Understanding the Challenges with Medical Data Segmentation for Privacy

Ellick M. Chan, Peifung E. Lam, and John C. Mitchell
Stanford University {emchan, pflam, mitchell}@cs.stanford.edu

ABSTRACT

Electronic Health Records (EHRs) are perceived as a path to significant improvement in healthcare, and patient privacy is an important consideration in the adoption of EHRs. Medical record segmentation is a technique to provide privacy and protect against discrimination for certain medical conditions such as STDs, substance abuse and mental health, by sequestering or redacting certain medical codes from a patient's record.

We present an initial study that describes an approach for segmenting sensitive medical codes to protect patient privacy and to comply with privacy laws. Firstly, we describe segmentation strategies for sensitive codes, and explore the link between medical concepts using sources of medical knowledge. Secondly, we mine medical knowledge sources for correlations between medical concepts. Thirdly, we describe an approach that a privacy attacker may use to infer redacted codes based off second order knowledge. More specifically, the attacker could use the presence of multiple related concepts to strengthen the attack. Finally, we evaluate defensive approaches against techniques that an adversary may use to infer the segmented condition.

1. INTRODUCTION

Electronic Health Records (EHRs) are perceived by some government agencies and law makers as a path to significant improvement in healthcare, and patient privacy is an important consideration in the adoption of EHRs [40, 39]. Health records are commonly used by medical professionals for the purpose of diagnosis, treatment, coordination of care, and billing. However, some records contain *sensitive information* [15] such as sexually transmitted diseases (STDs), mental health or substance abuse information that may be embarrassing or used to discriminate against the patient in ways that are prohibited by certain state or federal laws (Section 1.1). Many of these laws were originally enacted to encourage patients with serious but embarrassing conditions to come forward and seek medical treatment while retaining some level of privacy. From a medical perspective, studies have shown that some doctors can be susceptible to unconscious bias if this sensitive information is present [16, 18]. Therefore some privacy advocates argue that sensitive information should be hidden by default, and that patients should have full control over what information is released[29]; but other reports have suggested that patients could be ill-equipped to decide what information is medically relevant[15].

Healthcare reform through coordinated care will require increased coordination of activities and continuity of information via electronic health records (EHRs). However, patient privacy concerns codified in state and federal privacy laws governing sensitive medical conditions such as mental health, HIV and substance abuse limit the information that various parties such as health information exchanges (HIE) can handle without explicit patient consent.

In this paper, we study the delicate balance between privacy and medical relevancy and the extent to which sensitive information can be separated by *segmentation* – defined by the Office of the National Coordinator for Health IT (ONC) as *"the process of sequestering from capture, access or view certain data elements that are perceived by a legal entity, institution, organization, or individual as being undesirable to share"* [15] – to a special portion of the health record in order to protect the patient's privacy and to comply with the law. There currently are some ongoing efforts to implement segmentation [25, 35].

While a näive approach, which we call first-order segmentation, may simply isolate these sensitive codes to protect the patient from discrimination and stigma, we show that such approaches may not be effective against an advanced privacy adversary. Our threat model considers an adversary who has access to segmented EHRs and access to medical knowledge and literature, but does not have the capability to circumvent other security measures.

1.1 Privacy laws and HIE

Health information exchanges (HIE) [41] foster the interchange of medical information between organizations to support the use cases mentioned earlier. Often times, shared records need to cross state and jurisdictional boundaries, and when they do, the exchange must comply with potentially differing laws and expectations for sensitive information. Laws typically require consent for sensitive information. There are currently two predominant models for handling consent in an HIE: $opt-in$ and $opt-out$. In an opt-in model, patients must consent to have their information shared, otherwise it is not available by default. In the opt-out model, patient data is shared by default. If a patient wishes to be excluded from the HIE, they must specifically request to do so.

In order to serve patients that have chosen to withhold sensitive information, or have not expressed consent in an opt-in model, segmentation has been proposed by ONC [15]. Ideally, such an approach would be able to perform two functions: (i) identify sensitive information, and (ii) selectively remove or isolate sensitive information in a medically-relevant manner. We term such a system as a $predicate-reducer$ where the $predicate$ identifies the presence of sensitive information according to some privacy policy, and the $reducer$ separates the information without otherwise changing the meaning of the health record with respect to a particular task such as diagnosis, treatment, or research.

A näive implementation of a predicate-reducer might simply take a health record and isolate all ICD-9/ICD-10, LOINC, HL7, SnoMed [34] and other similar codes corresponding to sensitive information. Alternatively, an even simpler reducer may just isolate certain sections of the health record that are likely to contain sensitive information, e.g. psychotherapy notes. Such reducers may leave gaping holes in the record because they fail to understand the complex semantics and relationships between concepts necessary for informed medical decision making. Näive reduction has been informally called the "Swiss cheese" model of medical records because the information is full of holes.

A slightly more advanced implementation may dive into free-form clinical notes to search for English (or other foreign language equivalent) keywords pertaining to sensitive information. However, free-text approaches have been shown to be deficient because of limitations in the current state of the art in medical natural language processing (NLP) [13].

While a näive approach to segmentation using a predicate-reducer may provide first-order protection against simple inferences, there are second-order effects based on the use of medical information for decision making which may lead to unintentional deviations in the diagnostic or treatment process. Additionally, given sufficient non-sensitive information from a health record and sufficient domain knowledge, an adversary may be able to infer the presence of sensitive information from secondary codes such as medications or lab tests. For instance, HIV status may be suggested by findings that include low white blood cell counts, the presence of related diagnoses (e.g. Kaposi's sarcoma is often linked to AIDS), or the prescription of antiretrovirals.

The work in this paper is part of a larger effort to study the issues in medical data segmentation. In this paper, we address:

- Strategies and algorithms for identifying and segmenting sensitive medical codes.
- An analysis of how sensitive codes may be related to non-sensitive conditions.
- Techniques to infer sensitive conditions from non-sensitive codes, and possible defenses.

In subsequent papers, we plan to study and implement a more advanced inferencing model based on real patient data and statistics.

The rest of the paper is organized as follows. We define the predicate-reducer model for segmentation mathematically in Section 2, review a theoretical model for medical inference in Section 3, explore inferencing techniques in Section 4, discuss various arguments around segmentation and the medical decision making process in Section 5 and conclude in Section 6.

2. THE PREDICATE-REDUCER MODEL

Let Π be the set of policies and individual consent preferences at the Federal, State and organizational levels that define and govern the use of sensitive information. For each policy $\pi \in \Pi$, let $S(\pi, \sigma) \subseteq \sigma$ be the set of sensitive patient history information in σ governed by π.

The function of the predicate $P_\pi : \sigma \rightarrow \{true, false\}$ is to determine whether any sensitive code as defined in π is present in the health record, i.e., whether $S(\pi, \sigma) \bigcap \sigma \neq \emptyset$. The function of the reducer R_π is to isolate any sensitive information as defined in π, so that $R_\pi(\sigma) \bigcap S(\pi, \sigma) = \emptyset$. Note R_π can be composed and the composition can be proven to be commutative. For example, $R_{\pi MH}(R_{\pi HIV}(\sigma)) = R_{\pi HIV}(R_{\pi MH}(\sigma))$ isolates both HIV and mental health information. In practice, there may be a multitude of laws which may apply, and some conflict resolution may be necessary to determine how to apply the policies. We refer readers to [22] for some strategies on how to do so.

2.1 Naive segmentation

To illustrate how segmentation may be performed by a näive *predicate-reducer*, we present a simple example of a policy. In this example, suppose that a patient has some sensitive information present in his medical record that he does not consent to reveal. Assume that the medical record code is the only place where this sensitive information is present, and assume that the health record contains no free text or other information that can indirectly reveal the condition. In this case, a $reducer$ would simply isolate the offending codes. If the record uses standard ICD-9 codes[1], then the following strategy can be employed as a starting point:

1. Mental Health - Mental health codes are listed as a subtree `290-319` in the ICD-9 tree.
2. Substance Abuse - Section `305` in ICD-9 describes various forms of substance abuse, as defined in Federal Confidentiality of Alcohol and Drug Abuse Patient Records regulations (Part 2).
3. Sexually Transmitted Diseases (STDs) - ICD-9 section `099.9` represents various STDS.

The approach presented here is a näive simplified segmentation strategy. In practice, a health record may actually contain a mix of ICD-9, SnoMED, LOINC, HL7 and other codes depending on the vendor of the health system and their data formats. Also, there are aggregate ICD-9 codes that describe multiple conditions. For instance, Kaposi's Sarcoma[2], a common comorbidity with AIDS, with the ICD code `173.9` would also have to be isolated in addition to the AIDS ICD code `042`. These codes cross-cut multiple subtrees in the ICD-9 code system, so simply removing the subtrees of ICD-9 codes directly related to the sensitive categories of information is insufficient.

At present, there is a lack of any authoritative lists of sensitive conditions [15], but there are proposals for a sensitive code flag [27], which states whether a particular code is deemed sensitive.

3. HYPOTHETICO-DEDUCTIVE MODEL

In this section, we briefly review the hypothetico-deductive model [10, 21, 33] of medical decision making. This model helps diagnosticians reason through diagnostic problems by using a hypothesis testing cycle. A good understanding of this model can help the reader better grasp the principles behind the inferencing algorithms described in Section 4.

In this paper, we use the formal definition of a medical diagnostic problem from Reggia's formalization of the hypothetico-deductive

[1]ICD-9 codes are organized in a tree hierarchy, see the appendix
[2]A kind of malignant neoplasm of the skin.

model [32]. In brief, Reggia describes diagnosis as a process of finding a plausible explanation for some given set of manifestations such as medical signs and symptoms.

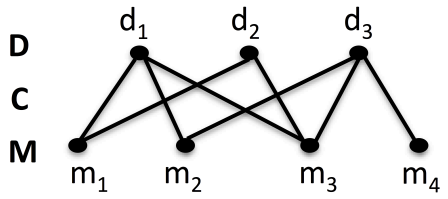

Figure 1: Generalized diagnosis model

More formally, let $\mathbf{D}$ be a finite set of all possible disorders that can occur, and let $\mathbf{M}$ be a finite set of all possible manifestations that can occur when one or more disorders are present. For example in medicine, $\mathbf{D}$ can represent all known diseases, and $\mathbf{M}$ would represent all possible symptoms, signs, and test results that can be caused by diseases in $\mathbf{D}$. We assume that diseases are well-defined by their manifestations.

To capture the intuitive notion that a disease causes manifestations, we assume knowledge of a relation $\mathbf{C} \subseteq \mathbf{D} \times \mathbf{M}$, where $\langle d_i, m_j \rangle \in \mathbf{C}$ represents "d_i can cause m_j". Note that $\langle d_i, m_j \rangle \in \mathbf{C}$ doesn't imply that m_i always occurs when d_i is present, but only that m_j may occur. For instance, a patient with the flu may experience symptoms such as fever, diarrhea, and coughing, but not all of these symptoms may be present at a particular point in time, or manifest at all.

Given $\mathbf{D}$, $\mathbf{M}$, and $\mathbf{C}$, the following sets can be defined for $d_i \in \mathbf{D}$ and $m_i \in \mathbf{M}$: $man(d_i) = \{m_i | \langle d_i, m_i \rangle \in \mathbf{C}\}$, and $causes(m_i) = \{d_i | \langle d_i, m_i \rangle \in \mathbf{C}\}$.

These sets reflect how human diagnosticians might represent medical knowledge. For instance, medical textbooks typically describe the set of $man(d_i)$ for each disease d_i. Diagnosticians typically refer to the "differential diagnosis" of a disease, which corresponds to the set $causes(m_j)$. If $man(d_i)$ is known for every disorder d_i, or if $causes(m_i)$ is known for every manifestation m_i, then the causal relation $\mathbf{C}$ can be completely determined. Note that in practice, not all diseases/manifestations are well-defined enough for these conditions to hold. For instance, some diseases have unclear etiology and/or presentation where the disease is named after general symptoms rather than the underlying cause; some of these diseases are classic *diagnoses of exclusion* such as *essential hypertension* and *fever of unknown origin* [12].

Let $man(\mathbf{D}) = \bigcup_{d_i \in \mathbf{D}} man(d_i)$ denote the set of manifestations for disorders in $\mathbf{D}$, and let $causes(\mathbf{M}) = \bigcup_{m_i \in \mathbf{M}} causes(m_i)$ denote the causes of manifestations in $\mathbf{M}$.

There is a distinguished set $\mathbf{M}^+ \subseteq \mathbf{M}$ which represents the manifestations which are known to be present. While $\mathbf{D}$, $\mathbf{M}$, and $\mathbf{C}$ are general knowledge about a class of diagnostic problems, $\mathbf{M}^+$ represents the manifestations occurring in a specific case. Using this terminology, a diagnostic problem is a 4-tuple $\langle \mathbf{D}, \mathbf{M}, \mathbf{C}, \mathbf{M}^+ \rangle$. For any diagnostic problem, $\mathbf{E} \subseteq \mathbf{D}$ is an *explanation* for $\mathbf{M}^+$ if: (E1) $\mathbf{M}^+ \subseteq man(E)$, (E2) $\mathbf{E}$ is fit by some definition of fitness, e.g. minimizing $|\mathbf{E}|$ for parsimony. For condition (E1), $\mathbf{E}$ covers $\mathbf{M}^+$, or the explanation covers all the manifestations noted. For (E2), human diagnosticians usually apply the heuristic of Occam's razor or the simplest explanation possible.

Figure 1 depicts all manifestations caused by d_i and all the possible disorders that cause m_j respectively. For instance, d_1 causes $\{m_1, m_2, m_3\}$, and m_4 is caused exclusively by d_3. If d_3 were a sensitive condition, knowing the presence of m_4 in this model would highly suggest d_3 as the causative agent, as no other diseases can cause this manifestation. However, the remaining manifestations, m_1 and m_3 can be explained by the presence of d_1 or d_2. This can provide plausible deniability for disease d_1 or d_2.

4. INFERENCING TECHNIQUES

The problem of inferencing sensitive conditions from segmented manifestations can be modeled as a best-match problem, for a set of manifestations and relations. More precisely, an inferencing problem can be represented as a 4-tuple $\langle \mathbf{D}, \mathbf{M}, \mathbf{C}, \mathbf{R}(\mathbf{M}^+) \rangle$, where $R(M^+)$ represents segmented manifestations. A good match may satisfy the plausibility and fitness criteria described in Section 3.

4.1 Identifying sensitive medical concepts and terms

We obtained a list of categories of protected patient health information from federal and state privacy laws including HIPAA[28], HITECH[39], and state laws[20]. The categories covered by the privacy laws include mental health diseases, sexually transmitted diseases, substance abuse, and genetic health conditions.

Let S be the set of diseases and manifestations considered sensitive. Then $S \bigcap D$ can be derived from reputable sources such as DSM-IV[2] for mental health and substance abuse, and CDC[5] for STDs. Some groups [35] are working on curating such lists for data segmentation. If a manifestation is linked to only sensitive diseases, then we consider itself sensitive. I.e., $causes(m_i) \subseteq S \rightarrow m_i \in S$.

4.2 Identifying correlations

Many concepts in medicine are often correlated, and our model can capture some of those correlations. More precisely, we extend the model in Section 3 to include a set $\mathbf{T}$ which represents all treatments (such as medications and surgical procedures) and a set of intervention relations $\mathbf{I} \subseteq \mathbf{D} \times \mathbf{T}$, where $\langle d, t \rangle \in \mathbf{I}$ represents "t is an intervention for d." We define $treatments(d) = \{t | \langle d, t \rangle \in \mathbf{I}\}$ as the set of treatments available for disease d, and $treats(t) = \{d | \langle d, t \rangle \in \mathbf{I}\}$ as the set of diseases that t treats.

Furthermore, we model adverse effects as the relation $\mathbf{A} \subseteq \mathbf{T} \times \mathbf{M}$, where $\langle t, m \rangle \in \mathbf{A}$ represents "treatment t can cause manifestations m" , and define $adverse(t) = \{m | \langle t, m \rangle \in \mathbf{A}\}$ the adverse effects that can be caused by t.

Based on this model, the following correlations can be used for inferencing:

- *Causality* - This is represented by the relation $\mathbf{C}$ in our model.
- *Common causality* - We say that manifestations m_1 and m_2 have common causality if there exists disease d such that $\langle d, m_1 \rangle \in C$ and $\langle d, m_2 \rangle \in C$.

- *Competing hypotheses* - This is intrinsically part of our model, and represented as the set of explanations $\mathbf{E}$ for the competing hypotheses.
- *Treatment* - Our model links treatment to disease and adverse effects though the relations $\mathbf{I}$ and $\mathbf{A}$ respectively.

Table 1 shows some of the relationships that can occur between concepts. Many of these relationships can be modeled as drug treatments, and side effects by $\mathbf{T}$ and $\mathbf{A}$ in our model respectively. Some of the drugs such as Risperidone and Carbamazepine are used primarily to treat mental disorders, and their use can reveal the presence of one or more mental health conditions, e.g. $treats(t) \subseteq S$. This information might be available to doctors and pharmacies for the purpose of handling prescriptions and checking for drug-drug interactions.

Furthermore, some treatments such as Citalopram have multiple uses, and can treat depression or hot flashes. Since there are multiple uses, identifying the drug doesn't completely suggest that its use is for the treatment of depression, i.e. $treats(t) \not\subseteq S$. To understand the ability to inference better, Figure 2 illustrates a space of medical concepts. Some concepts such as cervical cancer and Kaposi's sarcoma are considered non-sensitive, and they are represented by points outside of the set of sensitive information. Some concepts such as AIDS and Schizophrenia exist within the set of sensitive concepts S. Since sensitive concepts can be isolated, in order for an inference of the presence of a sensitive condition to be made, some manifestations of these conditions must be outside of the set of sensitive concepts, and some relations must cross the boundary from sensitive to non-sensitive manifestations to be detected, e.g. $\exists \langle d_i, m_j \rangle | d_i \in S, m_j \notin S$.

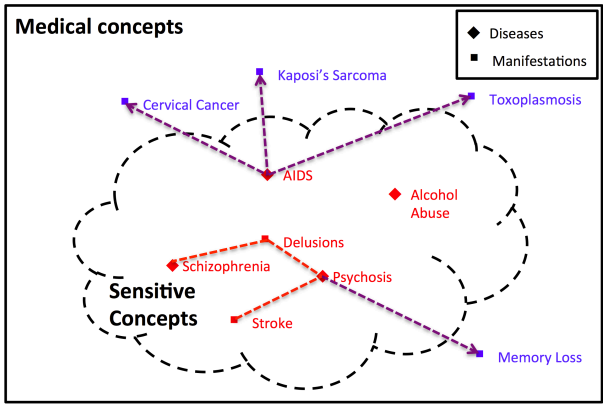

Figure 2: Medical concept space - This shows that sensitive diseases which have associations outside the sensitive concepts leave clues that may not be directly hidden by näive segmentation.

Identifying Links

Once the basic set of sensitive conditions S has been determined, the next task is to identify the direct relations I and A for all conditions in S. This can be done by consulting medical knowledge bases such as SnoMed[34], consulting the medical literature and mining other sources for correlations.

While simple links can help identify some direct relationships mentioned earlier, more complex inferences generally deal with multiple competing hypotheses. In order to determine whether Citralopram, mentioned earlier, is likely being used as an antidepressant or used to treat hot flashes, it would be helpful to know what other signs or symptoms are present. For instance, knowing that the patient has used another antidepressant in the past may suggest its functionality as an antidepressant, whereas knowing that the patient is menopausal might suggest its role in treating hot flashes, although it's possible that the patients may have both conditions simultaneously. Sometimes these secondary links are non-trivial, and involve multiple competing hypotheses. In order to help evaluate these hypotheses, we use the model and methods described in Section 3. Note that a good inference usually incorporates information about the strength of correlation of links and the degree to which competing hypotheses offer deniability of any particular hypothesis.

Initially, we attempted to use SnoMed to identify these correlations, and while we found that it was able to identify some of the relationships that we were interested in, many interesting relations could not be easily found in the knowledge base. This could be due to several reasons including a difficulty in navigating the concept map, encountering areas of the map that do not yet represent state of the art medical knowledge, and encountering areas where the medical associations are being debated or changed.

Although SnoMed was able to help us partially determine whether some associations were plausible, it gave little indication as to whether or not the association was *likely*. To attain better coverage, we devised a method to mine other sources of medical data for associations, and developed a set of techniques to process the information efficiently. Our technique is based on classic co-reference mining for associations. We chose this technique because it was simple to implement using existing technologies, fast, and fairly effective.

To build out our knowledge base, we incorporated information from:

- **PubMed** - 22 million articles from NLM and NIH.
- **PubMed Open Access subset** - XML data dump with full text for 615,000 articles.
- **Wikipedia** - We use a subset of 14,386 articles in Wikipedia which contain medical codes.
- **Google** - We use a generic Google search using AND queries.

For Wikipedia and Pubmed, we created our database from publicly available XML dumps of their databases. The compressed Wikipedia dump used from April 2013 was 9.7 GB, and the compressed PubMed dump was 8.2 GB in size. Each dataset was used to create a separate database for queries. We chose this design to avoid statistical noise based on the nature of the dataset. Our search indexes were created using the open source Xapian[42] software, which is a probabilistic information retrieval system.

These data sources help act as a proxy for disease descriptions to determine the relations C, I, and A. The next section discusses the specifics of how these correlations are computed.

Concept	Description	Links	Notes
Risperidone	Treats schizophrenia, bipolar disorder, and autism.	schizophrenia, bipolar disorder, autism, weight gain, insomnia, alopecia	Use of Risperidone usually implies treatment of a mental health disorder.
Aspirin	Pain reliever.	pain relief, fever reducer, anti-inflammatory, blood thinner	Aspirin has many uses, which makes it challenging to infer what condition the user intended to treat, although pain relief is the most common usage.
Carbamazepine	Anti-convulsant and mood-stabilizing drug. Treats epilepsy and bipolar disorder.	epilepsy, bipolar disorder, headaches, drowsiness	Primarily used to treat mental health disorders. Could be used off-label to treat Complex regional pain syndrome(ICD9: 337.21)
Citalopram	Primarily used as an SSRI to treat depression. Can also be used to treat hot flashes.	depression, hot flashes, anorgasmia, nausea, diarrhea	Can treat both sensitive and non-sensitive conditions.
Lamotrigine	Primarily used as an anticonvulsant drug to treat epilepsy and bipolar disorder. Can also treat migraines.	epilepsy, bipolar disorder, migraines	Can be used to treat mental health disorders or migraines.
Olanzapine	Atypical antipsychotic used to treat schizophrenia and bipolar disorder.	schizophrenia, bipolar disorder, insomnia, weight gain, dry mouth	Usually treats mental health conditions.
Topiramate	Anticonvulsant drug for treatment of epilepsy. Can be used to prevent migraines.	epilepsy, migraines, bipolar disorder, CBT	Can be used to treat non-sensitive migraines. If cognitive behavioral therapy (CBT) added, this drug could be used to treat Bulimia Nervosa.

Table 1: Direct relationships

4.3 Hypothesis Fitness Index

The purpose of the hypothesis fitness index is to evaluate the plausibility of a given hypothesis. Given segmented manifestations $R(M^+)$, the task is to compute the most plausible explanation $E \subseteq causes(R(M^+))$, and check if $E \bigcap S = \emptyset$. More specifically, a set of competing hypotheses would be generated and evaluated with respect to the hypothesis fitness index. The process of hypothesis generation will be discussed in more detail in Section 4.4.

Let $W = D \bigcup M \bigcup T$ be the set of medical concepts.

Concept Support Index

Let $H \subseteq W$ be a set of concepts representing a hypothesis that the patient has had the medical manifestations, diseases, and treatments in H. Let $h \in H$ be a particular concept in H, then the Concept Support Index with respect to a medical knowledge document doc is defined as:

$$CSI(h, doc) = \frac{Count(h, doc)}{\sum_{w \in W} Count(w, doc)} \tag{1}$$

$$CSI(H, doc) = \sum_{h \in H} CSI(h, doc) \cdot w_h \tag{2}$$

, where $w_h \in [0, 1]$, $\sum_{h \in H} w_h = 1$, and $Count(h, doc)$ counts the number of occurrences of h in doc.

Intuitively, the co-occurrences of multiple medical concepts within the same medical knowledge document indicates a possible correlation between these concepts, and the CSI provides a heuristic measure of the relevance of the document with respect to a particular concept, relative to other concepts.

Note that there are multiple possible definitions of $CSI(H, doc)$. The current definition at Equation 2 is chosen for its simplicity, and we plan to experiment other formulations for future work.

Deniability Index

The Deniability Index with respect to a document doc measures the support for hypotheses other than H, and is defined as $1 - CSI(H, doc)$.

Hypothesis Fitness Index

While the Concept Support Index provides a vote with respect to a single document, we need a way to calculate the support for a hypothesis over a set of documents. The Hypothesis Fitness Index provides a way to do so, and is defined as

$$HFI(H, Docs) = \sum_{doc \in Docs} CSI(H, doc) \cdot weight(doc, H) \tag{3}$$

where $weight(doc, H)$ is a weighting function. The weighting function takes into account factors such as the relevance of the document with respect to the hypothesis H, and possibly scaled by a function of the size of the result set $|Docs|$ returned for the query. One way to formulate the relevance factor could be BM25 [24, 38, 42], which is defined as

$$\text{BM25}(D, Q) = \sum_{q_i \in Q} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}, \tag{4}$$

where

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \tag{5}$$

$f(q_i, D)$ is the term frequency of q_i in D, $k_1 \in \mathbb{R}^+$, $b \in [0, 1]$, and $avgdl$ is the average document length of $Docs$.

Once the HFI is calculated for each hypothesis, the hypotheses are ranked in descending order by HFI value, and this comprises the set of inferences. For instance, Table 2 illustrates a simple fitness example based on Figure 1. The first two hypotheses have similar support, and they can be considered competing hypotheses with no clear leader. This provides plausible deniability for these two solutions. The third solution, while plausible, is not well supported and also not "parsimonious". Note that d_3 is always required to achieve cover of m_4, i.e. $m_4 \rightarrow d_3$, therefore the last solution is not plausible.

Hypothesis	Rank	HFI	Notes
$\{d_1, d_3\}$	1	0.91	This is the most supported hypothesis.
$\{d_2, d_3\}$	2	0.88	This could be a competing solution, which can provide plausible deniability because the HFI support is similar.
$\{d_1, d_2, d_3\}$	3	0.35	This solution is not well supported, and not "parsimonious".
$\{d_1, d_2\}$	4	X	m_4 is not covered, so this is not a solution.

Table 2: Fitness example for m_1, m_2, m_3, m_4

4.4 Approaches to infer segmented data using correlations

Given a patient's potentially segmented EHR, Algorithm 1 attempts to discover if some sensitive medical concepts may have been segmented based on the remaining non-sensitive manifestations that are observable to the algorithm.

To do so, the algorithm selects combinations of manifestations from the health record as queries. For instance, queries corresponding to the hypotheses in Figure 1 could include:"m_1 AND m_3","m_2 AND m_3", "m_1 AND m_2 AND m_3", and so on. Each query then searches for medical knowledge documents that contain these manifestations, and the search results are then used to generate hypotheses. These can correspond to the "differential diagnosis" hypothesis list (Section 3), and the hypotheses are evaluated using the techniques from Section 4.3.

However, as Figure 3 illustrates, a näive approach to exploring the potentially large query space could be slow or intractable. In practice, we have encountered cases where the query was longer than six terms, and each term could be selected from a list of hundreds of manifestations. To address this difficulty, we use probabilistic sampling methods to help us comb the space effectively.

On the other hand, a strict interpretation of the AND operator in the query may lead to a small result set that may be insufficient to support the hypotheses. Figure 4 illustrates this idea more concretely with an example where a hypothetical patient's record contains three terms: Toxoplasmosis (TX), weight loss (WL), and cervical cancer (CC). The top query, "$TX \wedge WL \wedge CC''$ identifies documents that concern all three terms, however, this may be too restrictive, as only 25 documents from PubMed Open Access are retrieved, and that may not provide sufficient perspective and support for the hypothesis. To expand the search, we relax the conditions slightly and perform subqueries that match all but one condition. E.g. "$TX \wedge WL''$, "$TX \wedge CC''$, "$WL \wedge CC''$. Each time this process is performed, the set of documents considered is expanded. In this example, expanding one level yields support from over 2,000 documents, much more than the original 25. In practice, we've seen that this greatly improves the ability of our algorithms to infer sensitive hypotheses.

Inferring sensitive concepts from an EHR

Given an EHR, our algorithms can be applied as follows. First, to generate the queries, we select query terms from the EHR with respect to a certain probability distribution. One way to do so is to use the frequency distribution of the underlying medical knowledge base. Intuitively, this is done as a heuristic to maximize the number of documents returned.

To calculate this distribution, we take a set of documents from a given knowledge base, such as PubMed, and a set of medical terms from medical term databases such as MESH, SnoMed, or PubChem, and build a set containing the medical terms and their probability distribution within the knowledge base.

Convergence is reached when the top k hypotheses in the cumulated results remain the same in successive iterations. Some studies show that human diagnosticians typically consider 4 ± 1 hypotheses[11], and we choose to experiment with k in the range from 3 to 5. Experimentation with larger values of k is left as future work.

$hypotheses \leftarrow \emptyset$;
repeat
  $query \leftarrow \emptyset$;
  for $j = 1 \rightarrow numTerms$ **do**
    /* select a concept from the EHR using a probability distribution */
    $x \leftarrow select\_concept(concept\_probs, EHR)$
    $query \leftarrow query \bigcup x$;
  end
  /* search for docs that contain the query terms */
  $sr \leftarrow search(query, knowledge\_base)$;
  /* Identifies hypotheses from medical concepts in documents */
  $hypotheses \leftarrow update\_hyp(hypotheses, sr)$;
  /* Evaluates hypotheses according to plausibility criteria */
  $results \leftarrow eval\_hypotheses(hypotheses) \bigcup results$;
until *convergence*;
$rank(results)$;

Algorithm 1: Inference algorithm

4.5 Results

Table 3 illustrates some of the results from our approach. The first set of queries finds results for "Rett Syndrome," a mental health/neurological condition listed in DSM-IV as a mental disorder. The queries themselves involved non-sensitive manifestations to suggest a sensitive result.

The next set of results illustrates an example found by our algorithm involving the inference of AIDS from four non-sensitive concepts, and testing the result across different data sources including PubMed, Google and Bing. We verified this result with medical experts, and they told us that two of the concepts were nondescript. Rotavirus is a common cause of diarrhea, and "weight loss" is nonspecific, so few inferences can be made based on those conditions alone. Furthermore, the presence of toxoplasmosis, a parasitic disease that is estimated to affect one third of the world's population[19] does little to suggest the cause on its own. Considering the final term, "cervical cancer" alone also does little to narrow down the exact cause, as the symptom can have many possible causes. Combining all of these nondescript concepts together narrows the list of plausible hypotheses down to AIDS. Our medical experts have independently verified this result before we told them that the query would yield AIDS.

Another notable example is the query for HIV involving "Hepatitis" and "Hepatitis C". One might expect that these two concepts may be synonyms, but upon closer inspection, it turns out that the more specific version, "Hepatits C" produced more specific results. Hepatitis C is associated with intravenous drug use, unsafe blood

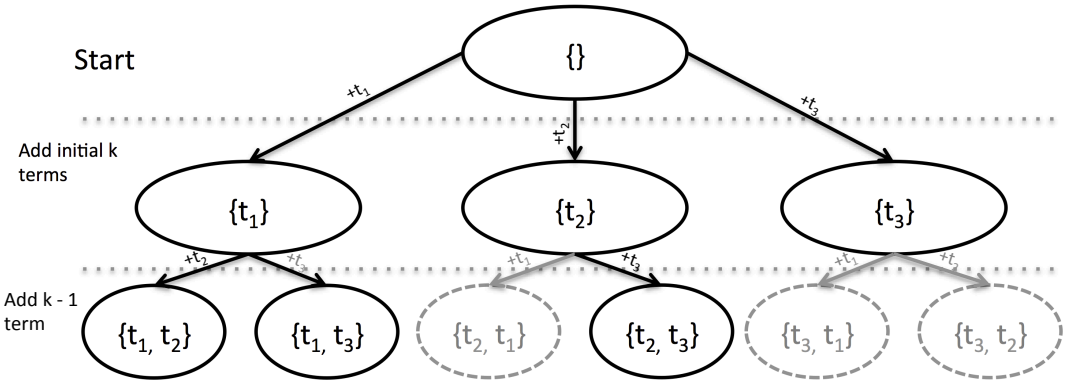

Figure 3: Term space exploration - Hierarchy of possible query terms

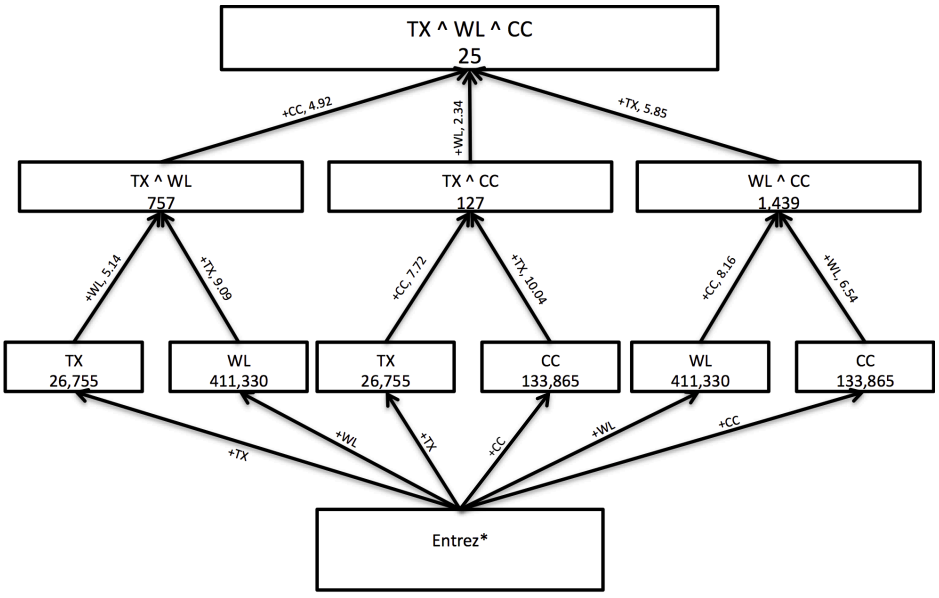

Figure 4: Query lattice

transfusions and anal penetrative sex - a sex practice associated with increased risk of HIV transmission. Our medical experts have also verified this result.

If the reader is interested to explore the results, they may perform the suggested queries in their search engine of choice. These queries happened to return the results noted as of May 6th, 2013, but there is no guarantee that the ranking or distribution may not change significantly, although we have not observed such a change during our development and testing. We recommend that the reader use the private browsing mode of their browser to avoid biasing the search results with personalized content derived from other browsing habits. Another tip is to use Tor connected to an exit node in the United States to get the same kinds of localized results we obtained.

4.6 Approaches to protect against inferencing

As we have discussed, the ability to generate a good inference depends on: (1) the non-sensitive manifestations of sensitive concepts, e.g., toxoplasmosis secondary to AIDS, (2) the Hypothesis Fitness Index (Section 4.3) of the hypotheses for the manifestations. Approaches to defend against inferencing may address these conditions.

One approach could be to reduce non-sensitive manifestations that can be linked with sensitive concepts. This approach has the benefit of potentially limiting the information necessary to make inferences. However, this could limit legitimate medical use, including diagnosis, treatment, and research.

Another approach is to reduce the relative strength of the leading hypothesis, or to strengthen alternative non-sensitive hypotheses to allow for plausible deniability. For example, Citalopram (Section 4.2) can treat depression or hot flashes. If depression is a leading hypothesis, then noting a history of hot flashes in the EHR could provide plausible deniability that the treatment is for hot flashes.

A more in-depth example might involve an AIDS patient where the primary disease is HIV infection which causes the AIDS syndrome which in turn leads to immunodeficiency and opportunistic infection. Possible causes or risk factors (and plausibly deniable explanations) for opportunistic infections include: organ transplant, chemotherapy, genetic predisposition, and antibiotic treatment.

An approach that has seen some success in the setting of differential privacy [7, 9, 8] is to introduce noise to prevent adversarial attacks, yet allow legitimate uses. One potential downside to this approach is that it is unclear what the threshold should be, and it may increase the risk that legitimate uses of the EHR are affected.

This area may require further research and some potential defense may draw from ideas in association rule hiding[14]. Some of these ideas include support-based and confidence-based distortion[3], blocking[36] and border-based approaches[37]. Another technique is to avoid the disclosure of information entirely, for example, private set intersection[6] could be used to determine drug-drug interactions while minimizing information disclosure.

5. DISCUSSION AND FUTURE WORK

Contextual integrity [4] has been proposed as an approach to the medical privacy problem. One extreme for medical privacy is to isolate all sensitive codes without regard to the patient's privacy preferences (consent) or the particular medical situation involved (e.g. emergency room). Contextual integrity suggests that the use of sensitive medical data should be governed by the purpose and underlying contextual cues such as the patient's privacy expectations, fears of stigma and other factors such as risk of misdiagnosis. Previous tools such as [22] can help the organization identify if a communication is compliant with the law, and we extend upon this approach by educating and empowering the patient. Contextual integrity can be modeled in R by adding a parameter $R(\sigma, c)$, where c represents the context.

Studies such as [17] have shown the relative importance of obtaining an accurate patient history, and concerns about segmenting patient history have led to a debate within the medical community about rejecting the use of data from segmented records on the grounds that medical professionals might be held legally liable[3] for misdiagnosis or medical error due to incomplete/incorrect information [23]. One possible effect of this situation is that doctors may proceed more cautiously in the presence of segmented information[4] and ask more verbal questions, as well as ordering more tests to compensate. However, we believe that there is a soft upper-limit to what doctors can test for due to limits in the amount of discomfort a patient is willing to endure, potential dangers in over-testing [1], and rationing of precious resources [30].

Also, as scientists study correlations in biology and medicine, it may become more challenging to hide the associations with sensitive conditions. One example of such databases is the OMIM database[26] which can be used to link certain sensitive conditions such as autism and depression with certain genes.

For future work, we plan to: (1) expand the knowledge base of articles considered, (2) develop a more advanced inferencing model which may incorporate more complex causal networks[31], (3) better curate and interpret the information in the articles to get a better approximation of the relations C, I, A, and (4) test our framework with real patient data by working with real-world healthcare institutions.

6. CONCLUDING REMARKS

We have demonstrated that while a näive approach to medical segmentation may address some first-order issues regarding privacy, complex medical and biological correlations may reveal second-order conditions that cannot be hidden easily with the näive approach. As a result, we believe that if data is to be segmented to protect privacy, it is best to consider deeper medical meaning and context when segmenting data, and using the segmented data in real-life medical situations.

Acknowledgment

We gratefully acknowledge support from ONC SHARPS under Grant No. HHS-90TR0003/01 and the National Science Foundation under the following grants: CNS-0831199 and CCF-0424422. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any sponsoring organization.

We thank Carl Gunter and Mike Berry for their initial conceptualization of the predicate-reducer model and for their helpful insights in working through the details. We also thank James Reggia for his

[3]It's unclear if the party that withheld information is at fault, or if the party that received the incomplete information is at fault.

[4]If the data has been marked as sensitive. Otherwise, doctors may not know that data has been isolated.

formalization of the hypothetico-deductive model and his helpful insights in working with the model. Brad Malin provided helpful resources along the way to guide us in our thought process, and Ivan Handler gave us a good perspective of these issues at a health information exchange level.

7. REFERENCES

[1] H. Abrams. The Overutilization of X-rays. *New England journal of medicine*, 300(21):1213–1216, 1979.

[2] A. P. Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. American Psychiatric Publishing, Inc., 2000.

[3] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure Limitation of Sensitive Rules. In *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on*, pages 45–52. IEEE, 1999.

[4] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and Contextual Integrity: Framework and Applications. In *IEEE Symposium on Security and Privacy*, 2006.

[5] Centers for Disease Control and Prevention. 2011 Sexually Transmitted Diseases Surveillance.

[6] D. Dachman-Soled, T. Malkin, M. Raykova, and M. Yung. Efficient Robust Private Set Intersection. *International Journal of Applied Cryptography*, 2(4):289–303, 2012.

[7] C. Dwork. Differential Privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[8] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and Differential Privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

[9] C. Dwork and A. Smith. Differential Privacy for Statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):2, 2010.

[10] A. Elstein, L. Shulman, and S. Sprafka. Medical Problem-Solving. *Academic Medicine*, 56(1):75, 1981.

[11] A. Elstein, L. Shulman, S. Sprafka, et al. *Medical Problem Solving: An Analysis of Clinical Reasoning*, volume 2. Harvard University Press Cambridge, MA, 1978.

[12] R. Engle Jr and B. Davis. Medical Diagnosis: Present, Past, and Future. I. Present Concepts of the Meaning and Limitations of Medical Diagnosis. *Archives of internal medicine*, 112:512, 1963.

[13] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated Encoding of Clinical Documents based on Natural Language Processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[14] A. Gkoulalas-Divanis and V. Verykios. *Association Rule Hiding for Data Mining*. Advances in Database Systems. Springer US, 2010.

[15] M. M. Goldstein and A. L. Rein. Data Segmentation in Electronic Health Information Exchange: Policy Considerations and Analysis. The George Washington University Medical Center, 2010.

[16] J. Groopman. *How Doctors Think*. Houghton Mifflin, 2007.

[17] J. Hampton, M. Harrison, J. Mitchell, J. Prichard, and C. Seymour. Relative Contributions of History-taking, Physical Examination, and Laboratory Investigation to Diagnosis and Management of Medical Outpatients. *British Medical Journal*, 2(5969):486, 1975.

[18] K. Henriksen and H. Kaplan. Hindsight Bias, Outcome Knowledge and Adaptive Learning. *Quality and Safety in Health Care*, 12(suppl 2):ii46–ii50, 2003.

[19] D. Hill and J. Dubey. Toxoplasma Gondii: Transmission, Diagnosis and Prevention. *Clinical microbiology and infection*, 8(10):634–640, 2002.

[20] L. E. Joy Pritts, Angela Choy and J. Hustead. The State of Health Privacy. Institute for Health Care Research and Policy Georgetown University, 1999.

[21] J. Kassirer, G. Gorry, et al. Clinical Problem Solving: A Behavioral Analysis. *Annals of Internal Medicine*, 89(2):245, 1978.

[22] P. Lam, J. Mitchell, and S. Sundaram. A Formalization of HIPAA for a Medical Messaging System. *Trust, Privacy and Security in Digital Business*, pages 73–85, 2009.

[23] S. Mangalmurti, L. Murtagh, and M. Mello. Medical Malpractice Liability in the Age of Electronic Health Records. *New England Journal of Medicine*, 363(21):2060–2067, 2010.

[24] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[25] Mary Mosquera. VA, SAMHSA test exchange of tagged substance abuse data. Government Health IT, 2012.

[26] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine. Online Mendelian Inheritance in Man, 2013.

[27] J. Moehrke. Proposal for confidentiality Code vocabulary. Institute for Health Care Research and Policy Georgetown University, 2011.

[28] Office for Civil Rights. Summary of the HIPAA privacy rule. US Department of Health and Human Services, 2003.

[29] D. Peel. Your Medical Records Aren't Secure. *The Wall Street Journal-Opinion*, 2009.

[30] E. D. Pellegrino. Rationing Health Care: The Ethics of Medical Gatekeeping. *J. Contemp. Health L. & Pol'y*, 2:23, 1986.

[31] Y. Peng and J. A. Reggia. Diagnostic problem-solving with causal chaining. *International Journal of Intelligent Systems*, 2(3):265–302, 1987.

[32] J. Reggia, D. Nau, and P. Wang. Diagnostic Expert Systems Based on a Set Covering Model. *International Journal of Man-Machine Studies*, 19(5):437–460, 1983.

[33] A. Rubin. Hypothesis Formation and Evaluation in Medical Diagnosis. Technical report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology. AI-TR-316, 1975.

[34] P. Ruch, J. Gobeill, C. Lovis, and A. Geissbuhler. Automatic Medical Encoding with SNOMED Categories. *BMC Medical Informatics and Decision Making*, 8(Suppl 1):S6, 2008.

[35] S & I Framework. Data Segmentation for Privacy Charter and Members, 2011.

[36] Y. Saygin, V. S. Verykios, and C. Clifton. Using Unknowns to Prevent Discovery of Association Rules. *ACM SIGMOD Record*, 30(4):45–54, 2001.

[37] X. Sun and P. S. Yu. A Border-based Approach for Hiding Sensitive Frequent Itemsets. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[38] K. M. Svore and C. J. Burges. A Machine Learning Approach for Improved BM25 Retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1811–1814. ACM, 2009.

[39] U.S. Congress. HiTech Act. TITLE XIII - Health Information Technology. 2009.

[40] U.S. Department of Health and Human Services. Benefits of Electronic Health Records (EHRs). 2012.

[41] J. Walker, E. Pan, D. Johnston, J. Adler-Milstein, D. W. Bates, and B. Middleton. The Value Of Health Care Information Exchange and Interoperability. *Health Aff*, Jan. 2005.

[42] Xapian. The BM25 Weighting Scheme.

Sensitive goal	Query	Results	Medical codes	Notes
Rett Syndrome	"wringing" AND "female" AND "constipation" AND "scoliosis"	6 articles suggest Rett Syndrome.	F84.2[6], R09.0[5], K59.0[6], 737.0[5]	Pubmed
Rett Syndrome	"wringing" AND "female" AND "constipation" AND "scoliosis"	1.73M results, 5 of top 10 results suggest Rett Syndrome, including NIH Medline.	F84.2[6], R09.0[5], K59.0[6], 737.0[5]	Google
AIDS	"Toxoplasmosis" AND "Weight loss" AND "Cervical cancer"	6 articles suggest AIDS. No solutions if "rotavirus" is also included in query.	042[5], 130[5], 783.21[5], 008.61[5], 180[5]	Pubmed
AIDS	"Toxoplasmosis" AND "Weight loss" AND "Rotavirus" AND "Cervical cancer"	203,000 results. 3 of top 5 results report AIDS in title, 4 out of top 5 report AIDS in content.	042[5], 130[5], 783.21[5], 008.61[5], 180[5]	Google
AIDS	"Toxoplasmosis" AND "Weight loss" AND "Rotavirus" AND "Cervical cancer"	2,940 results. 3 of top 8 results report AIDS in content or title.	042[5], 130[5], 783.21[5], 008.61[5], 180[5]	Bing
HIV	"Tuberculosis" AND "Hepatitis" AND "Cancer" AND "Hepatitis C"	1.98M results, 4 out of top 10 mention AIDS/HIV. Note that "Hepatitis" and "Hepatitis C" seem to be similar.	042[5], 010[5], 070.70[5], 573.3[5], 140[5]	Google
HIV	"Tuberculosis" AND "Hepatitis" AND "Cancer"	9.98M results, 3 out of top 10 mention HIV/AIDS. Search with only "Hepatitis".	042[5], 010[5], 573.3[5], 140[5]	Google
HIV	"Tuberculosis" AND "Hepatitis C" AND "Cancer"	1.5 M results, 4 of top 10 mention HIV/AIDS. Note that specifying "Hepatitis C" rather than "Hepatitis" produces fewer more specific results possibly because "Hepatitis C" is linked to risky sexual and drug use behavior.	042[5], 010[5], 070.70[5], 140[5]	Google
Chlamydia	"Pelvic inflammatory disease" AND "Urethritis" AND "Infertility" AND "Trachoma"	1.44M results, 7 out of 10 top results suggest Chlamydia.	099.41[5], 614[5], 597[5], 606[5], 076[5]	Google
Catatonia	"extreme excitement" AND "mutism" AND "grimacing" AND "waxy flexibility"	104 results, 8 of 10 top results suggest catatonia.	F20.2[6], D002375[7]	Google
Alcoholism	cancer AND "child abuse" AND "domestic violence" AND "heart failure"	2 of top 5 results suggest Alcohol.	F303[5], 140[5], D017579[7]	Google.
Schizophrenia	"Cognitive Behavioral Therapy" AND "delusion" AND "genetics" AND "metabolic syndrome" AND "Alzheimer's"	2 of top 5 results suggest Schizophrenia.	295[5], D015928[7], 297[5], 277.7[5], 331.0[5]	Google
Schizophrenia	"hallucination" AND "genetics" AND "paranoia"	Suggests Alzheimer's in 4 of top 5 results.	295[5], 7801.[5], 295.3[5]	Google
Alzheimer's disease	"hallucination" AND "genetics" AND "paranoia" AND "memory"	Suggests Alzheimer's in 3 of top 10 results. Note that this search is similar to the one above. The addition of the term "memory" suggests Alzheimer's and not "Schizophrenia".	331.0[5], 780.1[5], 295.3[5]	Google
Alzheimer's disease	"hallucination" AND "genetics" AND "paranoia" AND "memory"	Suggests Alzheimer's in 3 of top 5 results.	331.0[5], 780.1[5], 295.3[5]	Bing
Autism	"phenylketonuria" AND "gaze" AND "magnetoencephalography"	3 of top 10 results suggest Autism	299.0[5], 270.1[5], D015225[7]	Google

[**5**] - ICD-9, [**6**] - ICD-10, [**7**] - MeSH medical subject headings

Table 3: Example queries