STRATEGIC HEALTH IT ADVANCED RESEARCH PROJECTS ON SECURITY

sharps.org

Stanford
UNDERSTANDING THE CHALLENGES WITH MEDICAL DATA SEGMENTATION

Ellick M. Chan, Peifung E. Lam and John C. Mitchell
8/12/13
Health Information Exchange (HIE)

- Federal
  - HIPAA
  - HITECH

- State laws on
  - Mental Health
  - Substance Abuse
  - STDs
  - Genetic testing

- Organizational
Compliance approaches

Automated Policy

HIPAA Law

§ 164.502 Uses and disclosures of protected health information: general rules.
(a) Standard. A covered entity may not use or disclose protected health information, except as permitted or required by this subpart or by subpart C of part 160 of this subchapter.

(1) Permitted uses and disclosures. A covered entity is permitted to use or disclose protected health information as follows:

(i) To the individual;
(ii) For treatment, payment, or health care operations, as permitted by and in compliance with §164.506;
(iii) Incident to a use or disclosure otherwise permitted or required by this subpart, provided that the covered entity has complied with the applicable requirements of §164.502(b), §164.514(d), and §164.530(c) with respect to such otherwise permitted

Data segmentation

Health Record

- Medications
- Previous diagnoses
- Labs

Sensitive conditions

According to research by the California HealthCare Foundation, 15 percent of patients who know their information will be shared would hide information from their doctor, and another 33 percent would consider hiding information[1].
Example

Medications

- 1. Tylenol
- 2. Sudafed
- 3. AZT
- 4. Bactrim

Problem List

- 1. Headache
- 2. Sinus Infection
- 3. HIV positive
- 4. UTI

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.

Adapted from J. Halamka, 2012
Example

Medications

- 1. Tylenol
- 2. Sudafed
- 3. AZT
- 4. Bactrim

Problem List

- 1. Headache
- 2. Sinus Infection
- 3. HIV-positive
- 4. UTI

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.
Medications
- 1. Tylenol
- 2. Sudafed
- 3. AZT +side effects
- 4. Bactrim

Problem List
- 1. Headache
- 2. Sinus Infection
- 3. HIV-positive
- 4. UTI

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.

Zidovudine (INN) or azidothymidine (AZT) is a type of antiretroviral drug used for the treatment of HIV/AIDS.

Side effects: anemia, neutropenia, hepatotoxicity, cardiomyopathy, and myopathy

Adapted from J. Halamka, 2012
Example

Medications

☐ 1. Tylenol
☐ 2. Sudafed
☐ 3. AZT + side effects
☐ 4. Bactrim

Problem List

☐ 1. Headache
☐ 2. Sinus Infection
☐ 3. HIV positive
☐ 4. UTI

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.

Trimethoprim/sulfamethoxazole or co-trimoxazole (abbreviated SXT, TMP-SMX, TMP-SMZ or TMP-sulfa) is a sulfonamide antibiotic combination of trimethoprim and sulfamethoxazole, in the ratio of 1 to 5, used in the treatment of a variety of bacterial infections.

Adapted from J. Halamka, 2012
Example

Medications
- 1. Tylenol
- 2. Sudafed
- 3. AZT
- 4. Bactrim

Problem List
- 1. Headache
- 2. Sinus Infection
- 3. HIV positive
- 4. UTI

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.

Prophylaxis (preventative med) for immunocompromised patient?
Patient has urinary tract infection (UTI), plausibly deniable case.

Adapted from J. Halamka, 2012
Example

Medications
- 1. Tylenol
- 2. Sudafed
- 3. AZT
- 4. Bactrim

Problem List
- 1. Headache
- 2. Sinus Infection
- 3. HIV-positive
- 4. UTI

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.

Highest rate of same-sec couples in Provincetown, MA.


Adapted from J. Halamka, 2012
Example

Medications
- 1. Tylenol
- 2. Sudafed
- 3. AZT +side effects
- 4. Bactrim

Problem List
- 1. Headache
- 2. Sinus Infection
- 3. HIV-positive
- 4. UTI

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.

Candidiasis (thrush) - Candidiasis or thrush is a fungal infection (mycosis). Commonly causes mouth yeast infections, which manifest as white patches in the mouth. 15% of immuno-compromised patients may develop this.

Adapted from J. Halamka, 2012
Example

Medications
- 1. Tylenol
- 2. Sudafed
- 3. AZT + side effects
- 4. Bactrim

Problem List
- 1. Headache
- 2. Sinus Infection
- 3. HIV positive
- 4. UTI

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.

Headaches & HIV: 24/535 patients – 4.5%
CDC NHDS 2010 dataset. 115,000 patients.

Mononucleosis-like symptoms

Adapted from J. Halamka, 2012
Disorders

Treatments

Manifestations

Cause

Treat

Effects
Threat Model

- Attacker has direct access to redacted health record, medical literature
- Attacker does not have the computational capability to circumvent security mechanisms that protect the primary sensitive codes
<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
<th>Links</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risperidone</td>
<td>Treats schizophrenia, bipolar disorder, and autism.</td>
<td>schizophrenia, bipolar disorder, autism, weight gain, insomnia, alopecia</td>
<td>Use of Risperidone usually implies treatment of a mental health disorder.</td>
</tr>
<tr>
<td>Carbamazepine</td>
<td>Anti-convulsant and mood-stabilizing drug. Treats epilepsy and bipolar disorder.</td>
<td>epilepsy, bipolar disorder, headaches, drowsiness</td>
<td>Primarily used to treat mental health disorders. Could be used off-label to treat Complex regional pain syndrome (ICD9: 337.21)</td>
</tr>
<tr>
<td>Citalopram</td>
<td>Primarily used as an SSRI to treat depression. Can also be used to treat hot flashes.</td>
<td>depression, hot flashes, anorgasmia, nausea, diarrhea</td>
<td>Can treat both sensitive and non-sensitive conditions.</td>
</tr>
<tr>
<td>Lamotrigine</td>
<td>Primarily used as an anticonvulsant drug to treat epilepsy and bipolar disorder. Can also treat migraines.</td>
<td>epilepsy, bipolar disorder, migraines</td>
<td>Can be used to treat mental health disorders or migraines.</td>
</tr>
</tbody>
</table>
Formal model

**Hypothesis**

- \{d_1, d_3\}
- \{d_2, d_3\}
- \{d_1, d_2, d_3\}
- \{d_1, d_2\}

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

Hypothesis

\[ \{d_1, d_3\} \]

\[ \{d_2, d_3\} \]

\[ \{d_1, d_2, d_3\} \]

\[ \{d_1, d_2\} \]

Reggia’s set cover model
- Plausibility – set cover
- Likelihood – Occam’s razor and fitness
Explanation of manifestations

- Best explanation $E$ of manifestations:
  - Covers all observed manifestations $M^+$
  - Is the simplest (parsimonious) definition

- Heuristics for “best cover”
  - Minimality - $|E|$ is minimal
  - Criticism: minimal cardinality covers can be too restrictive
    - Occam’s razor vs Hickam’s dictum
  - Irredundancy – removing any disorder results in a non-cover of $M^+$
  - Relevancy – Every $d$ in $D$ must be causally associated with some $m$ in $M^+$
Medical concepts

- Diseases
  - AIDS
  - Schizophrenia

- Manifestations
  - Kaposi’s Sarcoma
  - Cervical Cancer
  - Stroke
  - Delusions
  - Alcohol Abuse
  - Memory Loss
  - Psychosis
  - Rotavirus
  - Alcohol Abuse
  - Delusions
  - Memory Loss
  - AIDS
  - Schizophrenia
  - Cervical Cancer
  - Stroke
  - Psychosis
  - Rotavirus
Delusion
Hallucination
Alcoholism
HIV+
Stroke
Memory loss

Psychosis
Alzheimer's Disease

Manifestations
Diseases

Source: PubMed, NIH.gov
Predicate-Reducer definition

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

\[ A(m) = A(R_{\pi}(m)) \quad \forall m \in M \]
Inference approach

- Input: Reduce(Diseases U Manifestations U Treatments)
- Output: Inferred Diseases

1. For each input, evoke hypotheses
2. Evaluate hypotheses
3. Rank hypotheses according to fitness

- Hypothesis fitness
  - Competing hypotheses, e.g. \( d_1 \) or \( d_2 \)
Algorithm overview

- **Salient Concepts**
- **Hypotheses**

- **Extract Concepts**
- **Retrieve Documents**
- **Extract and rank Hypo**
- **R(EHR)**
- **Docs**
To calculate this distribution, we take a set of documents from the base. Intuitively, this is done as a heuristic to maximize the number of documents returned. To generate the queries, we select query terms from the EHR with the hypothesis. We've seen that this greatly improves the ability of our algorithms to infer sensitive hypotheses. After this process is performed, the set of documents considered is expanded. In this example, expanding one level yields support from 25 documents from PubMed Open Access. E.g., if the query is restrictive, as only 25 documents from PubMed Open Access are available, this could be a competing solution, which may be synonyms, but upon closer inspection, it turns out that 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 manifestations of several medical conditions. To address this difficulty, we use probabilistic techniques from Section 4.3.

Algorithm overview

```plaintext
hypotheses ← ∅;
repeat
    query ← ∅;
    for j = 1 → numTerms do
        /* select a concept from the EHR using a probability distribution */
        x ← select_concept(concept_probs, EHR)
        query ← query ∪ x;
    end
    /* search for docs that contain the query terms */
    sr ← search(query, knowledge_base);
    /* Identifies hypotheses from medical concepts in documents */
    hypotheses ← update_hyp(hypotheses, sr);
    /* Evaluates hypotheses according to plausibility criteria */
    results ← eval_hypotheses(hypotheses) ∪ results;
until convergence;
rank(results);
```

Algorithm 1: Inference algorithm
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:

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

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

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

\[ HFI(H, Docs) = \sum_{doc \in Docs} CSI(H, doc) \cdot weight(doc, H) \]  

(3)

where \( weight(H, doc) \) is a weighting function. One such function could be BM25 [20, 30, 34], which is defined as

\[ BM25(D, Q) = \sum_{q_i \in Q} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|_{avgdpl}}{avgdl})}, \]  

(4)

where

\[ IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \]  

(5)

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Query</th>
<th>Results</th>
<th>Medical codes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rett Syndrome</td>
<td>“wringing” AND “female” AND “constipation” AND “scoliosis”</td>
<td>3 articles suggest Rett Syndrome.</td>
<td>F84.2, R09.0, K59.0, 737.0</td>
<td>Pubmed</td>
</tr>
<tr>
<td>Rett Syndrome</td>
<td>“wringing” AND “female” AND “constipation” AND “scoliosis”</td>
<td>1.73M results, 5 of top 10 results suggest Rett Syndrome, including NIH Medline.</td>
<td>F84.2, R09.0, K59.0, 737.0</td>
<td>Google</td>
</tr>
<tr>
<td>AIDS</td>
<td>&quot;Toxoplasmosis&quot; AND &quot;Hepatitis B&quot; AND &quot;Encephalopathy&quot; AND &quot;Progressive multifocal leukoencephalopathy&quot; AND &quot;Cryptococcosis&quot;</td>
<td>140,000 results. 5 of top 10 suggest AIDS.</td>
<td>130, 070.2, 348.30, 046.3, 117.5</td>
<td>Google</td>
</tr>
<tr>
<td>AIDS</td>
<td>...</td>
<td>18,000 results. &gt;8 of top 10 suggest AIDS.</td>
<td>130, 070.2, 348.30, 046.3, 117.5</td>
<td>Bing</td>
</tr>
</tbody>
</table>
Possible defenses

Deniability through relative strengths of hypotheses

- Hide non-sensitive EHR as well
- Enhance competing hypothesis, e.g. Citalopram
- Association rule hiding
A message from our sponsors...

We thank:

- Carl Gunter and Mike Berry - predicate-reducer model
- James Reggia - formalization of the hypothetico-deductive model
- Brad Malin - helpful resources
- Ivan Handler - health information exchange level
- Fisayo Ositelu - medical insight.
Questions?
Ask your doctor!