprising to see that many "standard" health care reports on different categories of disease do not have clear definitions of how the categories were defined. It is common for organizations to have rules, edits and coverage decision logic based on a set of codes where the definition and purpose of the code set is not clearly known by business stakeholders.

Defining the Category

It is essential for any analysis that every category is clearly defined so that users applying reasonable logical deduction can arrive at similar conclusions based on similar data. To get at a clear definition, the detailed nature of the category must be stated. The definition is not complete until it is clear what concepts are meant to be included or excluded.

Defining a Category

The following represents a sample definition of a category. The purpose of this definition is to attempt as clearly as possible to identify what coded data should be included or excluded. The definition could vary substantially, depending on the entity creating the definition. It is important that each entity understands the intent of the definition so that appropriated comparisons of similar disease categories can be made.

- ***Example Category Definition: "Burns"***

Burns refer to the clinical disorder of tissue damage that may be caused by some form of thermal energy.

- *Includes:*
  - *Heat related burns of all degrees*
  - *Electrical burns*
- *Excludes:*
  - *Friction burns*
  - *Chemical burns*
  - *Cold related burns or tissue damage*
  - *Sunburns*
  - *Radiation burns*

The above definition is simply an example of how the category "Burns" could be defined. Other entities might wish to include some of the excluded concepts or excluded some of the included concepts in their definition of "Burns". It is critical that any comparisons or analyses of "Burns", in terms of cost, outcomes, or any other analytic focus, are done with a clear understanding of what is or is not included in this category based on the category definition. The coded data may be markedly different depending on the intent of the category definition. Other categories such as "Diabetes", "Brain injury", "Depression", "Drug Induced Conditions" and most other areas of categorization may have significantly

different code sets and thus data aggregation depending on the definition that meets the intent of the entity creating the category. There are many different categorizations, but there is no right, wrong or universally accepted category definition.

FINDING THE RIGHT CODE SET

The first step in the process of data aggregation is the definition as illustrated above. This example definition lays out the map for which concepts should be included or excluded from the code set used for data aggregation. Without this clear definition, selecting the appropriate codes becomes an arbitrary effort that will result in non-comparable interpretations. Clarity of definition is essential to validate that the logic for aggregation is including data that should be included and excluding data that should not be included.

Given a clear definition, there is still a substantial challenge in finding all of the codes that should be included or excluded based on the defined concepts. The following are examples of the challenge in searching for codes that represent basic clinical concepts.

CONCEPT: "HIP FRACTURES"

A hip fracture is generally described as any fracture of the proximal or upper third of the femur. Fractures of the acetabulum or other part of the pelvis are generally not included in this classification but rather in a classification of "pelvic fractures"

In the process of searching for the right ICD-10 codes to include in data aggregation to the category "Hip Fracture", it might seem appropriate to query the codes for the terms "hip" + "fracture", but that query would only return 38 codes that are very uncommon types of hip fractures. There are actually 1,260 codes that should be included in this set. Figure 1 illustrates an example of different terms that are used in ICD-10 for various types of hip fractures based on this traditional definition.

M84359A - Stress fracture, hip, unspecified, initial encounter for fracture
M84559A - Pathological fracture in neoplastic disease, hip, unspecified, initial encounter for fracture
M84659A - Pathological fracture in other disease, hip, unspecified, initial encounter for fracture
M9701XA - Periprosthetic fracture around internal prosthetic right hip joint, initial encounter
S72001A - Fracture of unspecified part of neck of right femur, initial encounter for closed fracture
S72021A - Displaced fracture of epiphysis (separation) (upper) of right femur, initial encounter for closed fracture
S72031A - Displaced midcervical fracture of right femur, initial encounter for closed fracture
S72041A - Displaced fracture of base of neck of right femur, initial encounter for closed fracture
S72051A - Unspecified fracture of head of right femur, initial encounter for closed fracture
S72091A - Other fracture of head and neck of right femur, initial encounter for closed fracture
S72101A - Unspecified trochanteric fracture of right femur, initial encounter for closed fracture
S72111A - Displaced fracture of greater trochanter of right femur, initial encounter for closed fracture
S72121A - Displaced fracture of lesser trochanter of right femur, initial encounter for closed fracture
S72141A - Displaced intertrochanteric fracture of right femur, initial encounter for closed fracture
S7221XA - Displaced subtrochanteric fracture of right femur, initial encounter for closed fracture
S79001A - Unspecified physeal fracture of upper end of right femur, initial encounter for closed fracture

Fig. 1

As can be seen in this sampling of the 1,260 hip fracture codes, the description of the condition may not be what is expected by someone not familiar with the various descriptions of types of hip fractures. Very rarely is the word "hip" used in these descriptions, but they are still considered "hip fractures".

CONCEPT: "DOWN SYNDROME"

Down syndrome is a condition associated with a chromosomal abnormality that results in an extra chromosome 21 (trisomy).

Figure 2 illustrates that a search for the term "Down syndrome would leave out codes for "Trisomy 21" which by definition is "Down syndrome".

Q909 - Down syndrome, unspecified
Q900 - Trisomy 21, nonmosaicism (meiotic nondisjunction)
Q901 - Trisomy 21, mosaicism (mitotic nondisjunction)
Q902 - Trisomy 21, translocation

Fig. 2

The person identifying the codes for this code set would need to know that Trisomy 21 is Down syndrome, or results from analysis could leave out a substantial amount of data that should be included, based on the intent of the category. Payment policies might inappropriately pay or deny claims, depending on the intent of the policy, if codes are excluded or included inappropriately.

CONCEPT: "RENAL FAILURE"

Renal failure is a condition where the renal system is no longer able to maintain the critical function of waste excretion through the urine.

While there are various levels of disorders of renal function certain descriptions are generally included as "renal failure". Figure 3 illustrates some of the terminology used in the code set that would represent "renal failure"

N170 - Acute kidney failure with tubular necrosis
P960 - Congenital renal failure
N184 - Chronic kidney disease, stage 4 (severe)
N185 - Chronic kidney disease, stage 5
N186 - End stage renal disease

Fig. 3

As shown, sometimes the term "kidney failure" is used rather than "renal failure". Chronic kidney disease at stage 4 or stage 5 is generally considered renal failure, where stage 1, 2 and 3 are not included as renal failure. End stage renal disease is also renal failure. All 20 codes associated with these types of descriptions would need to be included in order to accurately report on "Renal Failure".

CONCEPT: "POSTPARTUM"

Postpartum conditions are related to delivery and occur in the timeframe after birth, generally considered within the first 6 weeks after delivery.

As illustrated in Figure 4, different terms are used to refer to postpartum type conditions. There are 1,260 codes that would meet this definition.

O712 - Postpartum inversion of uterus
O8611 - Cervicitis following delivery
Z379 - Outcome of delivery, unspecified
O9081 - Anemia of the puerperium
Z3800 - Single liveborn infant, delivered vaginally

Fig. 4

CONCEPT: "DRUG-INDUCED"

Drug-induced conditions refers to conditions that are attributable to the use of some pharmaceutical substance.

Identifying the codes that should be categorized to the concept of "Drug-induced" is especially complex. Based on this definition, there are 3,104 codes that could be included.

D521 - Drug-induced folate deficiency anemia
R330 - Drug induced retention of urine
M342 - Systemic sclerosis induced by drug and chemical
D61810 - Antineoplastic chemotherapy induced pancytopenia
D7582 - Heparin induced thrombocytopenia (HIT)
E0900 - Drug or chemical induced diabetes mellitus with hyperosmolarity without nonketotic hyperglycemic-hyperosmolar coma (NKHHC)
F1014 - Alcohol abuse with alcohol-induced mood disorder
F1114 - Opioid abuse with opioid-induced mood disorder
F1314 - Sedative, hypnotic or anxiolytic abuse with sedative, hypnotic or anxiolytic-induced mood disorder
F1414 - Cocaine abuse with cocaine-induced mood disorder
F1514 - Other stimulant abuse with stimulant-induced mood disorder
F1614 - Hallucinogen abuse with hallucinogen-induced mood disorder
F1814 - Inhalant abuse with inhalant-induced mood disorder
F1914 - Other psychoactive substance abuse with psychoactive substance-induced mood disorder
G2111 - Neuroleptic induced parkinsonism
H4060X0 - Glaucoma secondary to drugs, unspecified eye, stage unspecified
I952 - Hypotension due to drugs
L432 - Lichenoid drug reaction
L560 - Drug phototoxic response
L561 - Drug photoallergic response
N140 - Analgesic nephropathy
O2940 - Spinal and epidural anesthesia induced headache during pregnancy, unspecified trimester
O771 - Fetal stress in labor or delivery due to drug administration
P962 - Withdrawal symptoms from therapeutic use of drugs in newborn
T360X1A - Poisoning by penicillins, accidental (unintentional), initial encounter
T360X5A - Adverse effect of penicillins, initial encounter

Fig. 5

As shown, if the intent was to identify all data related to drug-induced conditions, these codes would be needed as part of the data set. There is no easy way to identify all codes without recognizing how these codes are described in the data.

THE CHALLENGE WITH EXISTING HIERARCHICAL CATEGORIES

There are many models for data aggregation that have been used for years to define categories of diseases. The Agency for Healthcare Research and Quality has used the Clinical Classification System for a wide range of analysis[3]. CMS has defined the Hierarchical Condition Categories (CMS HCCs) as a method of aggregating conditions to defined risk groups for the purpose of Medicare Advantage risk adjustment[4]. More recently under the Accountable Care Act and MACRA legislation a different and expanded version of Hierarchical Condition Categories (HHS HCCS) was introduced. ICD-10-CM categorizes codes in its tabular structure[5]. A large variety of government, academic, specialty and commercial entities also have categorization models around diseases. While it might seem that analysts could just select one of the categories to meet their healthcare business needs, which category or group of categories would they select? How will any category selected compared with similar categories used by other entities?

There are a number of challenges with the adoption of any of these categories:

- There is no universally standardized set of categories.
- For most of these categories there is no clear definition of the category or the intent of what concepts are to be included or excluded.
- Many of these category schemes are mutually exclusive so that a code must belong to one and only one category. Most of the ICD-10 codes are "combination codes"[6] so that many codes may be related to many distinct concepts. By putting a code in an exclusive category, data aggregated to that category may not be counted in other categories. If the code for "diabetic retinopathy" is counted under the category for "diabetes", it may or may not be counted under "eye disorders".
- The definition of some of these categories may not be clinically homogeneous. In other words, certain concepts are widely different in terms of the nature of the condition. The CMS HCC category, HCC12 (Breast, Prostate and Other Cancers and Tumors) includes a wide range of conditions representing a broad set of neoplasms that vary greatly in risk, severity, tissue type, location and other key factors.
- Most of these categories are at a level of granularity that makes it difficult to get at actionable information. Actions that are based on data generally require the ability to drill down to much greater detail.
- In the ICD-10 tabular structure, codes may not be found where there are expected to be. For example, when looking for codes related to pneumonia, the tabular category in ICD-10

[3] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
[4] https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html
[5] https://www.cms.gov/Medicare/Coding/ICD10/2017-ICD-10-CM-and-GEMs.html
[6] The term "combination codes", refers to the fact that a number of ICD codes reference multiple different concepts within a single code description.

includes 38 codes under “Influenza and Pneumonia”. Further analysis, however, reveals that there are 42 other codes related to some form of pneumonia in 14 other categories. For example, streptoccocal pneumonia is under the category of “streptoccocal diseases” rather than under the “influenza and pneumonia” category.

- Beyond these challenges, static categories can be difficult to modify and maintain over time as new codes and new medical concepts evolve, particularly as the number of categories and uses for these categories expands.
- Defining these categories requires a great deal of time, effort and commitment from clinical, coding and technology resources.

The Ontological Solution

An ontological model represents a mapping of concepts to other expressions[7] or concepts through defined relationships. Ontologies are different than hierarchical taxonomies in that they do not impose a rigid structure or categorically exclusive assignment to any expression or concept. For example, if the expression “Streptococcal Pneumonia” is categorized in a traditional hierarchical taxonomy, its representation in a specific category may be different depending on the focus of the taxonomy.

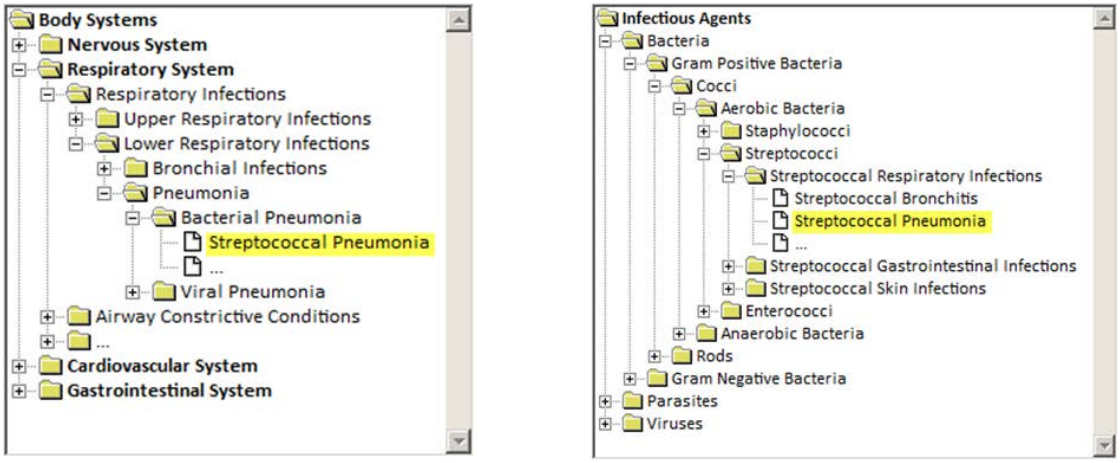

Fig. 6

As can be seen in this example, the classification would be different depending on the focus of a taxonomy based on a body systems structure as compared to a taxonomy based on infectious agents.

Ontologies do not put expressions or concepts into a hierarchical “bucket”. The ontological approach attaches metadata tags to an expression or concept so the expression can be included

[7] An expression in this context refers to any term, combination of terms or code that expresses a simple or complex thought

or excluded in many different types of categories based on these indexed (metadata) tags as seen in Figure 7.

Streptococcal Pneumonia	
Relationship	**Ontological Concept**
Is a type of	Pneumonia
Is a type of	Infection
Is a condition of	Pulmonary system
Is a condition of	Lung
Is caused by	Streptococcus
Is a	Communicable Disease

Fig. 7

In the ontological model, any expression can have an unlimited number of metadata tags through an unlimited number of relationships. This allows for a robust way to aggregate concepts "on-the-fly" into an intended category for an intended purpose, while maintaining consistent definitions. For example, if there is intent to create a category of pulmonary conditions other than infectious conditions, the metadata tag "Pulmonary system" can be selected as an included concept and the metadata tag for "Infection" can be selected as an excluded concept. Any combination of included concepts and/or excluded concepts can be used to create a defined category for any purpose. This allows for a very robust and rapid method for categorization to meet any purpose while maintaining a consistent thread in the process around the core metadata tags that define the category. In this way, research and validation can be focused on mapping codes to concepts and reusing that effort to create categories to meet a specific business or analytic intent.

THE MDMETA ONTOLOGY

Health Data Consulting has partnered with a Baltimore based technology company to create MDMeta LLC. The MDMeta product currently applies metadata tags to all ICD-9 and ICD-10 codes. These tags are applied based on clinical research, the official description of the code, and other official references that identify concepts that are included in the definition of the coded expression. In some instances there may be only a limited number of tags and in other instances a large list of tags is needed to capture all relevant concepts. The following two examples show how tags are applied to codes.

- S52531B - Colles' fracture of right radius, initial encounter for open fracture type I or II

  Metadata tags:

  - Fracture (*This represents an injury that is a fracture of bone*)
  - Injury (*This is a health condition resulting from trauma*)
  - Bone (*This condition is associated with bone type tissue*)
  - Forearm *(This fracture involves a bone of the forearm)*
  - Radius (*This is a condition of the radius bone*)
  - Wrist (*This condition involves the region of the wrist*)
  - Unilateral (*This condition involves only one side of the body*)
  - Right (*This condition involves the right side of the body*)
  - Displaced (*By definition Colles' type fractures are displaced*)
  - Extra-articular (*By definition Colles' type fractures are outside of the joint*)
  - Distal (*By definition Colles' type fractures involve the lower end of the radius*)
  - Colles' fracture (*This is a named type of fracture called a 'Colles' fracture')*
  - Open fracture (*This represents a fracture that is exposed, external to the body, through a wound*)
  - Gustilo type 1 or 2 (*This type of open fracture is of a less severe open fracture classification*)
  - Initial encounter (*This is the first encounter for active treatment of injury by the treating clinician*)

  Each of these tags represents concepts that explicitly or implicitly are a part of the overall coded expression. While the concepts "unilateral", "forearm", "Gustilo", "distal", "displaced", "wrist" and "extra-articular" are not explicitly stated in the description, clinically they are relevant to this coded condition. If an analyst wished to evaluate the impact on costs or outcomes of all fractures of the distal radius that were outside of the joint (extra-articular) compared to inside the joint (intra-articular), it is a simple matter of selecting the appropriate metadata tags to create the appropriate analytic category. The mapping to these distinct tags has already been researched and completed.

- J15212 - Pneumonia due to Methicillin resistant Staphylococcus aureus

  Metadata tags:

  - Pulmonary (*This condition involves the pulmonary system*)
  - Lung (*The condition is anatomically associated with the lung*)
  - Lower respiratory (*This condition would be classified as a lower vs. upper respiratory type*)
  - Infection (*This is a condition caused by an infectious agent*)
  - Pneumonia (*This condition is defined as a pneumonia*)
  - Bacteria (*This is a bacterial type infection*)
  - Gram positive [bacteria] (*This infection is caused by a gram positive type bacteria*)
  - Staphylococcus aureus (*The specific bacteria involved is staph aureus*)
  - Antibiotics [penicillins] (*This condition references antibiotic use of the penicillin type*)
  - Methicillin-resistant (*The infection is resistant to the use of methicillin*)
  - CMS_HCC114 (*This condition has been classified to HCC114 in the CMS Hierarchical Condition Category scheme for risk adjustment*)

  Each of these distinct metadata tags can be used to include or exclude this and similar codes for aggregation to meet a specific purpose. For example, if an analyst wished to look at data related to claims for hospital admission related to all "Methicillin-resistant" "Lower respiratory" "Infections", simply selecting the appropriate tags will define the appropriate categorization of codes needed for the analysis. If a business analyst wished to know which codes were associated with a CMS HCC, that metadata tag selection will return all relevant codes. Codes can be mapped in the ontology to any metadata tag that can be defined at any level of granularity.

CURRENT MDMETA ONTOLOGY SCOPE

The development of the MDMeta ontology product has been an ongoing effort of research and mapping over many years. This ontological product exists currently in a SQL database, but is designed in a way that is technology independent. Access to the ontology is currently provided through a cloud-based database delivered through a state of the art universally accessible API (Application Programming Interface). Currently there are 9,500 distinct concept-based Metadata tags with over 900,000 code-to-tag associations. The definition of each MDMeta tag concept is available in the database. Additionally there are approximately 3,000 "alias" terms for the defined metadata tags. The ontology includes all ICD-9 and ICD-10 diagnostic codes as well as all current updates.

Ontologies by their nature are an ongoing effort. Thousands of hours have been spent researching coding scenarios and clinical concepts to continuously improve the scope and quality of Metadata tag mappings. The design of the technology allows for expansion to include all code and expression types and all metadata relationship types. The ongoing expansion and improvement of the ontology is an MDMeta commitment. Ongoing updates are published in realtime in the cloud-based database.

A proof of concept demonstration is available by contacting Joe Nichols MD at jnichols@mdmeta.com