DECISION SUPPORT FOR DATA SEGMENTATION (DS2)

SHARPS.org

Presented by Carl Gunter and Mike Berry

May 20, 2014

Strategic Health IT Advanced Research Projects (SHARP)

- **SHARP Area 1** – Security and Health Information Technology (SHARPS), University of Illinois
- **SHARP Area 2** – Patient-Centered Cognitive Support (SHARPC), University of Texas at Houston
- **SHARP Area 3** – Health Care Application and Network Design (SMART), Harvard University
- **SHARP Area 4** – Secondary Use of EHR Information (SHARPN), Mayo Clinic of Medicine
- **NIH Affiliate** – Medical Device "Plug-and-Play" Interoperability Program (MDSHARP), Massachusetts General Hospital, supported by NIH/NBIB Quantum Grant

The Six Challenges

SIX RESEARCH CHALLENGES

FOR THE SECURITY AND PRIVACY OF HEALTH INFORMATION TECHNOLOGY

1. Access controls and audit
2. Encryption and trusted base
3. Automated policy
4. Mobile health (mHealth)
5. Identification and authentication and
6. Data segmentation and de-identification

Challenge #6: Data Segmentation & De-Identification

- Patients feel that some types of health data are especially sensitive: records related to mental health, drug abuse, genetics, sexually transmitted diseases, and others
- There is a desire to transmit this type of information only when necessary
- Data segmentation: breaking the EHR into parts
- This is a hard problem in many cases
  - History with de-identification (segmenting the personally identifying parts of the record)
  - Case study: HIV
- How do we balance feasibility, privacy, and clinical impact?
- Research challenge: how to segment and measure

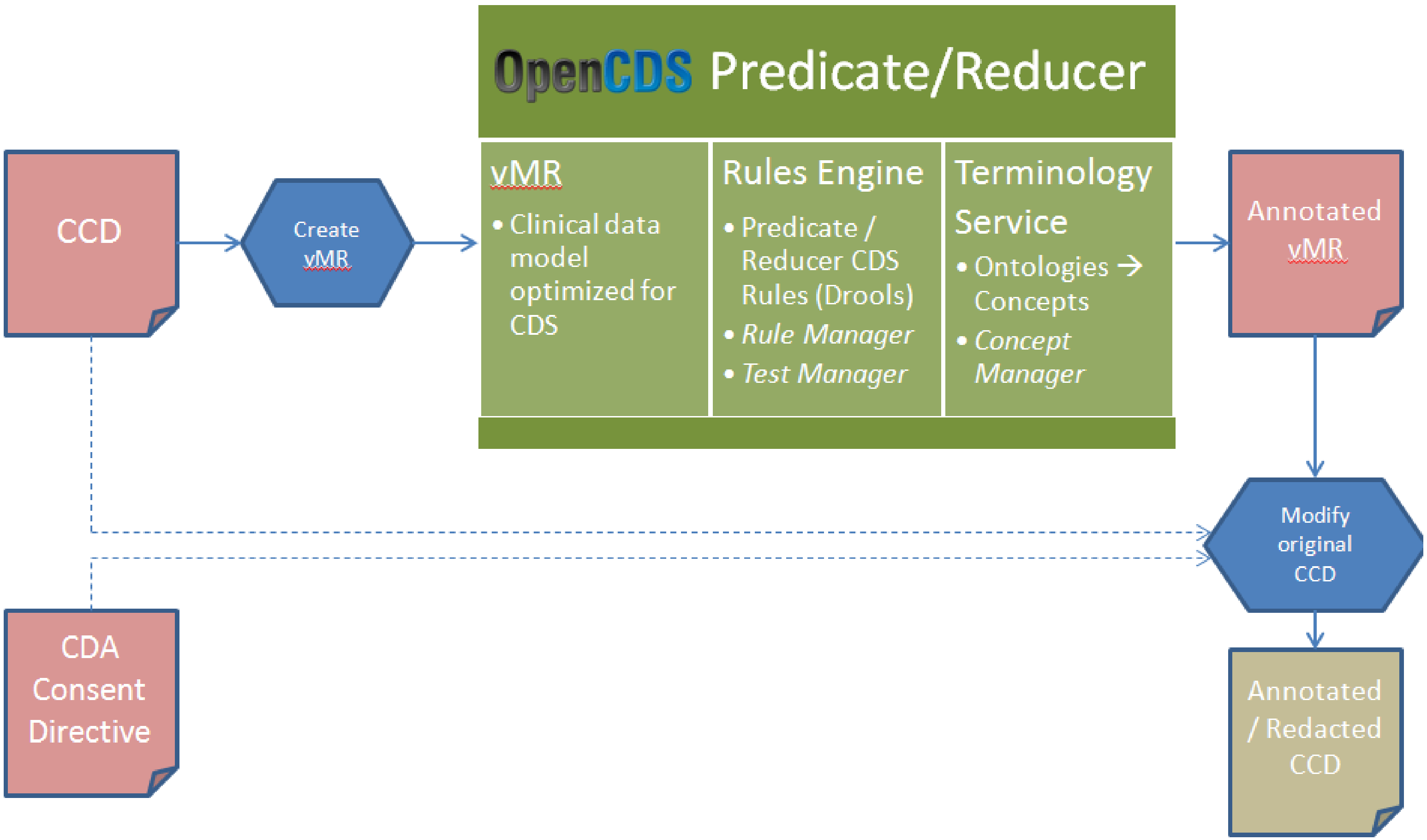
DS4P, DS2, and ILHIE Prototype

- Data Segmentation for Privacy (DS4P):
  - ONC S&I Framework initiative
  - Published use cases and an Implementation Guide
  - Sponsored pilot projects to demonstrate the effectiveness of standards-based data segmentation in the context of the use cases
- Decision Support for Data Segmentation (DS2):
  - Research-oriented project to address one of the challenges in certain types of segmentation:
    - Sequestration of a condition and its related clinical facts sometimes leaves clues – residual facts such as comorbidities and co-occurrences – that could still reveal the condition to an informed observer
    - DS2 views this as a decision support problem, that is, to treat inference as a function that can be specified by a computer program based on information like machine learning over a body of records.
- ILHIE Prototype: A prototype implementation of DS2 for an HIE architecture.

DS2 Concepts

- Predicate:
  - Identify if a clinical document has a particular type of sensitive data in it
- Reducer:
  - Redact portions of the clinical document until corresponding predicate is satisfied
- Safety Checker:
  - Check care plan against non-redacted clinical document

Predicate/Reducer Architecture



Predicate/Reducer Example

SHARPS

Problem List	Medication List
• HIV infection • Candidiasis of lung • Bacterial infection, unspecified	• Combivir • Norvir • Procrit • Azithromycin • Fluconazole

1. An HIV predicate might evaluate the CCD and return "True" (an HIV condition is present in the document).
2. A simple HIV reducer might remove the HIV diagnoses and medications (HIV infection, Combivir, Norvir).
3. The HIV predicate might evaluate the newly redacted CCD and return "True" again because of the co-occurrent clinical facts remaining in the record which might lead an observer to infer that the patient has HIV .
4. A more complex HIV reducer might continue to remove co-occurrent clinical facts until the predicate returns "False." Ultimately, in addition to the HIV, Combivir, and Norvir, it might also remove the Candidiasis, Fluconazole, and Procrit – even though none of them are directly related to HIV.

DS2 Focus and Outcomes

SHARPS

- Our DS2 study focused primarily on the development of predicates and reducers based on machine learning over databases of (portions of) EHRs.
  - For instance, we used such records from the problem lists of the Northwestern Memorial Hospital and from sources like the CDC National Hospital Discharge Survey.
- The work is described in detail in a pair of reports:
  - One focused on contextual integrity (policy) considerations
  - The other focused technical and architectural considerations, like how to realize predicates and reducers on top of the OpenCDS system.

Inference Analyzer

SHARPS

Look up/refresh this patient | Add new concept

Look up random patient: Random Patient with HIV infection

Probability of...

5: HIV infection

85%

Data: Fictional; Predicate: Bayesian Network trained on NMH problem list data

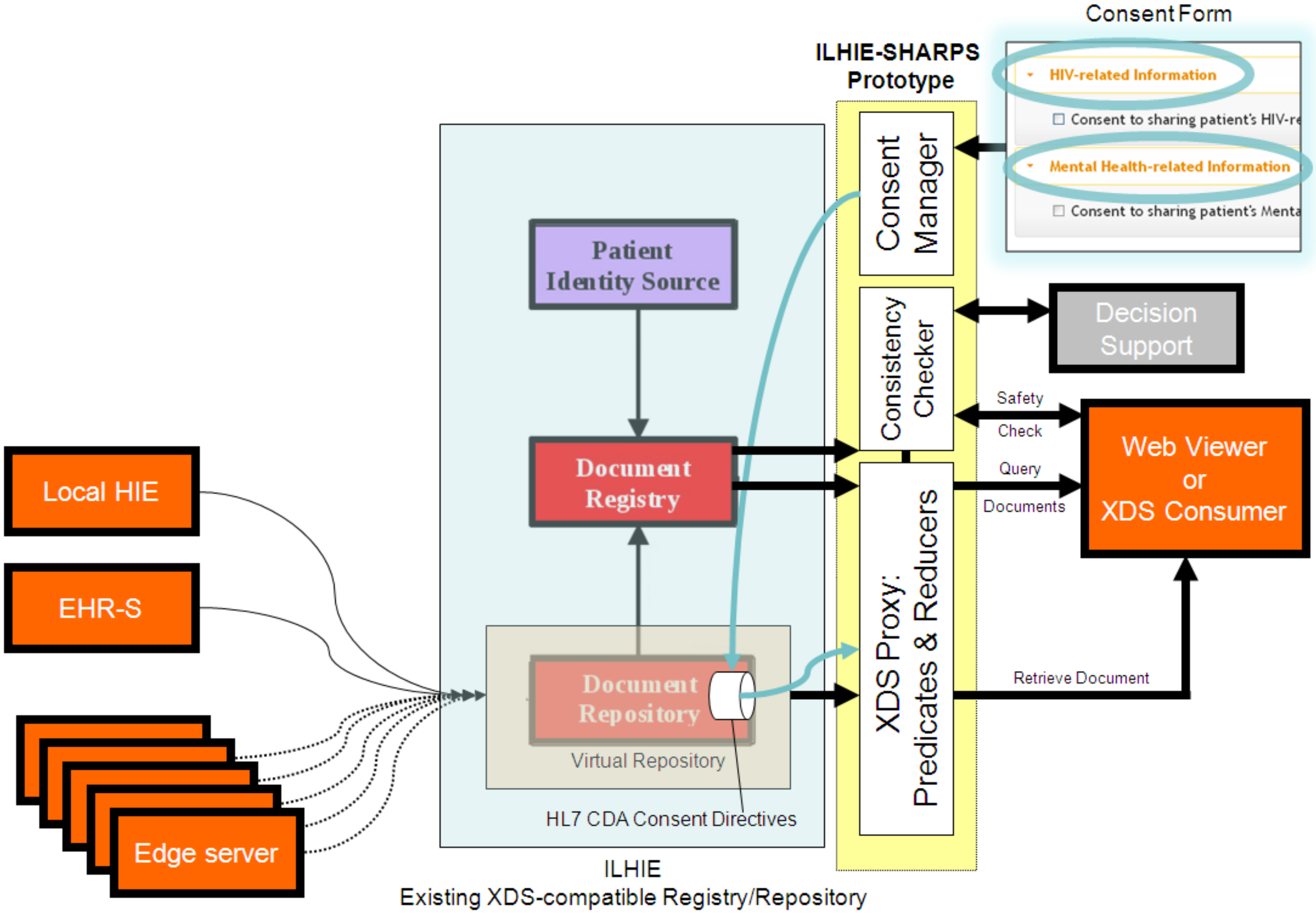
Patient Problem List (Fictional Patient)

ICD9 Code	ICD9 Description	CCS Category	Relative Risk	Phi Coefficient	Probability if redacted	Redact	Edit
042	Human immunodeficiency virus [HIV] disease	HIV infection (5)					
091.50	Syphilitic uveitis, unspecified	Sexually transmitted infections (not HIV or hepatitis) (9)	10.870	0.046	16.9%	Redact	
070.33	Chronic viral hepatitis B without mention of hepatic coma with hepatitis delta	Hepatitis (6)	8.504	0.069	66.4%	Redact	
682.9	Cellulitis and abscess of unspecified sites	Skin and subcutaneous tissue infections (197)	2.532	0.011	55.7%	Redact	
054.71	Visceral herpes simplex	*Herpes simplex (3006)	2.387	0.014	74.5%	Redact	
296.36	Major depressive affective disorder, recurrent episode, in full remission	Mood disorders (657)	2.127	0.022	80.2%	Redact	
401.9	Unspecified essential hypertension	Essential hypertension (98)	1.080	0.003	85.1%	Redact	

Results

- We compared a rules-based reducer utilizing an ontology with one based on DS2 machine learning inference using 150 randomly chosen NMH problem lists with HIV, and observed:
  - 78% had no redactions beyond HIV itself with either reducer (asymptomatic).
  - 7% of HIV patients had redactions beyond HIV itself that were the same with both reducers (HIV case definition)
  - 11% of HIV patients were more heavily redacted in the DS2 reducer (HIV case definition + common co-occurrences such as hepatitis)
  - 4% of HIV patients were more heavily redacted in the rules-based reducer (HIV case definition with low correlation such as herpes simplex)

ILHIE Prototype Architecture



Challenges

- Unstructured text
- Explaining probabilistic / machine learning-based redaction to patients and doctors
- Data quality
- Ability of observers to infer specific conditions based on clinical context may vary across different types of providers, levels of training, and medical specialties.
- Observer may have access to additional context not evaluated by a predicate, such as a medical record obtained prior to, or after, the predicate evaluation.

Other uses of DS2

SHARPS

- In addition to privacy-oriented segmentation, Predicate/Reducer technology could potentially be used for:
  - Providing a ***clinical summary evaluation tool*** to suggest possible "missing" diagnoses on patient records for quality review or for point-of-care clinical decision support.
  - Providing an ***inverse reducer*** that applies a Predicate/Reducer to filter a record to retain *only* the conditions related to a target condition.
    - Send diagnoses and medications to public health (or clinical research studies) related only to specific conditions
    - Assist physicians when generating referrals to specialists in order to highlight subsets of the clinical summary most relevant to a particular specialty or referral reason.

Learn More about DS2

- Website: http://sharps-ds2.atlassian.net/
- In addition to the prototype, the open source DS2 software repository includes a suite of related tools for:
  - working with clinical documents (CCDs and vMRs)
  - training classifiers and calling them from OpenCDS (open source clinical decision support framework)
  - checking CCDs for drug-drug interactions
  - visualizing the effectiveness and the impact of probabilistic redaction through a web-based "inference analyzer.“