Adapting clinical NLP methods for multi-site medical products research

David S. Carrell, PhD
Kaiser Permanente Washington Health Research Institute

June 15, 2017, FDA White Oak Campus, Silver Spring, MD

Overview

- Secondary use of EHR data & text
- Accessing clinical text
- Interpretive challenges
- Multi-site clinical NLP implementation

Secondary use of EHR data & text

- Reusing data/text for some other purpose
- Seemingly simple tasks can be challenging

Secondary use of EHR text

Electronic exhaust of healthcare delivery

- Intended use:
  - Care of *one patient*
  - Through an EHR interface
  - By *clinicians*
  - In one healthcare system
  - In context of other encounters
  - Limited access (privacy)
- Use for purposes not originally intended may be challenging

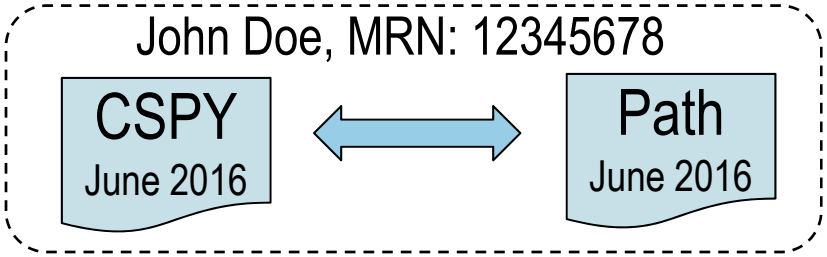
KAISER PERMANENTE®

Secondary use of EHR text

Electronic exhaust of healthcare delivery

- Information needed for patient care
  ≠ information needed for research

  Example: Does a patient have pre-cancerous colon polyps?



  Clinical: Easily resolved in a patient chart.

- **Research**: Meta data *do not link* pathology reports to their colonoscopy procedures
  - *Probabilistic* matching

KAISER PERMANENTE®

Secondary use of EHR text – Pathology reports

Electronic exhaust of healthcare delivery

- Finding colonoscopy-related pathology reports requires NLP

  Rule: Report contains ≥1 of:

  ```
  anal verge, ascending colon, ascending mass, ascending
  polyp, Cecal, Cecum, Colon, Colonic, descending colon,
  ic valve, ileocecal valve, ileum, rectal, rectosigmoid,
  rectum, sigmoid, sigmoid polyp, splenic flexure, terminal
  ileum, transverse colon, transverse polyp
  ```

- *Most* (not all) colonoscopy and pathology reports have same date

Pathology report counts by number of days before/after colonoscopy procedure date.											
Pathology date	Days before CSPY					Same date	Days after CSPY				
	-5-14	-4	-3	-2	-1		+1	+2	+3	+4	+5-14
N Path Rpts	7	1	3	0	7	1630	50	52	29	4	3

18 92% 138 (8%)

KAISER PERMANENTE®

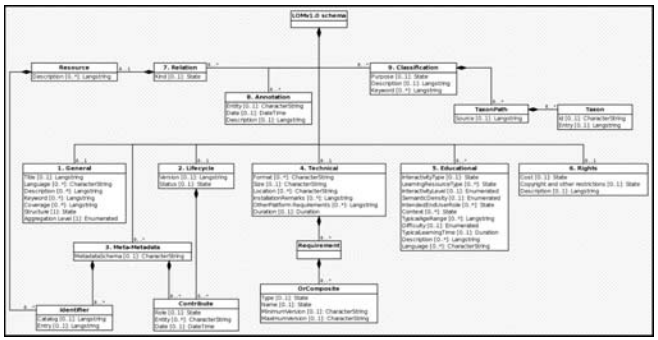
Secondary use of EHR text – "duplicate" notes

It's like déjà vu all over again --Yogi Berra

- As the legal medical record *every version of* a note is preserved
  - Every edit/save generates a new copy
  - EHR displays only the most recent version (primary use)
  - Tricky to de-duplicate in the back end database (secondary use)

3,025 ≥ 5 versions

Versions per clinical note (MSE corpus)		
N versions	**Count**	**Percent**
1	1,594,182	82%
2	107,412	6%
3	26,629	1%
4	7,151	0%
5	2,027	0%
6	618	0%
7	224	0%
8	73	0%
9	28	0%
10	13	0%
11	13	0%
12	10	0%
13	3	0%
14	6	0%
15	2	0%
18	2	0%
20	2	0%
27	2	0%
45	1	0%
53	1	0%
	1,934,518	



KAISER PERMANENTE®

Accessing clinical text

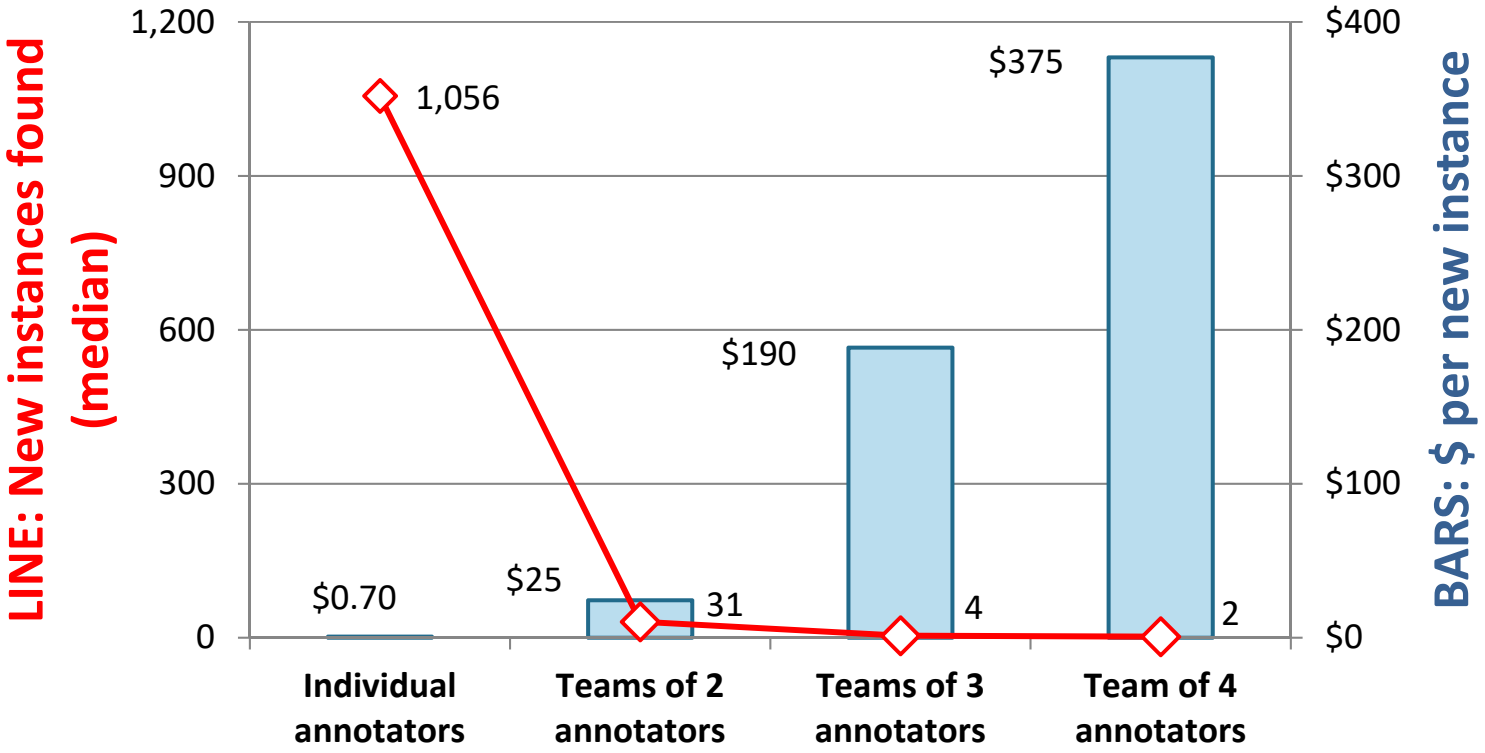
- Need for privacy, de-identification
- Text availability
- Text degradation

Accessing clinical text

- All clinical text may contain highly sensitive information
  - Unlike structured data, can't just suppress sensitive content
  - Risks to patient privacy
  - Institutional risk
- Clinical text seldom exits local firewall
  - Lack of freely available corpora impedes technical progress, collaboration
- Text de-identification
  - 94-98% effective
  - Requires technical sophistication, knowledge
  - Added cost

Accessing clinical text – de-identification cost

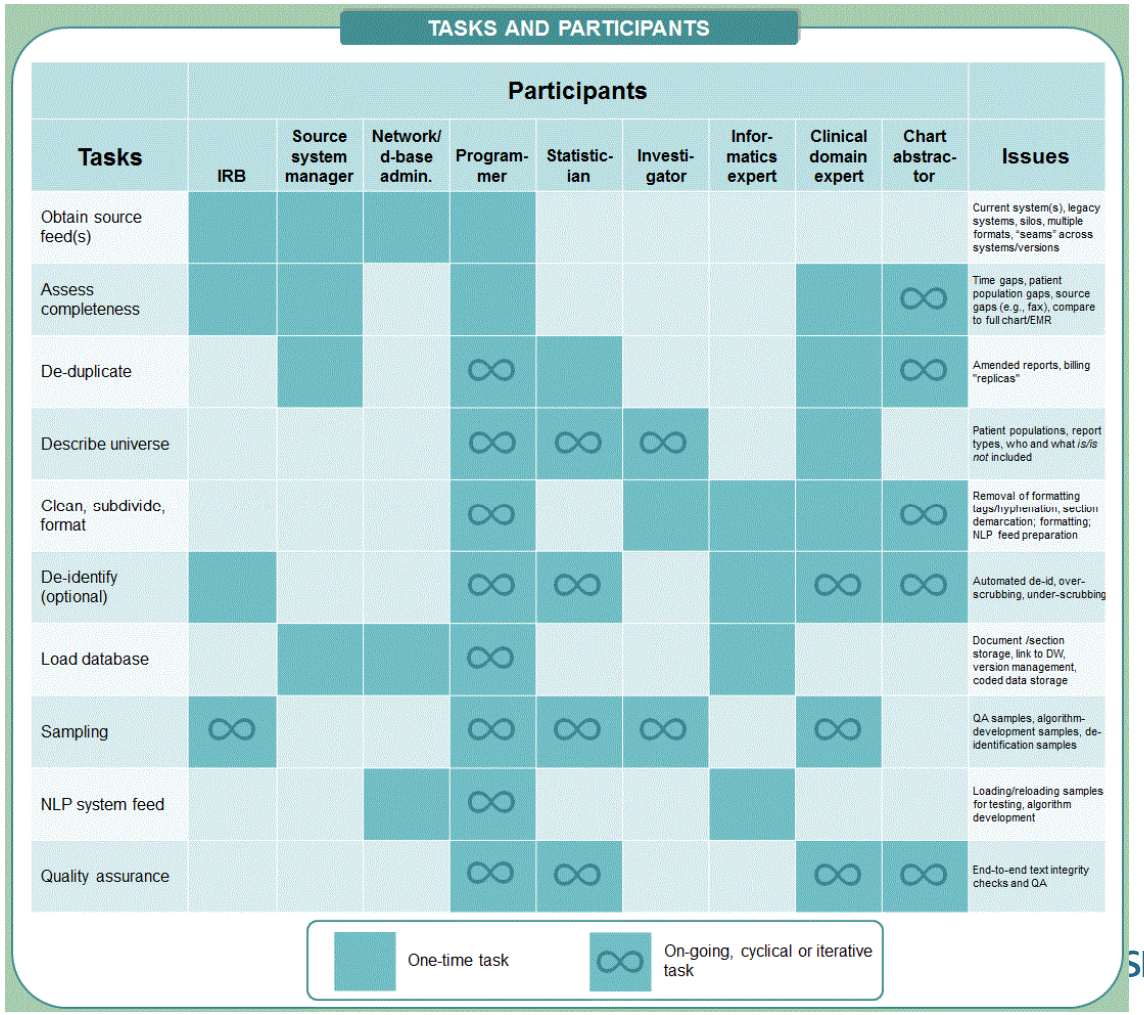
- Cost of manually annotating personally identifiable information (PII)
- 100 Family Practice notes (1,093 PII instances)



KAISER PERMANENTE®

Accessing clinical text – The process

- Corpus assembly: multi-step process highly reliant on local expertise



Accessing clinical text – What text exists?

- Availability of clinical text varies by site
- Study of prescription opioid abuse/addiction/overdose in four sites

KPNW	Optum	TennCare	KPW
▪ Epic EHR ▪ Outpatient ▪ Hospital, ER ▪ In-house substance use treatment	▪ Subset of patients w/ EHR ▪ Outpatient ▪ NLP-extracted term lists ▪ Claims data	▪ Epic EHR ▪ Outpatient ▪ Hospital, ER ▪ At *one* location (VUMC)	▪ Epic EHR ▪ Outpatient clinics ▪ ER (partial)

Accessing clinical text – degraded text

EHR version

SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son John

Today she was in usual state of health. was going to a restaurant, and on way in "just hit her suddenly" with double vision.

{deletions}

OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2 97% Estimated Body mass index is 21.92 kg/(m^^2)

{deletions}

LABS / STUDIES:
HEMATOCRIT (%)
Date Value Low High Status

NENTE®

Accessing clinical text – degraded text

The Worst You Ever Gave Me Was the Best I Ever Had-- Frank Sinatra

EHR version	Secondary use database version
SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son John Today she was in usual state of health. was going to a restaurant, and on way in "just hit her suddenly" with double vision. {deletions} OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2 97% Estimated Body mass index is 21.92 kg/(m^^2) {deletions} LABS / STUDIES: HEMATOCRIT (%) Date Value Low High Status 3/21/2011 46 34 49 Final 3/29/2010 41 34 49 Final 6/18/2010 39 34 49 Final 10/4/2007 47 34 49 Final 8/20/2008 41 34 49 Final 8/23/2007 37 34 49 Final {deletions} ASSESSMENT / PLAN: 1. Double vision - acute onset in patient with HTN, COPD and Hx TIA. No other symptoms of posterior circulation issues but does have acute double vision worrisome for posterior circulation. I didn't see notes of prior aniscoria but given overall mental status ? more related to cataracts. Recommend ER eval ASAP for TIA / CVA Spoke with Barb, RN at Pleasant Valley Medical Center ER and gave heads up 2. HEADACHE - mild. To ER for imaging Christine Jones, M.D. (123) 456-7890 (123) 456-7891 FAX	SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son John Today she was in usual state of health. was going to a restaurant, and on way in \"just hit her suddenly\" with double vision. {deletions} OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2\r\n97% Estimated Body mass index is 21.92 kg/(m^^2) {deletions} LABS / STUDIES: HEMATOCRIT (%) Date Value Low High Status 3/21/2011 46 34 49 Final 3/29/2010 41 34 49 Final 6/18/2010 39 34 49 Final 10/4/2007 47 34 49 Final 8/20/2008 41 34 49 Final 8/23/2007 37 34 49 Final {deletions} ASSESSMENT / PLAN: 1. Double vision - acute onset in patient with HTN, COPD and Hx TIA. No other symptoms of posterior circulation issues but does have acute double\r\nvision worrisome for posterior circulation. I didn't see notes of prior aniscoria but given overall mental status ? more related to cataracts. Recommend ER eval ASAP for TIA / CVA Spoke with Barb, RN at Pleasant Valley Medical Center ER and gave heads up 2. HEADACHE - mild. To ER for imaging Christine Jones, M.D. (123) 456-7890 (123) 456-7891 FAX

Accessing clinical text – degraded text

The Worst You Ever Gave Me Was the Best I Ever Had-- Frank Sinatra

EHR version	Secondary use database version
SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son John Today she was in usual state of health. was going to a restaurant, and on way in "just hit her suddenly" with double vision. {deletions} OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2 97% Estimated Body mass index is 21.92 kg/(m^^2) {deletions} LABS / STUDIES: HEMATOCRIT (%) Date Value Low High Status 3/21/2011 46 34 49 Final 3/29/2010 41 34 49 Final 6/18/2010 39 34 49 Final 10/4/2007 47 34 49 Final 8/20/2008 41 34 49 Final 8/23/2007 37 34 49 Final {deletions} ASSESSMENT / PLAN:	SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son John Today she was in usual state of health. was going to a restaurant, and on way in \"just hit her suddenly\" with double vision. {deletions} OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2\r\n97% Estimated Body mass index is 21.92 kg/(m^^2) {deletions} LABS / STUDIES: HEMATOCRIT (%) Date Value Low High Status 3/21/2011 46 34 49 Final 3/29/2010 41 34 49 Final 6/18/2010 39 34 49 Final 10/4/2007 47 34 49 Final 8/20/2008 41 34 49 Final 8/23/2007 37 34 49 Final {deletions} ASSESSMENT / PLAN: 1. Double vision - acute onset

Accessing clinical text – degraded text

I never said most of the things I said --Yogi Berra

- Progress note in EHR:

  COPD oxygen and steroid dependent on oxycontin and oxycodone for dyspnea and pain follow up in 1 week

- Back-end database version:

  COPD oxygen and steroid dependent on oxycontin and oxycodone for dyspnea and pain follow up in 1 week

- False positive NLP hit in abuse/addiction study

Interpretive challenges in clinical text

- The synonymy problem
- The copy-and-paste problem
- The freedom problem in free-text

Interpretive challenges: *Synonymy*

I know it when I see it -- Justice Potter Stewart

- Example: *Clostridium difficile infection* ("c diff")
  - 1 t
  - 6 r

Negatic
Affirmati
Affirmati
Affirmati
Negate
Negate
dil diarrhea

Negation Status of NLP-Extracted Mammogram Findings

NLP-Extracted Finding	Negated	Affirmed	Total
mass or lump	8,691 (95%)	453 (5%)	9,144
calcification	6,739 (87%)	1,046 (13%)	7,785
architectural distortion	1,875 (78%)	542 (22%)	2,417
asymmetry	100 (17%)	486 (83%)	586
distortion	137 (66%)	71 (34%)	208
Total	17,542 (87%)	2,598 (13%)	20,140 (100%)

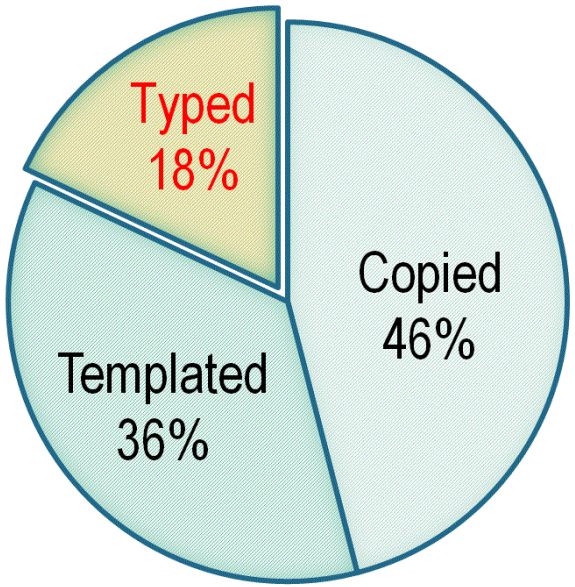
	N	%
	1	0%
	1	0%
	1	0%
	1	0%
	1	0%
	1	0%
	1	0%
is	1	0%
	1	0%
	1	0%
	1,241	100%

KAISER PERMANENTE®

Interpretive challenges: Copy-and-paste

I never said most of the things I said --Yogi Berra

- EHR text has a copy-and-paste problem
- Most valuable text is manually entered
- <20% of a note's text is manually typed
  - Wang, JAMA 2017
- The text we most care about is written:
  - In a hurry
  - While multi-tasking
  - Usually without editing
  - Often without punctuation



KAISER PERMANENTE®

Interpretation: Medication side effect mentions

KPW medication side effects project (underway)

- 89,377 patients with 113,564 new anti-depressant medication episodes
- 21,602 patients with 25,439 new anti-psychotic medication episodes
- 2005-2016
- Goal: Capture side effects described in Family Practice and Behavioral Health notes

Interpretation: Medication side effect mentions

The importance of context

Found in the notes of patient starting a new anti-depressant:

- "Zoloft caused increased anxiety"
- "feels panicked"
- "not able to sleep"
- "better experience with newer medication (Lexapro)"

DAY 0

SUBJECTIVE: … [NN] year old [MALE/FEMALE] with recurrent depression … Early AM wakening ... no energy … Hx of depression in past with poor medication experience. Tried Imipramine with bladder sx, Zoloft caused increased anxiety (felt unable to cope and poor sleep), Trazadone didn't help with depression. Sister coping with depression with better experience with newer medication (Lexapro) …

ASSESSMENT/PLAN: Depression, severe … start with Celexa

Day 4

SUBJECTIVE: ... feels panicked by not able to sleep …

ASSESSMENT/PLAN: Acute sleep disturbance … add short term Alprazolam for acute anxiety and sleep

Day 18

SUBJECTIVE: ... checking back on depression/anxiety

ASSESSMENT/PLAN: Depression/anxiety improving. Continue current dose of Celexa … Alprazolam

Day 39

SUBJECTIVE: … Feels like has returned to "normal"

KAISER PERMANENTE®

Interpretation: Typical side effect mentions

- Attribution?
- Severity?
- Rare MSEs?

I am glad the nortriptyline is helping with your pain and burning Since it also seems to be causing severe dry mouth I recommend decreasing your dose …

stopped the Sertraline a week ago secondary to worsening SI after increasing the dose

He had difficulty achieving orgasm on fluoxetine.

Last visit changed from Prozac to Wellbutrin. Decreased libido is not better.

He is experiencing a diminished libido but says this is tolerable right now.

is concerned about **weight gain**

less anxious … Happy about **weight gain**.

Was on higher dose but was causing side effects

SE: Unable to cry; early a.m. Awakening; occasionally feels nauseated

Interpretation: Abbreviation SE (side effect)

Word sense of "SE" in a sample of 50 clinic notes randomly selected from 18,542 notes containing "SE" 2005-2016.[1]

Sense of "SE"	Mentions	Percent	Estimate in corpus
Clinic address (quadrant)	17	34%	6,304
Patient address (quadrant)	16	32%	5,933
Hypothetical MSE[3]	10	20%	3,708
MSE[4]	3	6%	1,113
MSE confirmed abse			
Misspelling[6]			
Spanish pronoun[7]			

(1) The 18,542 notes were ^
medications during the 180

"started on fluoxetine … Mood is somewhat better, however … notes significant spaciness and inability to concentrate at work … symptom occurred in the first week, but has worsened on the higher dose … ASSESSMENT: major depression on fluoxetine with significant SE unlikely to resolve PLAN: Switch to citalopram"

KAISER PERMANENTE®

Interpretation: *Hypothetical* mentions

- Hypotheticals are 3-5 times more common than actual side effect mentions

Dry mouth is a common side effect of many prescription and nonprescription drugs

You question if metoprolol causes dry mouth, this is possible but unlikely

Potential side effects could include but not limited to: fatigue, mild to severe somnolence, increased appetite/potential weight gain, dizziness, orthostatic hypotension, elevated liver function tests, or paradoxical reactions such as insomnia, anxiety, agitation, panic attacks, insomnia, irritability, hostility, or worsening depression or suicidal ideation.

If side effects are tolerable patient will continue Rx and they will likely abate

Benefits, risks, SE, alternative of no meds yes

Interpretation: List-style reporting

Benefits, risks, SE, alternative of no meds yes

EHR:

ASSESSMENT:

[redacted]

PLAN:

[redacted]

-Symptom(s): see subjective portion of note and above
-Goal(s): reduction in symptoms and improved functioning
-Intervention: psychotropic medication, also see instructions

Benefits, risks, SE, alternative of no meds yes
Rx course & natural history w & w/o treatment yes
Review studies available yes
Refill policy, Stimulant policy no
Patient resources provided yes

KAISER PERMANENTE®

Interpretation: List-style reporting

- Checklist style (is the "X" *before or after the* symptom?)
- Note misspelled "abominal"
  - In original (and 591 other notes since 2011)
  - Spelling distance algorithms may help (e.g., allow 1 missing letter)
  - But, allowing 1 missing letter: "addiction [to] opioids" = "addition [of] opioids"

ROS (x indicates positive finding) (Remainder of comp review of systems is negative except as noted in HPI) [ ] Fatigue, weakness [ ] Loss of appetite [ ] Weight loss [ ] Weight gain [ ] Fever [ ] Night sweats [ ] Trouble swallowing [ ] Heartburn [ ] Indigestion [ ] Nausea, vomiting [**x**] abominal pain, discomfort [ ] Change in bowel habits [ ] Constipation [ ] Diarrhea [ ] Blood with bowel movement [ ] Fecal incontinence [ ] Eye problems [ ] Nose bleeds [ ] Mouth sores [ ] Cough [ ] Shortness of breath [ ] Excessive snoring, sleep apnea [ ] Chest pain [ ] Irregular heartbeat or palpitations [ ] Swelling in ankles [ ] Confusion [ ] Numbness [ ] Depression or anxiety [ ] Trouble sleeping [ ] Trouble with urination [ ] Joint problems [ ] Itching or rash [ ] Skin problems.

Multi-site clinical NLP implementation

- Sending algorithms to the text
- Bringing text to the algorithms
- Salient issues in multi-site NLP

Multi-site NLP implementation strategies

Send algorithms to the text or bring text to the algorithms

Sending algorithms to the text

- Advantages
  - Maintains local control over text (no data sharing)
  - More likely to discover bugs/errors
- Disadvantages
  - Additional software engineering
  - Variation in sites' technical proficiency
  - Site differences unavoidable

Bringing text to the algorithms

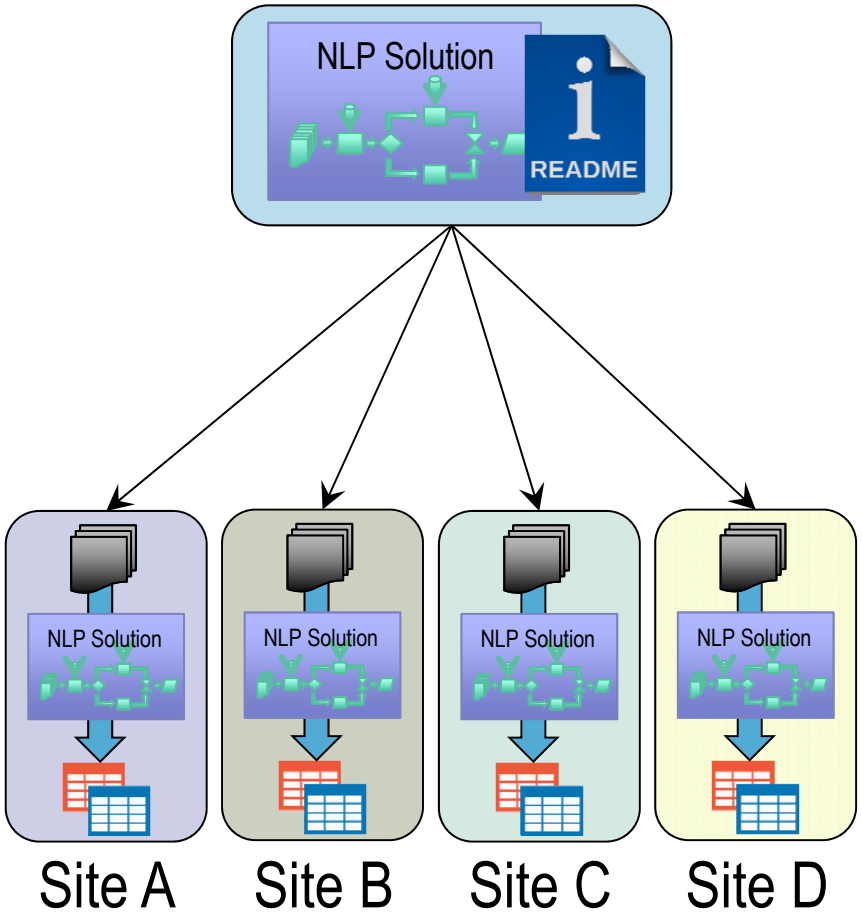
- Advantages
  - Uniform processing
  - Simplifies NLP system tailoring
- Disadvantages
  - De-identification need (local)
  - Loss of local control over text (mitigated by DUA)

- *Local assembly of corpora always necessary*

Multi-site NLP implementation strategies

Sending algorithms to the text

- Advantages
  - Maintains local control over text (no data sharing)
  - More likely to discover bugs/errors
- Disadvantages
  - Additional software engineering
  - Variation in sites' technical proficiency
  - Site differences unavoidable



KAISER PERMANENTE®

Multi-site NLP: Sending algorithms to text

Portable NLP system used in the eMERGE consortium

- Task: Process free-text imaging reports from 7 eMERGE consortium sties to support GWAS study of CAAD
- Rule-based system captured numeric ("… `50-69% stenosis` …") and qualitative ("… `completely occluded` …") descriptions
- Designed for portability and e
  - Self-installing JAVA applicatior
  - Simple GUI, data output
  - Trouble-shooting features
- Columbia, Geisinger, Harvard
- Three revisions needed
  - 1) Original, 2) Check exam type, 3) Ignore normative range boilerplate

```
==================================
CAROTID STENOSIS REFERENCE:
-distal internal carotid artery
diameter as the denominator for
stenosis measurement:
 MILD = <50% stenosis.
MODERATE =  50-69% stenosis.
SEVERE =  70-89% stenosis.
HAIRLINE/CRITICAL = 90-99% stenosis.
OCCLUDED = 100% stenosis.
==================================
```


KAISER PERMANENTE®

Multi-site NLP: Sending algorithms to text

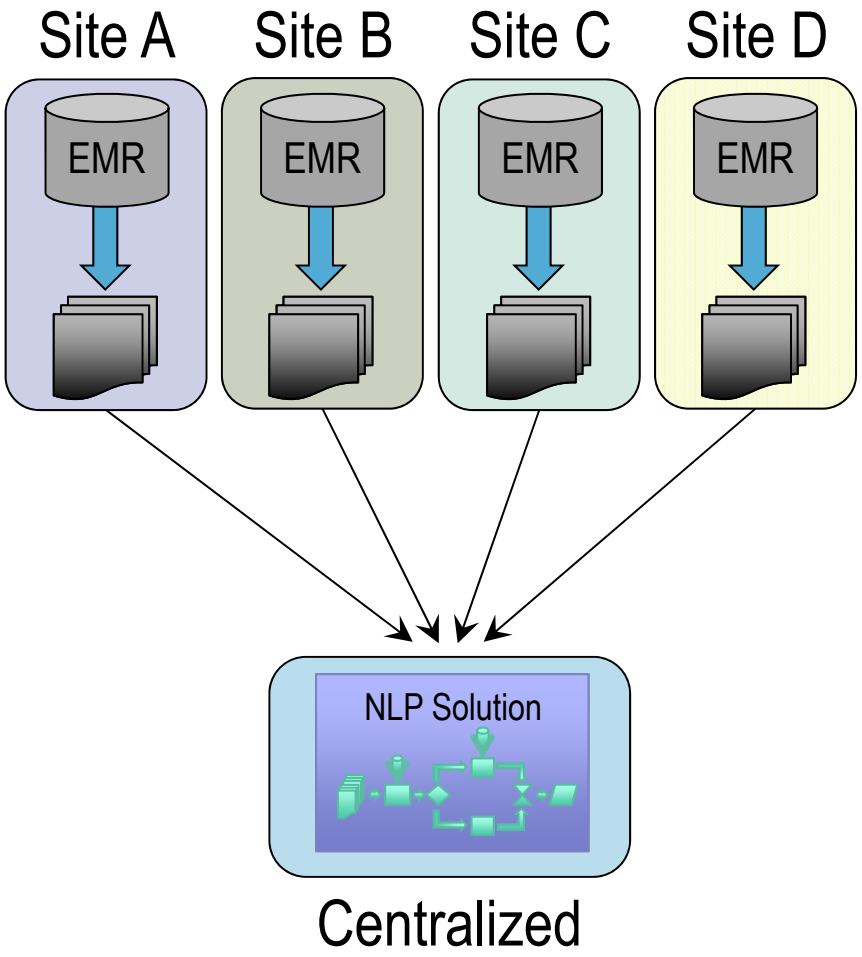
Portable NLP system used in the eMERGE consortium

Issues:

- Design NLP system for portability
- Accommodate site differences in sources, language
- Tailoring needed at each site
- Tailoring across firewalls is tricky (*tele-informatics*)
- Validation needed at each site
- Simpler NLP tasks are the best candidates

KAISER PERMANENTE

Multi-site NLP: Bringing text to algorithms



Bringing text to the algorithms

- Advantages
  - Uniform processing
  - Simplifies NLP system tailoring
- Disadvantages
  - De-identification need (local)
  - Loss of local control over text (mitigated by DUA)

KAISER PERMANENTE®

Multi-site NLP: Bringing text to the algorithms

Centralized NLP in a 4-site colonoscopy study (Ateev Mehrotra, PI)

- Task: Calculate colonoscopy quality metrics in diverse settings
- Centralized adaptation of an existing NLP system
  - Centralized, manually-annotated training and validation corpora
  - All software engineering at lead site
- Each site assembles, de-identifies corpora
  - 2 years of colonoscopy & associated pathology reports
  - De-identification via open-source Miter MIST or commercial De-ID®
  - DUAs & IRB reviews
- KPW, UNC, UPMC, Central Illinois Endoscopy (CIE)
- *2+ years* to adapt existing NLP system!

DC1
DC2

Multi-site NLP: Bringing text to the algorithms

Centralized NLP in a 4-site colonoscopy study (Ateev Mehrotra, PI)

Characteristics of the sites to which the original NLP system was adapted								
Site	Characteristic							
	Location	Practice type	EHR type	Compen-sation	No. of MDs	No. of CSPYs	Report characteristics	
							Median word count (CSPY/Path)	Unqiue UMLS concepts (CSPY/Path)
KPW	West	HMO	Compre-hensive	Salary	18	12,098	980/716	8,920/988
CIE	Mid-W	Private	GI specialty	Fee-for-service	11	13,036	504/214	1,132/873
UNC	South	Univ.	Comp. & GI	Salary + incentive	53	19,062	733/247	1,643/1,684
UPMC	NE	Univ. & private	Comp. & GI	FFS, salary + incentive	119	73,990	388/595	4,093/3,977

KAISER PERMANENTE®

Slide 34

DC1 David Carrell, 6/9/2017

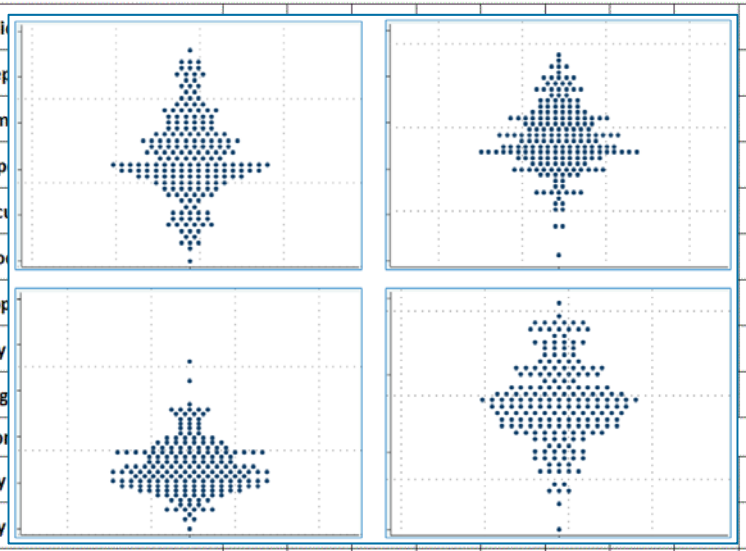
DC2 David Carrell, 6/9/2017

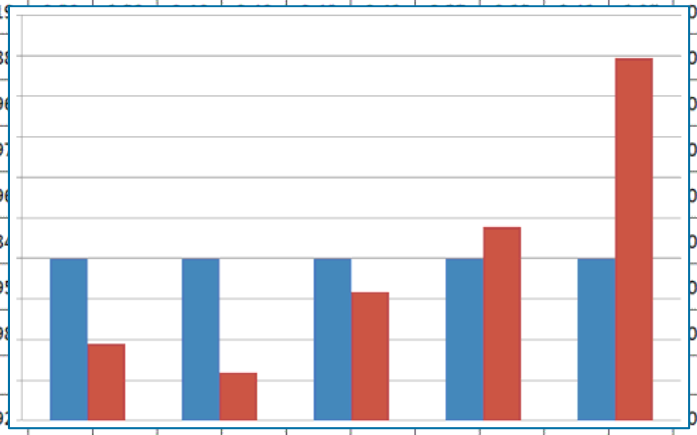
Multi-site NLP: Bringing text to the algorithms

Centralized NLP in a 4-site colonoscopy study (Ateev Mehrotra, PI)

Table 2. Performance of the NLP system for 13 colonoscopy quality metrics as measured by accuracy, average recall, precision, and F measure by study site.(1)

Quality metric	Accuracy					Recall					Precision					F measure				
	All Sites	UPMC	UNC	CIE	KPW	All Sites	UPMC	UNC	CIE	KPW	All Sites	UPMC	UNC	CIE	KPW	All Sites	UPMC	UNC	CIE	KPW
From colonoscopy report																				
Indi									0.93	0.93	0.90	0.93	0.92	0.88	0.88	0.90	0.90	0.91	0.91	0.90
Prep									0.1											0.46
Fam									0.8											0.89
App									0.9											0.99
Cecu									0.9											0.99
Ileo									0.9											0.99
Biop									0.8											0.85
Any									0.9											0.91
Larg									0.9											0.98
Fro																				
Any									0.9											0.95
Any									0.94	0.91	0.96	0.91	1.00	0.98	0.91	0.95	0.91	0.97	0.96	0.91
Any villous/dysplastic chgs.	0.99	0.92	0.94	0.79	0.92	0.90	0.80	1.00	0.95	0.80	0.87	0.92	0.82	0.86	0.92	0.88	0.86	0.90	0.90	0.86
Location in colon	0.87	0.99	1.00	1.00	0.99	0.85	0.92	0.92	0.78	0.92	0.85	0.91	0.90	0.80	0.91	0.85	0.91	0.91	0.79	0.91





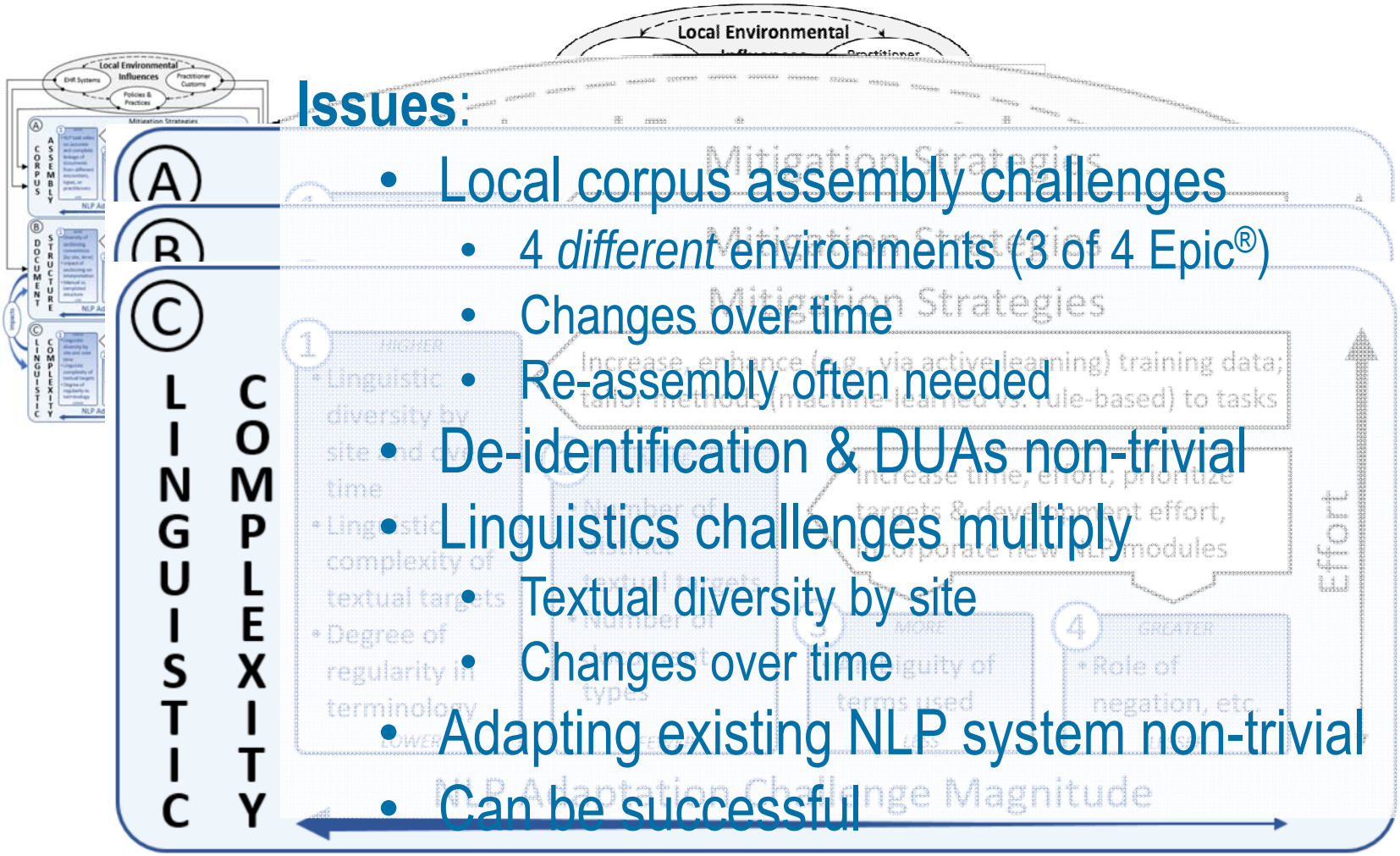
(1) Study sites: UPMC = University of Pittsburgh Medical Center, UNC = University of North Carolina, CIE = Central Illinois Endoscopy , KPW = Kaiser Permanente Washington.

(2) These variables had >2 possible outcomes and were therefore evaluated using average recall, average precision, and average F measure, defined as the arithmetic mean of the respective performance metric calculated for each outcome category individually.

KAISER PERMANENTE®

Multi-site NLP: Challenges and strategies

Many challenges in clinical NLP have little to do with NLP – Josh Denny



Issues:

- Local corpus assembly challenges
  - 4 *different* environments (3 of 4 Epic®)
  - Changes over time
  - Re-assembly often needed
- De-identification & DUAs non-trivial
- Linguistics challenges multiply
  - Textual diversity by site
  - Changes over time
- Adapting existing NLP system non-trivial
- Can be successful

Summary

Many challenges in clinical NLP have little to do with NLP– Josh Denny

- "Secondary use" implies many challenges, often unanticipated
- Assembling the right clinical corpus takes effort, local expertise
- Clinical text availability/quality often less than ideal
- Challenges of interpretation, especially in multi-site context
- Simpler NLP tasks → higher likelihood of success

Questions / Discussion

David S. Carrell, PhD
Kaiser Permanente Washington Health Research Institute

June 15, 2017, FDA White Oak Campus, Silver Spring, MD

Abstract

Adapting clinical NLP methods for multi-site medical products research

David S. Carrell, PhD

Medical product clinical trials and postmarketing safety surveillance are increasingly coordinated across multiple institutional settings where secondary use of electronic health record (EHR) data makes large-scale ascertainment of outcomes more efficient. Many important outcomes are captured only in unstructured clinical narrative. Harmonizing information extracted from unstructured text in these settings entails challenges similar to those encountered when combining structured EHR data from geographically and institutionally diverse delivery systems. The adage emerging from these efforts, that "all data are local," is at least as relevant to unstructured clinical data as it is to more widely used structured EHR data. This presentation will describe salient issues confronted when adapting and applying natural language processing (NLP) methods across multiple institutional settings. Seemingly simple tasks, such as assembling complete and representative clinical corpora, can be surprisingly challenging. Idiosyncratic characteristics of clinical documentation, including language usage, document structure, and content, makes the application of NLP methods in multi-site settings an endeavor that requires forethought and attention to detail. These and related issues will be illustrated with examples from recent NLP projects in several clinical domains, including a project now underway to extract from clinical progress notes information about patient-reported medication side effects.

Other issues …

- Character set clashes
- Good PDFs / Bad PDFs (scanned images of text)
- Dictated/transcribed notes (misunderstandings)
- Open source vs. proprietary software
- Machine learned algorithms and HIPPA PHI
- Co-reference resolution
- Negation
- Epic “CareEverywhere” (= “DataNowhere”)