BioMedical NLP

Class 3/4 - Clinical data, EHRs
NLP Master's Programme, University of Bucharest

Lect. Dr. Ana Sabina Uban
auban@fmi.unibuc.ro

- slides credit: MIT, David Sontag & Pete Szolovits

Electronic Health Records

The purpose of a patient record is "to recall observations, to inform others, to instruct students, to gain knowledge, to monitor performance, and to justify interventions."

Stanley Reiser (1991)

Clinical data

Who (e.g., physician, nurse, pharmacist, front desk, professional biller, etc.)

When (e.g., when the lab was drawn, when the results were available, etc.)

Why (e.g., clinical care, billing, auditing, legal record, etc.)

What (e.g., structured, semi-structured, or unstructured data)

Electronic Health Records

Patient Care

Diagnoses

Medication

Allergies

Laboratory Tests

Radiology Images

Provider notes

Other Roles of the EHR

Billing record

- what was performed?
- by whom?
- for what purpose?

Legal record

- who recorded data?
- who saw what when?
- Audit logs

Electronic Health Records

EHR vs EMR

EMR Electronic Medical Record

EHR Electronic Health Record

EHR:

- more comprehensive

- includes other health data like radiology reports, lab reports, billing information, notes from multiple physicians

How Doctors Feel About Electronic Health Records

Poll among Primary Care Physicians (2019):...

Why automatic/ML methods? Problems

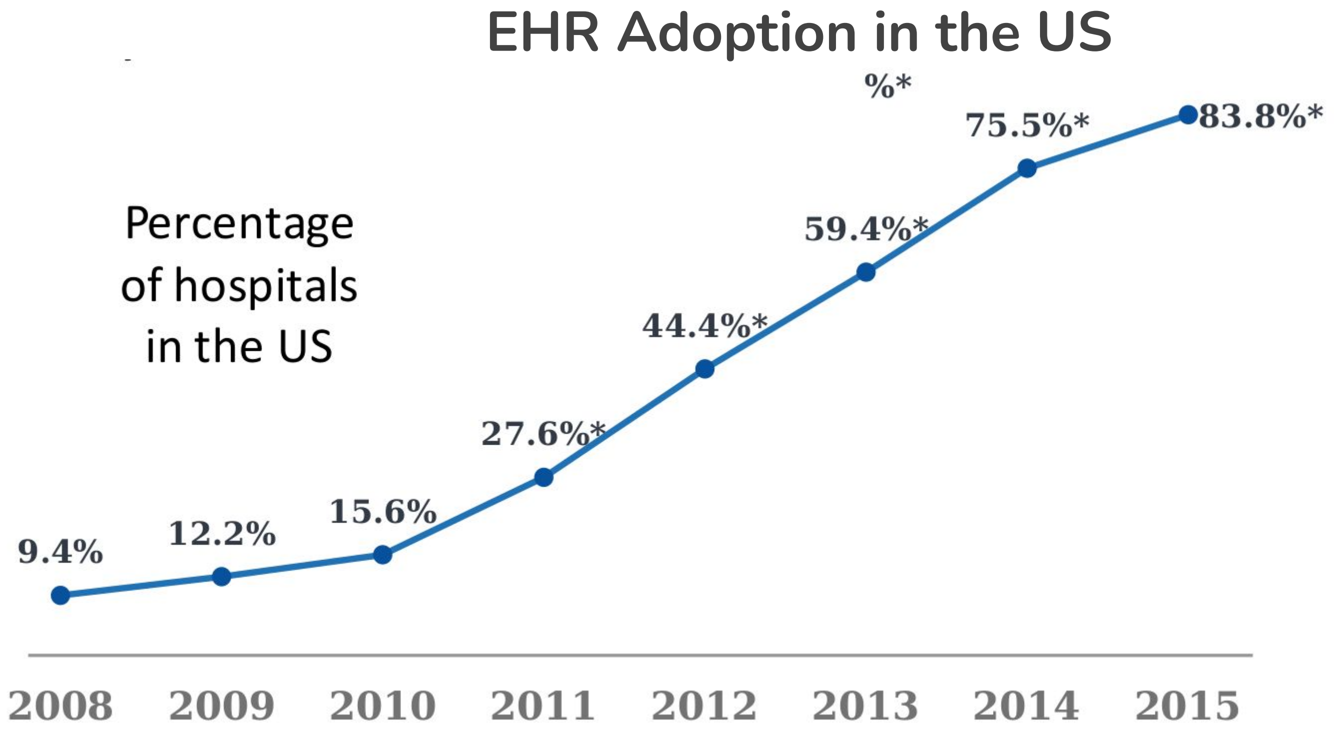
- Cost of health care expenditures in the US are over $3 trillion, and rising
- Despite having some of the best clinicians in the world, chronic conditions are:

– Often diagnosed late

– Often inappropriately managed

- Medical errors are pervasive

Why automatic/ML methods? Why now?



Courtesy of Health and Human Services. Image is in the public domain.

Why automatic/ML methods? Why now?
Large datasets

Why automatic/ML methods? Why now? Standardization of clinical data

Diagnosis codes: ICD-9 and ICD-10 (International Classification of Diseases)

Laboratory tests: LOINC codes

Pharmacy: National Drug Codes (NDCs)

Unified Medical Language System (UMLS): millions of medical concepts

Why automatic/ML methods? Why now? Standardization of clinical data

ICD-10	Site	Comments
C00 - 96	All sites	Includes the following D-diagnoses; D32-D33, D35.2-35.4, D42-D43, D44.3-D44.5 and D45-47
C38	Mediastinum, pleura	Excludes mesotheliomas (which are included in C45)
C44	Skin, non-melanoma	Excludes basal cell carcinoma
C56	Ovary	Excludes borderline tumours
C64	Kidney except renal pelvis	Excludes non-invasive papillary tumours
C65	Renal pelvis	Includes non-invasive papillary tumours
C66	Ureter	Includes non-invasive papillary tumours
C67	Bladder	Includes non-invasive papillary tumours
C68	Other and unspecified urinary organs	Includes non-invasive papillary tumours
C70	Meninges	Includes benign tumours (D32-33, D42-43)
C71	Brain	Includes benign tumours (D32-33, D35.2-35.4, D42-43, D44.3-44.5)
C72	Spinal cord, cranial nerves and other parts of central nervous system	Includes benign tumours (D32-33, D42-43)
C75	Other endocrine glands and related structures	Includes benign tumours (D44.3-44.5)
C92	Myeloid leukaemia	Includes myelodyplastic syndrome (D46)
C95	Leukaemia of unspecified cell type	Includes polycythemia vera (D45) and other unspecified tumours in lymphatic or hematopoietic tissue (D47)

Why automatic/ML methods? Why now? Standardization of clinical data

https://www.healthcareitleaders.com/blog/the-18-most-bizarre-icd-10-codes-infographic/

Why automatic/ML methods? Why now? Industry interest in ML & Healthcare

Major acquisitions to get big data for ML:

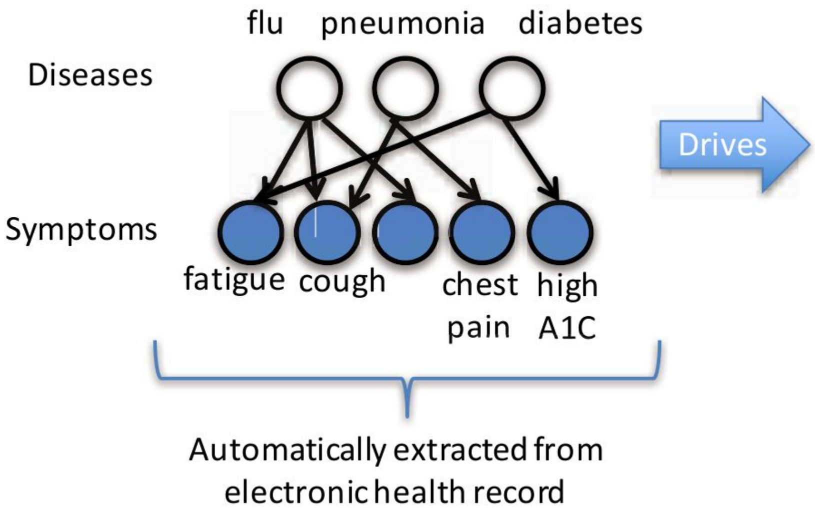
– Merge ($1 billion purchase by IBM, 2015): medical imaging

– Truven Health Analytics ($2.6 billion purchase by IBM, 2016): health insurance claims

– Flatiron Health ($1.9 billion purchase by Roche, 2018): electronic health records (oncology)

What will the ER of the future be like?

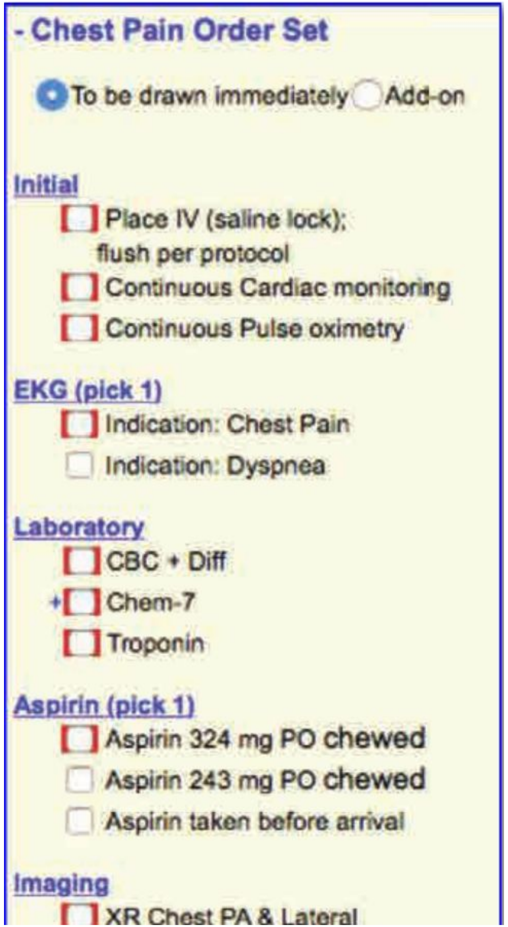
Behind-the-scenes reasoning about the patient’s conditions (current and future)



- Better triage
- Faster diagnosis
- Early detection of adverse events
- Prevent medical errors

What will the ER of the future be like?

Anticipating the clinician's needs


- Chest Pain Order Set

To be drawn immediately Add-on

Initial

Place IV (saline lock); flush per protocol

Continuous Cardiac monitoring

Continuous Pulse oximetry

EKG (pick 1)

Indication: Chest Pain

Indication: Dyspnea

Laboratory

CBC + Diff

Chem-7

Troponin

Aspirin (pick 1)

Aspirin 324 mg PO chewed

Aspirin 243 mg PO chewed

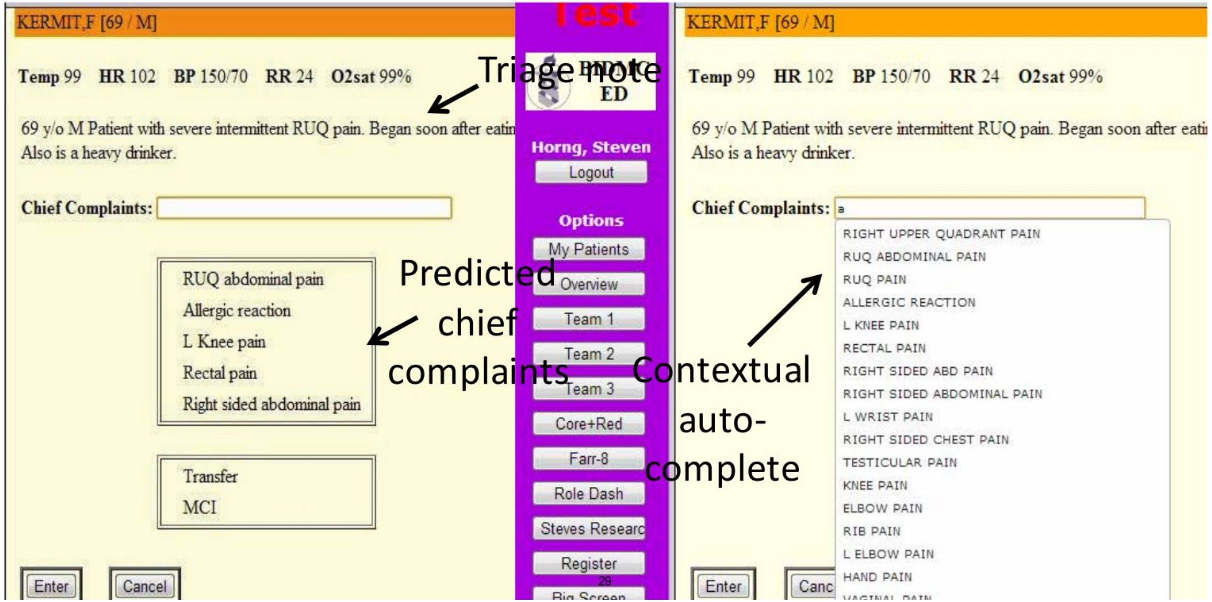
Aspirin taken before arrival

Imaging

XR Chest PA & Lateral

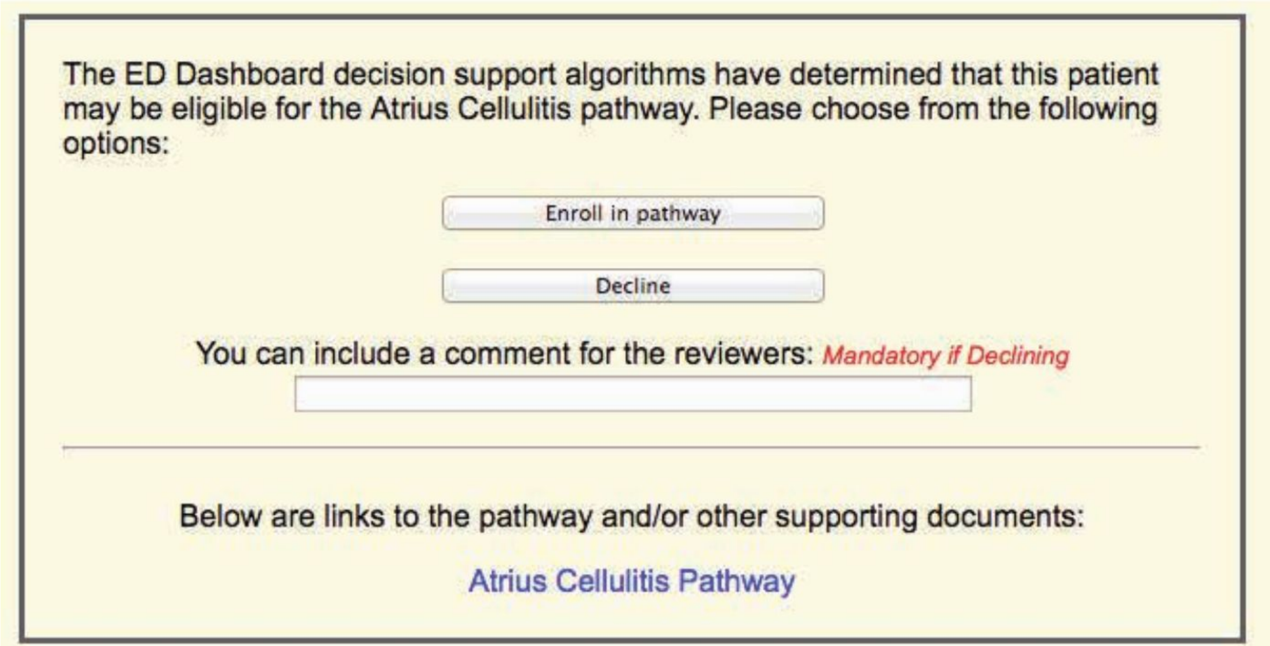
What will the ER of the future be like?

Automated documentation and billing



What will the ER of the future be like?

Propagating best practices


The ED Dashboard decision support algorithms have determined that this patient may be eligible for the Atrius Cellulitis pathway. Please choose from the following options:

Enroll in pathway

Decline

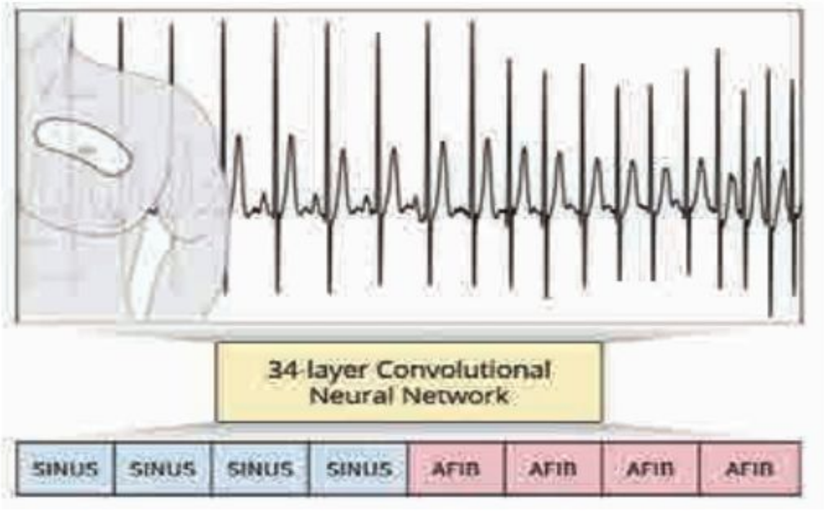
You can include a comment for the reviewers: *Mandatory if Declining*

Below are links to the pathway and/or other supporting documents:

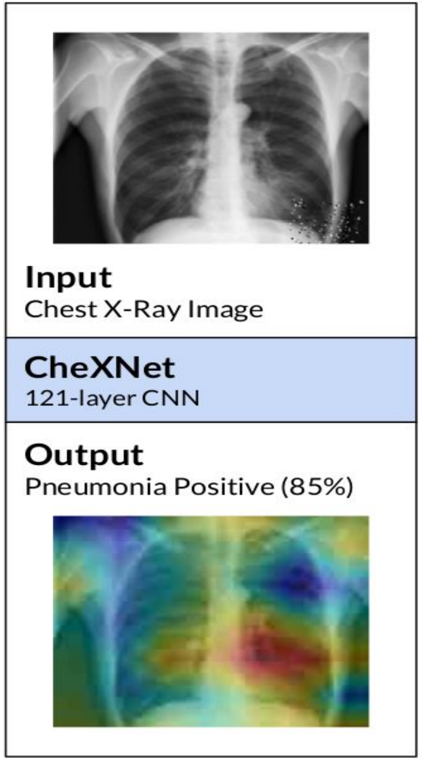
Atrius Cellulitis Pathway

What will the ER of the future be like?

Reducing the need for specialist consults

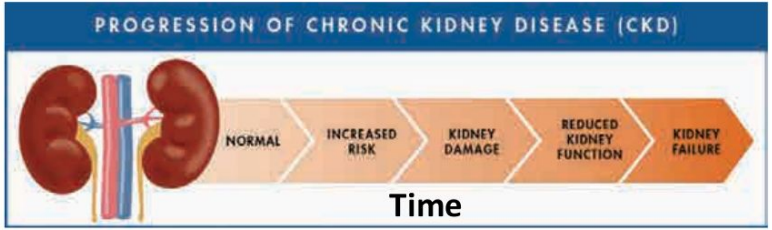


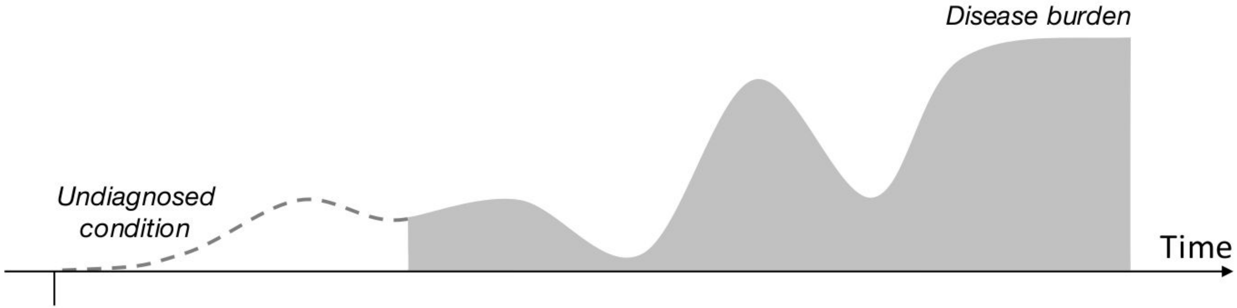
Arrhythmia?



What is the future of how we treat chronic disease?

Predicting a patient's future disease progression





What is the future of how we treat chronic disease?

- Early diagnosis, e.g. of diabetes, Alzheimer's, cancer
- Continuous monitoring and coaching, e.g. for the elderly, diabetes, psychiatric disease
- Discovery of new disease subtypes; design of new drugs; better targeted clinical trials

What makes healthcare different?

- Life or death decisions

– Need robust algorithms

– Check and balances built into ML deployment

– (Also arises in other applications of AI such as autonomous driving)

– Need fair and accountable algorithms

- Many questions are about unsupervised learning

– Discovering disease subtypes, or answering questions such as “characterize the types of people that are highly likely to be readmitted to the hospital”?

- Many of the questions we want to answer are causal

– Naïve use of supervised machine learning is insufficient

What makes healthcare different?

- Very little labeled data
  – Motivates semi-supervised learning algorithms
- Sometimes small numbers of samples (e.g. a rare disease)
  – Learn as much as possible from other data (e.g. healthy patients)
  – Model the problem carefully
- Lots of missing data, varying time intervals, censored labels

What makes healthcare different?

Difficulty of de-identifying data

– Need for data sharing agreements and sensitivity

• Difficulty of deploying ML

– Commercial electronic health record software is difficult to modify

– Data is often in silos; everyone recognizes need for interoperability, but slow progress

– Careful testing and iteration is needed

Types of Data

- Demographics

  Age, sex, socio-economic status, insurance type, language, religion, living

  situation, family structure, location, work, ...

- Vital signs

  Weight, height, pulse, respiration rate, body temperature, ...

Types of Data

- Medications

  Prescriptions, over-the-counter drugs, illegal drugs, alcohol, …

  Medication reconciliation

- Laboratory

  Components of blood, urine, stool, saliva, spinal fluid (CSF), ascitic fluid, joint fluid, bone marrow, lung, …

Types of Data

- Pathology

  Qualitative and quantitative examination of any body tissue, e.g., biopsy samples, surgical “scraps”

- Cell-level measurements, e.g., cell-surface antigens

Types of Data

Microbiology — organisms grown, typically from cultures

- Testing sensitivity to various antibiotics, at various dilutions

    Input/Output (fluids)

- Notes
- Discharge summary
- Attending and/or Resident
- Nurse

Types of Data

- Nurse
- Specialist
- Radiology, Pathology, ECG, Nutrition, Respiratory, Social work, …
- Consultant
- Referring physician
- Emergency Department

Types of Data

Billing

- Diagnoses (ICD-{9, 10})
- Procedures (CPT and ICD)
- Diagnosis Related Groups (DRG) [~ abstraction of ICD]

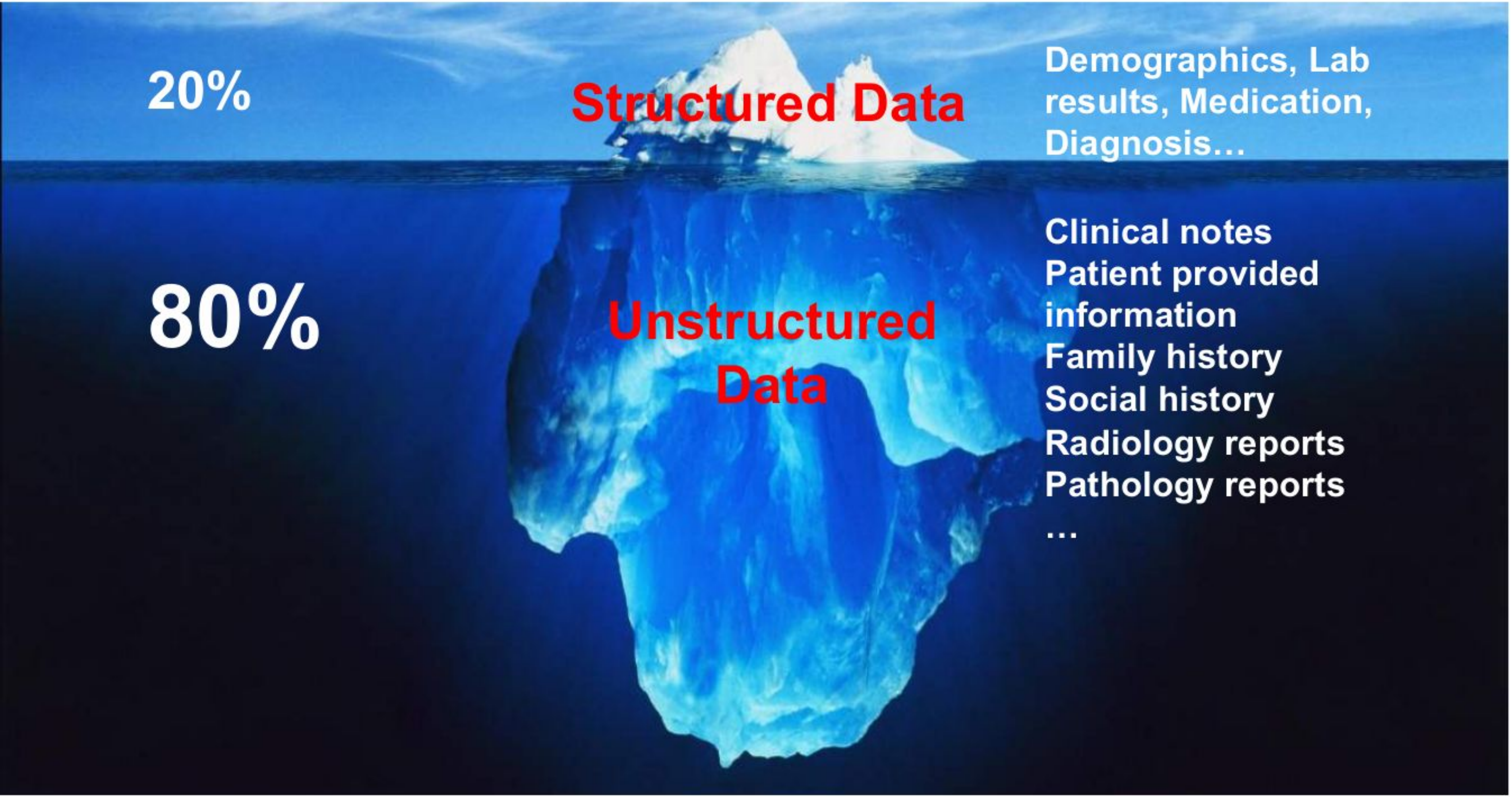
Administrative

- Service
- Transfers

Types of data

Imaging

- X-ray
- Ultrasound
- CT
- MRI
- PET
- Retinal
- Endoscopy
- Photographs


20%
Structured Data
Demographics, Lab results, Medication, Diagnosis…
80%
Unstructured Data
Clinical notes
Patient provided information
Family history
Social history
Radiology reports
Pathology reports
…

Relational databases

NLP Tools Required

Diagnosis codes

Fake ID	ENTRY_DAT	CODE
34068	5/13/2001	41.85
37660	8/6/2002	79.99
140680	8/31/2003	79.99
23315	5/14/2003	112
75936	7/9/2004	117.9

Lab tests

Fake ID	TEST	ENTRY_DAT	VALU
3536	pO2	1/23/1996	314
72921	LDL	2/5/1996	34
102460	pCO2	1/26/1996	45
135043	HDL	1/25/1996	35
135432	MonAb	1/24/1999	0.16

Structured

Problem lists:

---- Medications known to be prescribed:
Keppra 750 mg 1/2 tab q am and pm
Dexilant 60 mg by mouth daily
aspirin 325 mg 1 tablet by mouth daily
clopidogrel 75 mg tablet 1 tablet by mouth daily

---- Known adverse and allergic drug reactions:
Sulfa Drugs

---- known significant medical diagnoses:
Seizure disorder
Aneurysm
Heartburn

---- Known significant operative and invasive procedures:
2003 Appendectomy
2005 Stents put in **DATE [Aug 29 05]

Semi-structured

Clinical notes

EXAM: BILATERAL DIGITAL SCREENING MAMMOGRAM WITH CAD, **DATE[Mar 16 01]; COMPARISON: **DATE[Jul 01 01] TECHNIQUE: Standard CC and MLO views of both breasts were obtained. FINDINGS: The breast parenchyma is heterogeneously dense. The pattern is extremely complex with postsurgical change seen in the right upper outer quadrant and scattered benign-appearing calcification seen bilaterally. A possible asymmetry is seen in the superior aspect of the left breast. The parenchymal pattern otherwise remains stable bilaterally, with no new distortion or suspicious calcifications. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy.. LEFT: Possible developing asymmetry superior aspect left breast for which further evaluation by true lateral and spot compression views recommended. Ultrasound may also be needed.. RECOMMENDATION: Left diagnostic mammogram with additional imaging as outlined above.. A left breast ultrasound may also be needed. BI-RADS Category 0: Incomplete Assessment - Need additional imaging evaluation. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy....

Unstructured

MIMIC-III Dataset

MIMIC-III integrates **deidentified**, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, accessible to researchers internationally under a **data use agreement**.

Allows clinical studies to be reproduced and improved.

MIMIC-III Dataset

MIMIC-III database: populated with data acquired during routine hospital care, no associated burden on caregivers and no interference with their workflow.

Data was downloaded from several sources, including:

- archives from critical care information systems.
- hospital electronic health record databases.
- Social Security Administration Death Master File.

(Source: https://physionet.org/content/mimiciii/1.4/)

MIMIC-III Dataset - Structure

SUBJECT_ID - patient

HADM_ID - hospital admission

Hospital Stay Data

ADMISSIONS - unique hospitalizations

ICDSTAYS - unique ICU stays

PATIENT - unique patients in data

CALLOUT - movement from ICU

TRANSFERS - movement in hospital

Critical Care Unit Data

CAREGIVERS - unique caregivers

CHARTEVENTS - all observations

DATETIMEEVENTS - all date observations

NOTEEVENTS - clinical notes

INPUTEVENTS_CV - fluid intake in CV

INPUTEVENTS_MV - fluid intake in MV

PROCEDUREEVENTS_MV - procedures in MV

OUTPUTEVENTS - fluids excreted

Hospital System Data

CPTEVENTS - billing

DIAGNOSES_ICD - billing

PROCEDURES_ICD - billing

DRGCODES - billing

LABEVENTS - laboratory data

MICROBIOLOGYEVENTS - laboratory data

PRESCRIPTIONS - medications

MIMIC-III Dataset

Structure

https://mimic.mit.edu/docs/iii/tables/

https://mit-lcp.github.io/mimic-schema-spy/

https://mit-lcp.github.io/mimic-schema-spy/relationships.html

MIMIC-III Dataset

How to access MIMIC data:
https://mimic.mit.edu/docs/gettingstarted/

- complete a recognized course in protecting human research participants that includes Health Insurance Portability and Accountability Act (HIPAA) requirements.
- sign a data use agreement, which outlines appropriate data usage and security standards, and forbids efforts to identify individual patients.

Approval requires at least a week.

Demo link: https://physionet.org/content/mimiciii-demo/1.4/

MIMIC - IV : free text notes

https://mimic.mit.edu/docs/iv/modules/note/

The Note module contains deidentified free-text clinical notes for hospitalized patients.

MIMIC-Note is currently not publicly available and the structure is subject to change.

MIMIC example

In MIMIC dataset, we see asterisks in places of names, dates, locations etc. Here those entities have been replaced with synthetics names, dates, locations etc. to make it look like a piece that reads like a real text:

MIMIC example

Mr. Blind is a 79-year-old white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center. The patient developed hematemesis November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm. The patient's hematocrit was stable and he was given no further intervention.

The patient attempted a gastrografin swallow on the 21st, but was unable to cooperate with probable aspiration. The patient also had been receiving generous intravenous hydration during the period for which he was NPO for his esophageal tear and intravenous Lasix for a question of pulmonary congestion.

On the morning of the 22nd the patient developed tachypnea with a chest X-ray showing a question of congestive heart failure. A medical consult was obtained at the Valtawnprinceel Community Memorial Hospital. The patient was given intravenous Lasix.

Note: orange=demographics; blue=patient condition, diseases, etc.; red=procedures, tests; magenta=results of measurements; yellow=time

Example: predicting disease of patient from EHR data

2010 project: tried to understand what are the genetic correlates of **rheumatoid arthritis** (RA). => Research Patient Data Repository (RPDR) of Massachusetts General and Brigham Partners Healthcare, find the patients who had been **billed** for rheumatoid arthritis.

=> positive predictive value of having a billing code for RA: ~ 19%...

- billing codes were not created to specify what was actually wrong with the patient; instead the billing codes were meant to tell insurance companies/medicare that how much of the payment is reserved

=> use 3 billing codes: predictive value to about 27%.

Example: predicting disease of patient from EHR data

Using **narrative text** instead:

- they used a system called HITEx that extracted entities from narrative text health care provider notes, radiology reports, pathology reports, discharge summaries, and operative reports,
- + diagnoses notes, medications, laboratory data and radiology findings.
- + hand-curated list of alternative ways of saying the same thing
- dealt with negation.
- model used: **logistic regression**.

=> PPV **94%** (combination of:)

Using codified data (e.g. lab values, demographics) only to predict whether a patient has rheumatoid arthritis lead to a PPV of **88%**. Using NLP on clinical text (nursing notes, discharge summaries etc.) gave a PPV of **89%.**

Named Entity Recognition

BRIEF HISTORY: The patient is an (XX)-year-old female with history of <problem> previous stroke </problem> ; <problem> hypertension </problem> ; <problem> COPD </problem> , stable ; <problem> renal carcinoma </problem> ; presenting after <problem> a fall </problem> and possible <problem> syncope </problem> .
While walking , she accidentally fell to her knees and did hit <problem> her head on the ground </problem> , near <problem> her left eye </problem> .
<problem> Her fall </problem> was not observed , but the patient does not profess <problem> any loss of consciousness </problem> , recalling the entire event.
The patient does have a history of <problem> previous falls </problem> , one of which resulted in <problem> a hip fracture </problem> .
She has had <treatment> physical therapy </treatment> and recovered completely from that .
<test> Initial examination </test> showed <problem> bruising </problem> around the left eye , normal lung examination , normal heart examination , normal neurologic function with a baseline decreased mobility of <problem> her left arm </problem> .
The patient was admitted for <test> evaluation </test> of <problem> her fall </problem> and to rule out <problem> syncope </problem> and possible <problem> stroke </problem> with <problem> her positive histories </problem> .
<test> DIAGNOSTIC STUDIES: All x-rays </test> including <problem> left foot , right knee , left shoulder and cervical spine </problem> showed no <problem> acute fractures </problem> .
<problem> The left shoulder did show old healed left humeral head and neck fracture </problem> with <problem> baseline anterior dislocation </problem> .
<test> CT of the brain </test> showed no <problem> acute changes </problem> , <problem> left periorbital soft tissue swelling </problem> .
<test> CT of the maxillofacial area </test> showed no <problem> facial bone fracture </problem> .
<test> Echocardiogram </test> showed normal left ventricular function , <test> ejection fraction </test> estimated greater than 65% .

Relation Extraction

Determine relationships between entities or events

"We used hemofiltration to **treat** a patient with digoxin overdose that was complicated by refractory hyperkalemia." [PMID: 3718110]

Relationship: Hemofiltration-**TREATS**-Patients

Negation Identification (NegIde)

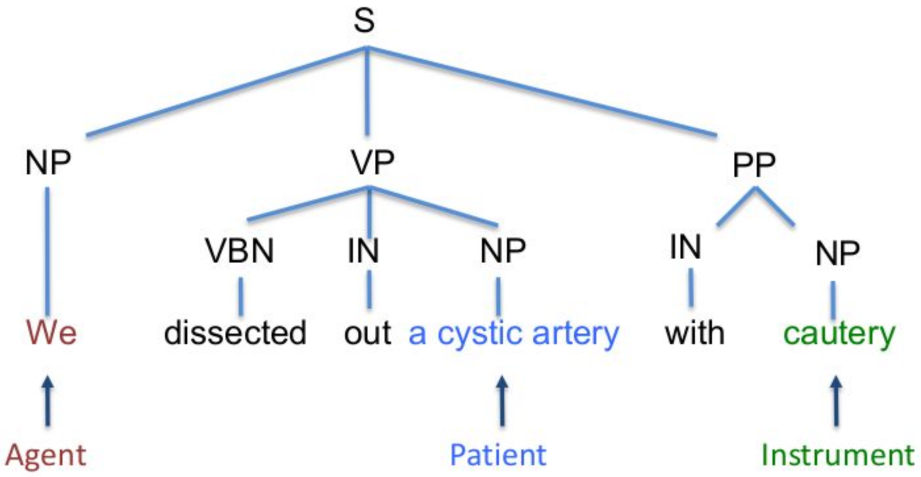
Identify pertinent Negatives from narrative clinical reports

- “The chest X-ray showed no infiltrates…”
- “The patient denied experiencing chest pain”
- “ no murmurs, rubs or gallops”
- “murmurs, rubs and gallops are absent”

Semantic Role Labeling

Detect the semantic role played by each noun phrase associated with the verb of a sentence

- Agent: NP before the verb
- Patient: NP after the verb
- Instrument: NP in a Prepositional Phrase (PP)



Information Extraction

Automated extraction of family and observation predications from unstructured text

- Supplied text: "Heart disease on the father side of the family. Mother has arthritis.”
- Extracted elements:

  Constituent: family {FAMILY HISTORY: FAMMEMB}

  Constituent: observation {Heart disease: C1576434}

  Constituent: family {father side of the family: Paternal*}

  Constituent: family {Mother: MTH}

  Constituent: observation {arthritis: C1692886}
- Predications :

  Family Member{father side of the family}, Observation{Heart disease},Negated{false}

  Family Member{Mother}, Observation{arthritis}, Negated{false}

NLP Tools

Spark NLP : https://nlp.johnsnowlabs.com/demos (requires license for some functions…)

scispacy https://allenai.github.io/scispacy/

Negation (**NegEx**): https://code.google.com/archive/p/negex/ free!

Semantic relations (**SemRep**): Broad-coverage biomedical relation extraction with SemRep | BMC Bioinformatics | Full Text
https://github.com/lhncbc/SemRep (requires UMLS license/agreement…)

Bio-medical Language Models: e.g. https://huggingface.co/bionlp

Information extraction tasks & methods in NLP…

More: see slides on general Intro to NLP and types of tasks & algorithms … (*A Very Quick Introduction to NLP 2*)

Review of ML for Clinical Text Data Applications [1]

EHR-based phenotyping:identify cohorts of individuals that share certain clinical characteristics, events, and service patterns:
=> observational and interventional studies
=> prospective recruitment into clinical trials
=> health services research, public health surveillance
=> comparative effectiveness research.

Review of ML for Clinical Text Data Applications [1]

For instance, historical trial patient enrollment decisions were used to demonstrate the potential of NLP to increase **trial screening efficiency by 450%** and reduce workload associated with patient cohort identification by 90% (e.g. events identified from EHRs: falls, long bone fractures)

For **prognosis**, text classification results were used to predict 3-month survival, likelihood of intracranial hemorrhage, development of coronary artery disease, prognosis based on cancer staging

Review of ML for Clinical Text Data Applications [1]

Lower-level tasks such as **coreference resolution** and **WSD** were not associated with any particular clinical application, but with enabling other higher-level NLP tasks.
NER can be used to support structuring text into predefined templates. (majority of NER studies were related to NLP community challenges, focused on entities such as medical problems, tests, and treatments; disorders; and protected health information)

Review of ML for Clinical Text Data Applications [1]

IE in general, wider clinical applications: **prognosis** and **care improvement:**

- cancer stage detected from clinical narratives, used for prognosis
- extraction of symptoms experienced by patients during chemotherapy => improve patient care through modifying treatments and recognizing and managing symptoms
- extraction of information about assessments and medications used to improve management and outpatient treatment of patients suffering from chronic heart failure

Review of ML for Clinical Text Data Applications [1]

Triage (sorting patients into groups based on their need for or likely benefit from medical treatment):

- clustering was used to identify latent groups of lymphoma patients from their pathology reports.
- Automatically generated clusters of radiology reports coincided with major topics in radiology investigations

However, triage was not found to be a common clinical application of NLP.

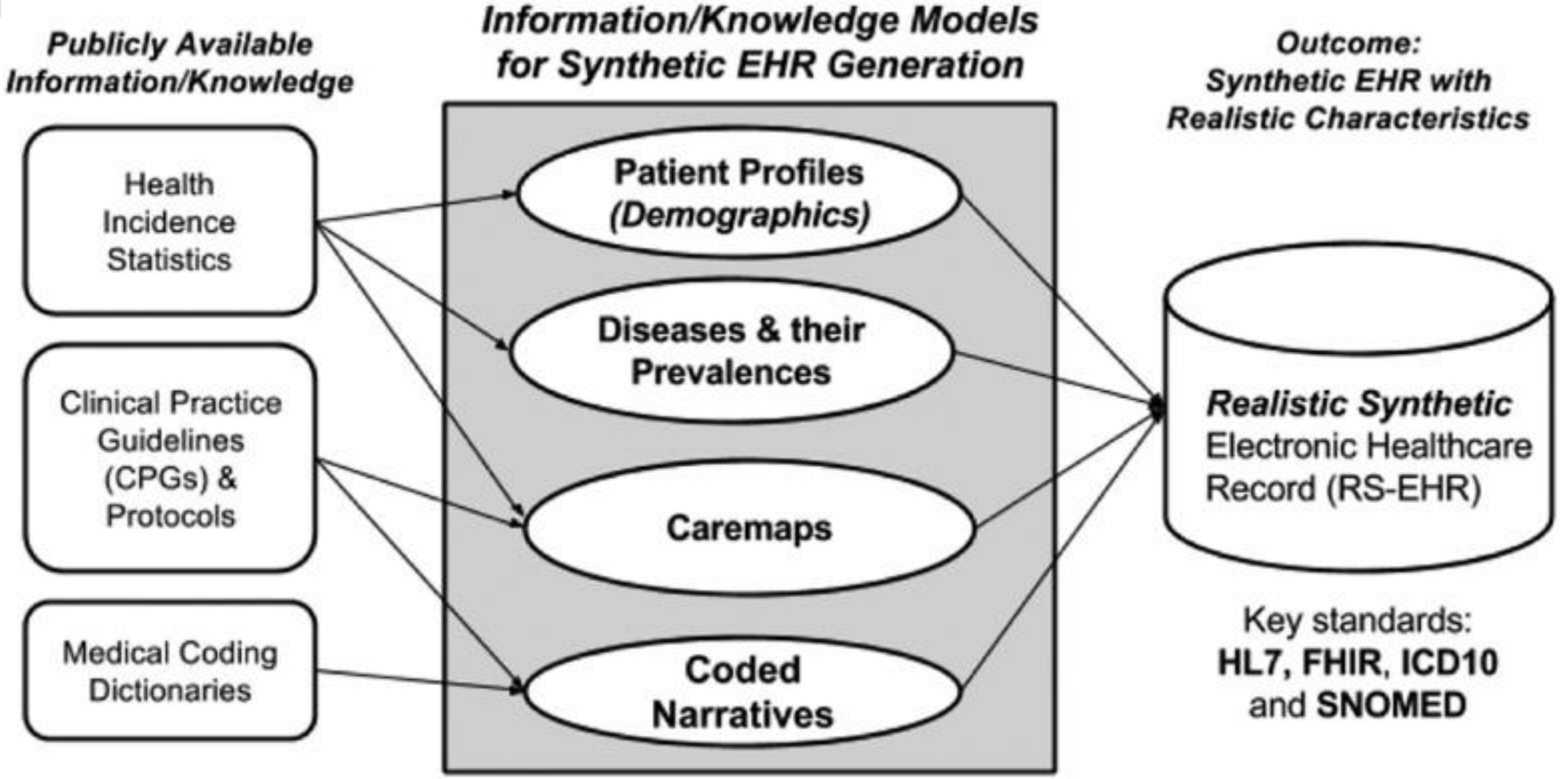
Review of ML for Clinical Text Data Applications [1]

Challenges:

- preservation of patient privacy and the annotation bottleneck
- training datasets have problems associated unrepresentative samples (may not reflect the distribution of characteristics of the target problem)
- most datasets used in the included studies originated from few institutions; format and style of clinical notes varying substantially across institutions => a significant drop in performance was observed when training a model in one institution and testing it in another

Alternative: synthetic data

Synthetic data: Synthea [2]



Review of ML for Clinical Text Data Applications [1]

[1] Spasic, Irena, and Goran Nenadic. "Clinical Text Data in Machine Learning: Systematic Review." *JMIR medical informatics* vol. 8,3 e17984. 31 Mar. 2020, doi:10.2196/17984
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7157505/

Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record [2]

[2] Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association : JAMIA*, *25*(3), 230–238. https://doi.org/10.1093/jamia/ocx079

https://synthetichealth.github.io/synthea/#technology-landing

Medical NER

BRIEF HISTORY: The patient is an (XX)-year-old female with history of <problem> previous stroke </problem> ; <problem> hypertension </problem> ; <problem> COPD </problem> , stable ; <problem> renal carcinoma </problem> ; presenting after <problem> a fall </problem> and possible <problem> syncope </problem> .
While walking , she accidentally fell to her knees and did hit <problem> her head on the ground </problem> , near <problem> her left eye </problem> .
<problem> Her fall </problem> was not observed , but the patient does not profess <problem> any loss of consciousness </problem> , recalling the entire event.
The patient does have a history of <problem> previous falls </problem> , one of which resulted in <problem> a hip fracture </problem> .
She has had <treatment> physical therapy </treatment> and recovered completely from that .
<test> Initial examination </test> showed <problem> bruising </problem> around the left eye , normal lung examination , normal heart examination , normal neurologic function with a baseline decreased mobility of <problem> her left arm </problem> .
The patient was admitted for <test> evaluation </test> of <problem> her fall </problem> and to rule out <problem> syncope </problem> and possible <problem> stroke </problem> with <problem> her positive histories </problem> .
<test> DIAGNOSTIC STUDIES: All x-rays </test> including <problem> left foot , right knee , left shoulder and cervical spine </problem> showed no <problem> acute fractures </problem> .
<problem> The left shoulder did show old healed left humeral head and neck fracture </problem> with <problem> baseline anterior dislocation </problem> .
<test> CT of the brain </test> showed no <problem> acute changes </problem> , <problem> left periorbital soft tissue swelling </problem> .
<test> CT of the maxillofacial area </test> showed no <problem> facial bone fracture </problem> .
<test> Echocardiogram </test> showed normal left ventricular function , <test> ejection fraction </test> estimated greater than 65% .

Medical NER

NER was the earliest topic tackled in the modern genomic era of BioNLP

Medical NER vs general NER: **diversity** of semantic types of named entities

General: PERSON, ORGANIZATION, LOCATION

Medical: many....

Table 3.1 A sample of the semantic classes of named entities that must be recognized in biomedical NLP. Note the surface similarities between many of the examples. Adapted from Jurafsky and Martin (2008).

Semantic class	**Examples**	**Systems**
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*	Settles (2005); Bada & Hunter (2007)
Cell types	*primary T lymphocytes, natural killer cells, NK cells*	Settles (2005); Johnson *et al.* (2006); Bada & Hunter (2007)
Chemicals	*citric acid, 1,2-diiodopentane, C*	Johnson *et al.* (2006); Corbett, Batchelor, & Teufel (2007)
Drugs	*cyclosporin A, CDDP*	Rindflesch *et al.* (2000)
Genes/proteins	*white, HSP60, protein kinase C, L23A*	Yeh *et al.* (2005)
Malignancies	*carcinoma, breast neoplasms*	Jin *et al.* (2006)
Disorders	*amyotrophic lateral sclerosis*	Aronson (2001a)
Mouse strains	*LAFT, AKR*	Caporaso *et al.* (2005)
Mutations	*C10T, Ala64 → Gly*	Caporaso *et al.* (2007)
Populations	*judo group*	Demner-Fushman & Lin (2007)

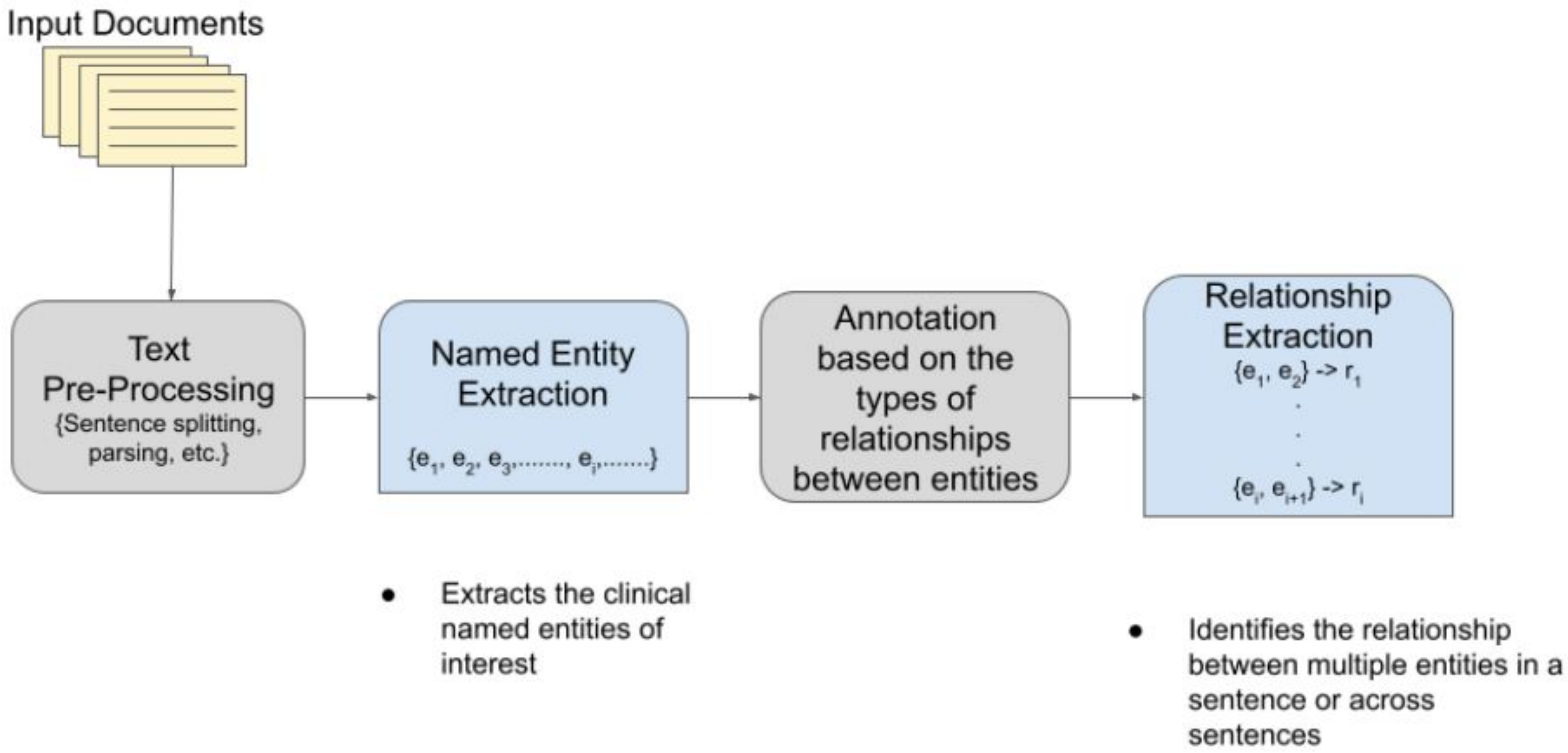


Figure 1. Association between Named Entity Recognition and Relationship Extraction.

Medical NER - example approaches & evaluation

Methods:

- dictionary based
- rule-based
- ML-based

Medical NER

Historically, many medical NER systems are dictionary-based due to richness of domain knowledge resources e.g. UMLS Metathesaurus.

Still, do not fully cover all known named entities + terminology for new diseases, conditions, drugs" e.g. 2003 SARS (Severe Acute Respiratory Syndrome) outbreak, the only sense of SARS in the UMLS Metathesaurus was SARS gene (seryl-tRNA synthetase).

=> corpus-based methods

Medical NER - example approaches

Weakly-supervised approach to recognition of diseases in MEDLINE abstracts based on *iterative pattern learning* [3]:

- recognition of disease names is bootstrapped using several seed patterns pertaining to diseases and commonly found in publications.
- new patterns are learned from the extracted diseases, and new disease names are discovered iteratively.
- problems: could potentially learn spurious patterns and erroneously label instances, e.g. in skills needed to manage patients with complex problems, “complex problems” could be labeled as a specific disease name
- Performance: .60 - .80 F1 score (depending on ranking cutoff).

[3] Xu, R., Supekar, K., Morgan, A., Das, A., & Garber, A. (**2008**). Unsupervised method for automatic construction of a disease dictionary from a large free text collection. In AMIA annual symposium proceedings (Vol. 2008, p. 820). American Medical Informatics Association.

Medical NER - example approaches & evaluation

2007: community-wide evaluation that involved assigning ICD-9 billing codes pediatric radiology reports (Pestian et al. 2007)

i2b2 shared task (2007) involved recognition of all entities needed for de-identification of clinical data in accordance with HIPAA rules.

Other challenges: extraction of medications and other attributes of a drug prescription, and extraction of disorders and diagnostic and therapeutic procedures

i2b2 NLP challenges: largely focused on NER in clinical text, exploring nuances of NER annotation boundaries, annotator consensus and importance of entity types and attributes.

Medical NER - example approaches & evaluation

Recent results

2010 i2b2 shared task: SVM-based supervised learning algorithm performed the best with an F1-score of **0.737**

2011 MADE1.0: competition for detecting Adverse Drug Events (ADEs) from EHR. NER task - best micro-averaged F1-score of **0.892**.

Review [4]:

[4] Bose, P., Srinivasan, S., Sleeman, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. Applied Sciences, 11(18), 8319.

NER approaches 2015-2021 [4]

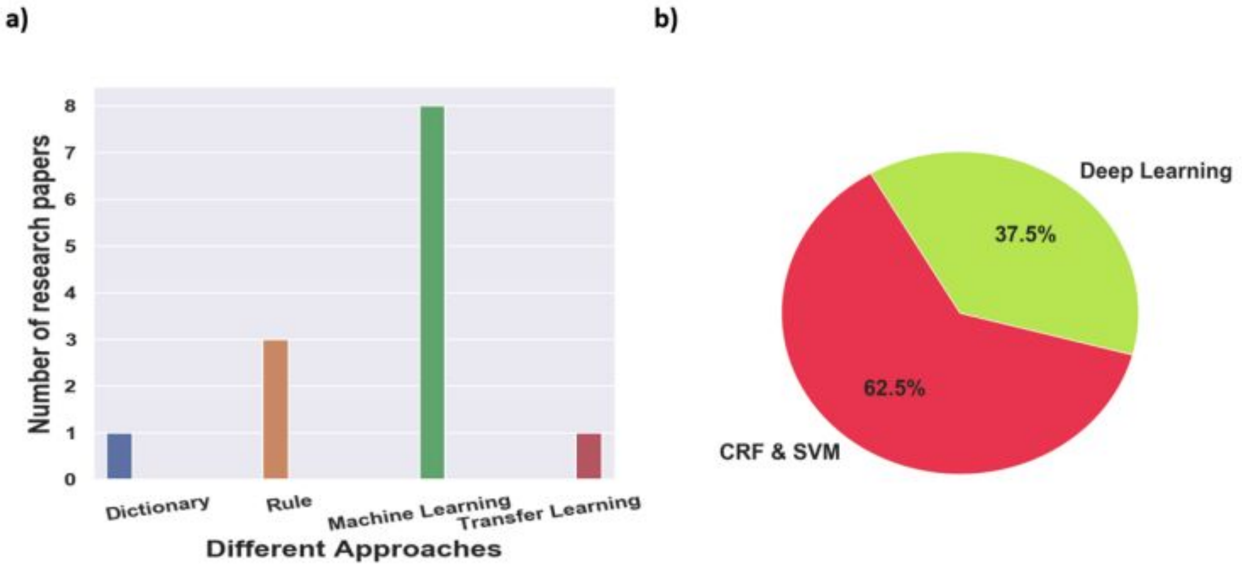


Figure 3. (**a**) Representation of the various clinical NER models based on different approaches for this survey paper and (**b**) percentage of NLP models identified based on different machine learning approaches.

Famous Google Paper in Nature

[HTML] Scalable and accurate deep learning with electronic health records

A Rajkomar, E Oren, K Chen, AM Dai, N Hajaj… - NPJ digital …, 2018 - nature.com

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning …

☆ Salvați 99 Citați Citat de 1527 ori Articole cu conținut similar Toate cele 17 versiuni

[HTML] nature.com

Full View

https://www.nature.com/articles/s41746-018-0029-1

Next

Scientific document analysis

Ethical perspective: fake news, privacy regulations (HIPPA), anonymization,...