Medical Data Segmentation for Privacy

Ellick Chan, Peifung E. Lam, and John C. Mitchell

Introduction

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.

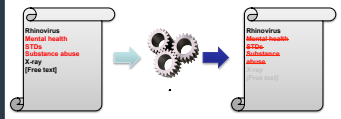
We present an initial study that describes an approach for segmenting sensitive medical codes to protect patient privacy and to comply with privacy laws.

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 possible defensive approaches against techniques that an adversary may use to infer the segmented condition.

Segmentation

A – Medical algorithm
π – Policy determines sensitive code s
M – Medical record
Predicate P(M, π) – Determines if s ∈ M
Reducer R(M, π) – Removes s from M

Ideal reducer $A(m) = A(R_\pi(m)) \; \forall m \in M$



References

[1] Elick M Chan, Peifung E Lam, John C Mitchell. Understanding the challenges with Medical Data Segmentation for Privacy. HealthTech 2013.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] J.Groopman. How Doctors Think. Houghto nMifflin,2007.

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

Segmentation Example

Medical Record

Medications
M1. Tylenol
M2. Sudafed
M3. AZT
M4. Bactrim

Letter
I hope you and your partner had a great weekend in Provincetown and that the thrush has improved with the mouthwash sample I gave you.

Problem List
P1. Headache
P2. Sinus Infection
P3. HIV positive
P4. UTI

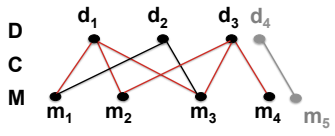
Adapted from J. Halamka, 2012

"Easy" guide to segmentation

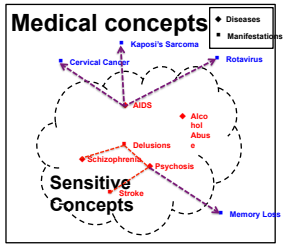
1. Hide P3 – HIV positive status
2. Hide M3 – AZT is a treatment for HIV
3. Should Bactrim (M4) be hidden? It could treat AIDS-related infections. UTI can also be treated.
4. Does "partner" imply homosexuality?
5. Provincetown is a popular destination for gays
6. Thrush or Candidiasis is a common mouth infection that AIDS patients may get
7. Headaches (P1) and sinus infections (P2) are common in AIDS patients, and Tylenol (M1) and Sudafed (M2) can be used to treat these symptoms.
8. Although AZT (M3) can be hidden, it's side effects cannot. Symptoms such as Anemia, neutropenia, hepatotoxicity, and cardiomyopathy may be present.

Inferencing Model

Hypothesis
{d1, d3}
{d2, d3}
{d1, d2, d3}
~~{d1, d2}~~

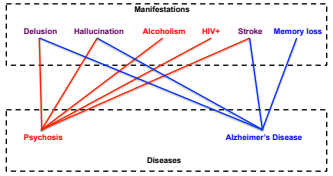


Reggia's set cover model
- Plausibility – set cover
- Likelihood – Occam's razor and fitness



Results

Condition	Query	Results	Medical codes	Notes
Rett Syndrome	"wringing" AND "female" AND "constipation" AND "scoliosis"	3 articles suggest Rett Syndrome.	F84.2, R09.0, K59.0, 737.0	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, R09.0, K59.0, 737.0	Google
AIDS	"Toxoplasmosis" AND "Hepatitis B" AND "Encephalopathy" AND "Progressive multifocal leukoencephalopathy" AND "Cryptococcosis"	140,000 results. 5 of top 10 suggest AIDS.	130, 070.2, 348.30, 046.3, 117.5	Google
AIDS	...	18,000 results. >8 of top 10 suggest AIDS.	130, 070.2, 348.30, 046.3, 117.5	Bing



Source: PubMed, NIH.gov

Algorithm

$hypotheses \leftarrow \emptyset$;
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

Fitness

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:

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

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

, 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.

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

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 by the query. One way to formulate the relevance factor could be BM25 [24, 38, 42], which is defined as

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

where

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

$f(q_i, D)$ 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 $Docs$.

Defenses

Deniability through relative strengths of hypotheses

- Hide non-sensitive EHR as well
- Enhance competing hypothesis, e.g. Citalopram can treat hot flashes or depression.
- Association rule hiding

Thanks to:

NSF
SHARPS Strategic Healthcare IT Advanced Research Projects on Security