EMERSE at UCSF

Unlocking our De-Identified Unstructured Data!

Leslie Yuan
UCSF Clinical & Translational Science Institute
October 19, 2021

A bit of history

2016

Conversations with David H about EMERSE

2019

UCSF's algorithm (Philter) for machine-redaction (M-R) of notes is ready

Feb: UCSF will use EMERSE with M-R notes

Apr: Sub-award received from UMich for UCSF EMERSE Team

2020

Apr: Decision to launch an additional COVID- EMERSE instance

Apr-May: EMERSE is running for internal QA

Jul: Launch!
EMERSE & COVID-EMERSE

2021

Feb: Campus-wide webinar

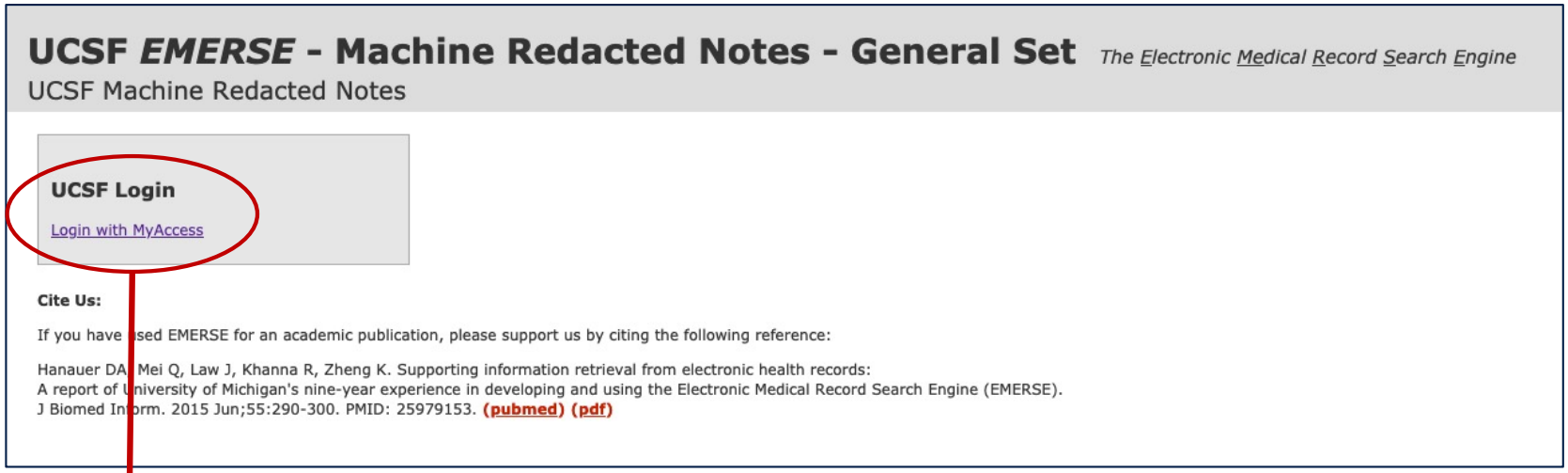
Nov-Dec: Re-load **certified** de-id notes into EMERSE

No IRB approval required for use!

De-identified Notes at UCSF + EMERSE

105 million clinical notes for over 2.4 million patients

and the COVID-19 Set (not fully de-id) has 260K+ patients



Integrated with UCSF Single-Sign-On ☺

Access to the Solr API

In addition to the EMERSE UI, UCSF users have access to the Solr API

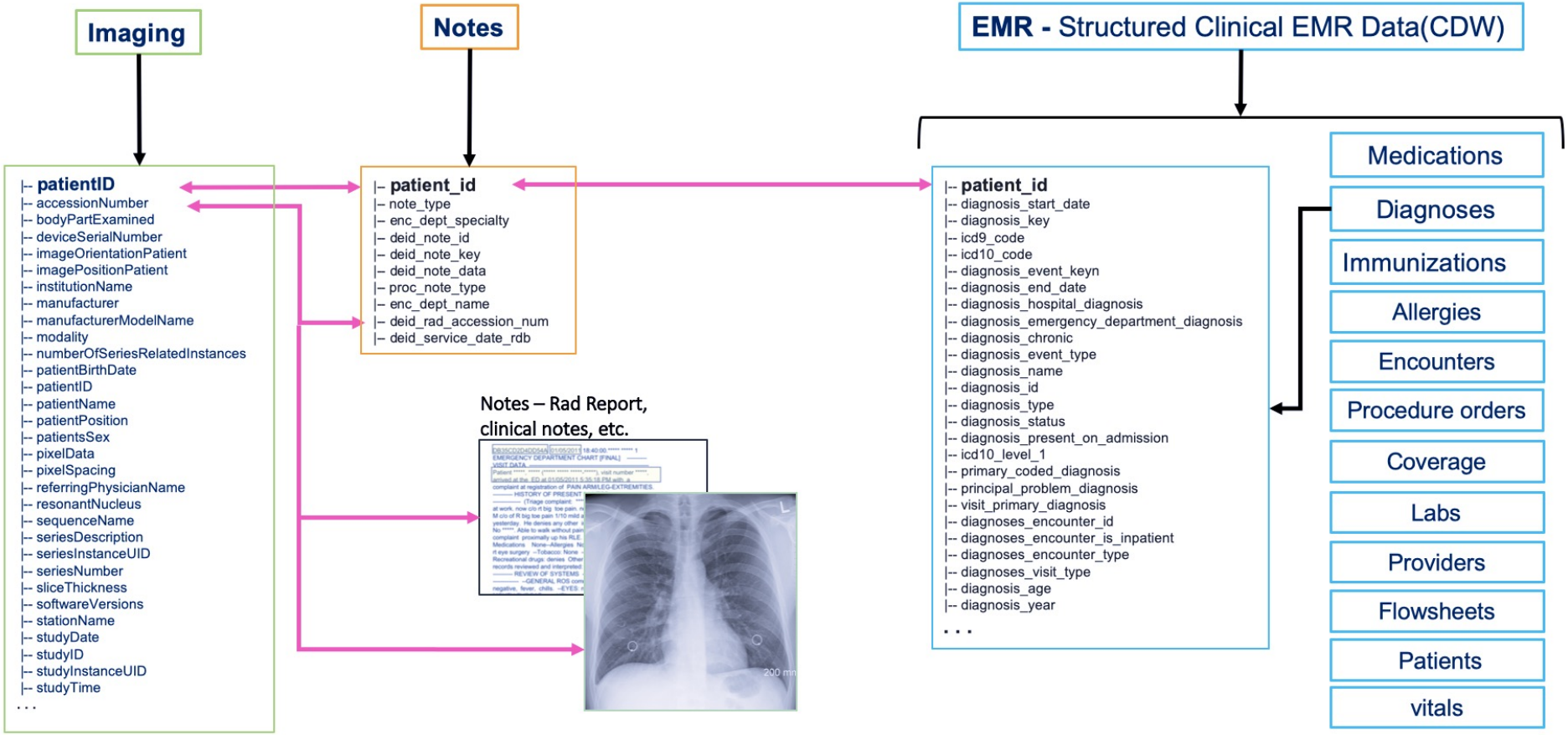
- UCSF always planned to give direct programmatic access to de-id notes to support research
- Researchers were not able to access the notes via the original methods we were considering
- Solr provides efficient and flexible text search and retrieval functionality through programmatic access

About UCSF's de-identified EHR data

- Same identifiers are used across all UCSF de-identified EHR data assets
  - Structured data
  - Unstructured data*
  - Imaging data*
- Multiple Query Tools allow import / export of patient identifiers
  - Allows non-technical and technical researchers to easily look at same patients across datasets

* *De-id certification in progress; notes and image subsets available by year-end 2021 to enable use without IRB.*

Traversing UCSF's de-identified EHR data

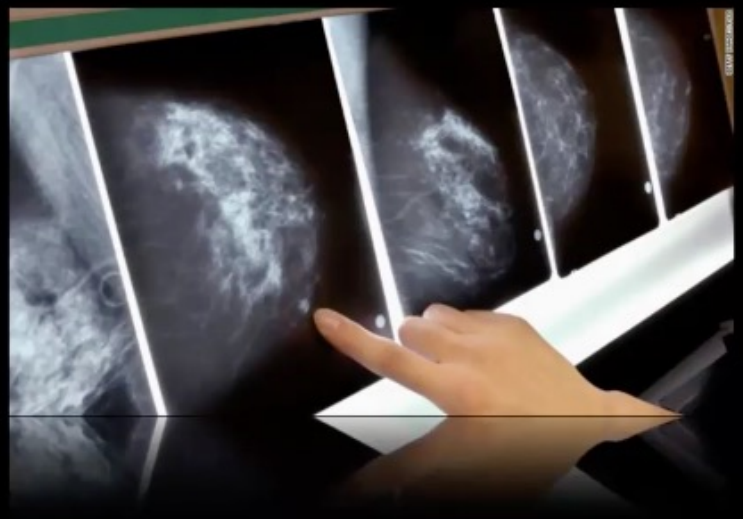


How EMERSE has benefited UCSF Research Teams

1. **Preliminary data review.** Easily review representative notes up front to assess whether information of interest is present and how it is encoded so we could then write code to programmatically pull it in bulk
2. Quickly see the **distribution of relevant terms**. The color coding permitted us to see distributions by how many colors showed up on the summary page
3. **QC during data review.** QC of training data for deep learning models. Data discrepancies were easily reviewed by pulling up notes to pinpoint areas needing attention
4. Enabled easy **self-serve check** on what de-id data are available before going to the trouble of pulling data from additional *identified* data sources.

Breast BI-RADs Research Use Case using EMERSE

Goal: Automatic BI-RADS classification from MRI images



A.I.

BI-RADS	Findings
0	**Incomplete** assessment
1	**Negative**
2	**Benign**
3	**Probably benign** (< 2%)
4	**Suspicious** (risk: 2-94%)
5	**Probably malign** (risk > 94%)
6	**Malign** (biopsy proven)

UCSF Department of Radiology and Biomedical Imaging

ci² center for intelligent imaging

ic² intelligent imaging computational core

How is EMERSE helpful in the BI-RADS study?

Easy, fast, visual validation of ”terms” used in notes; enables confident downstream queries for export of data to be used in deep learning models

1. From UCSF De-ID Clinical Data Warehouse, identify unique patients with unique accession numbers referring to MRI exams

2. **Import** the patient IDs into EMERSE to see associated De-ID notes
3. **Validate and confirm** via EMERSE that BI-RAD scores are in the notes
    1. Whether terms are used, where and how used
    2. Colored visuals using **term bundle shows distributions of**: “BI-RAD 0 – BI-RAD 6”

4. Once term use is confirmed, direct query against notes data using Solr API allows export of data needed for training a deep learning model

Additional areas of research enabled by EMERSE @ UCSF

- Dentistry
  - “Orofacial Injuries/Intimate Partner Violence” – Dr Sepideh Banava
- COVID-19 outcomes
- Interstitial lung disease
- Ophthalmology
- Nursing
- Oncology

[We could] kickstart the study by having a fast, self-serve interface for looking at the data.

If you don't even know what's in a data set, it's risky to invest the time to pull it and write code to extract it. This is a really critical part of any study ramp up IMO ... looking at the data.

UCSF Researcher,
Dept of Radiology & Biomedical Imaging

Thank you to the University of Michigan EMERSE Team!

From the UCSF EMERSE Team


Eric


Leslie


Oksana


Lakshmi


Gundolf


Sharat


Jason


Dima

UCSF

University of California
San Francisco