CogStack

KING'S HEALTH PARTNERS

KING'S College LONDON

MedCAT

Medical **C**oncept **A**nnotation **T**oolkit

(github.com/CogStack/MedCAT)
(github.com/CogStack/MedCATtrainer)
(github.com/CogStack/MedCATservice)

Tom Searle
tom.searle@kcl.ac.uk

Based off original slides by Zeljko Kraljevic (zeljko.kraljevic@kcl.ac.uk)

HDRUK

NHS
National Institute for Health Research

King's College Hospital NHS Foundation Trust

NHS
South London and Maudsley NHS Foundation Trust

Motivation

- Query examples
  - All patients with Atrial Fibrillation (AF, Afib, A. Fibrillation, ...)
    - Includes spelling mistakes, acronyms, shorthand etc.
  - Contextualised mentions, i.e.:
    - Status = Affirmed
    - Experiencer = patient
    - Temporality = Current
- Use-cases
  - Clinical research questions
    - Phenotyping
    - Disease prevalence
    - Down stream aggregate analysis of symptom / finding / diagnosis / drug / procedure etc.
  - Clinical coding
  - Clinical trials - cohort selection / recruitment

Two Main Tasks

- Document Annotation
  - Entity Extraction (NER)
  - Entity Linking (+L)
- Concept Classification

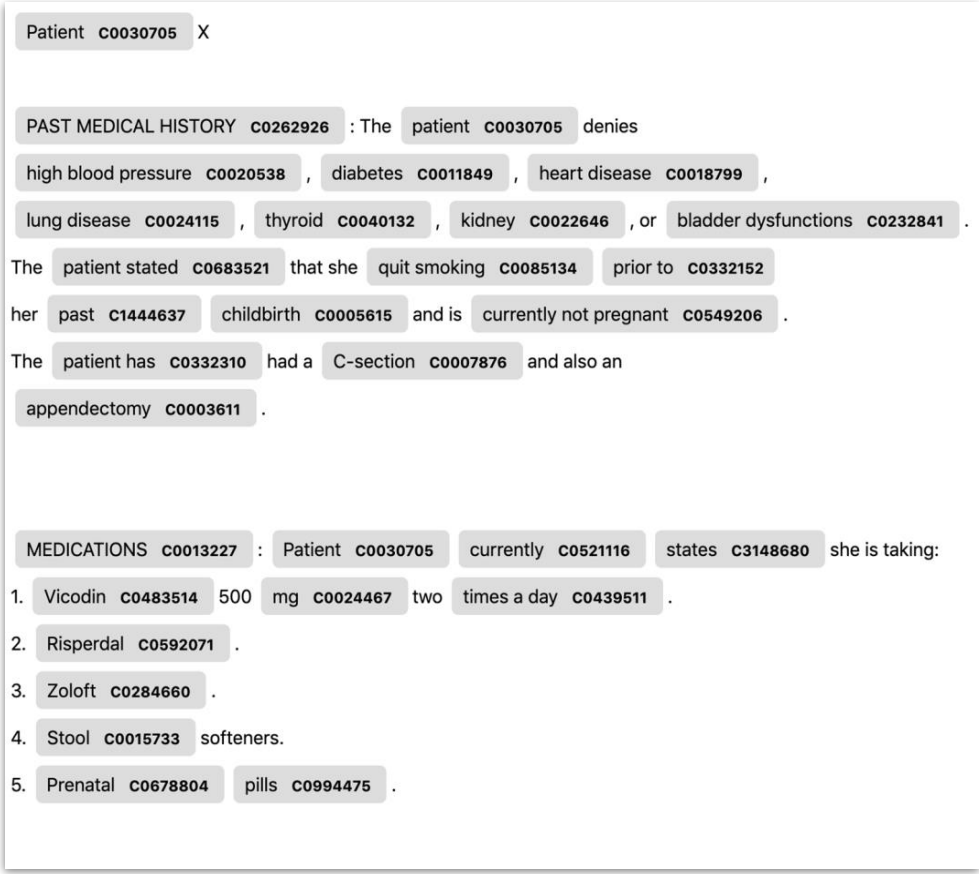
Patient X

PAST MEDICAL HISTORY: The patient denies high blood pressure, diabetes, heart disease, lung disease, thyroid, kidney, or bladder dysfunctions. The patient stated that she quit smoking prior to her past childbirth and is currently not pregnant. The patient has had a C-section and also an appendectomy.

MEDICATIONS: Patient currently states she is taking:
1. Vicodin 500 mg two times a day.
2. Risperdal.
3. Zoloft.
4. Stool softeners.
5. Prenatal pills.

Two Main Tasks

- Document Annotation
  - Entity Extraction (NER): defines start / end boundaries for a concept
  - Entity Linking (+L): links the text span with a single concept sourced from a configured knowledge base
- Concept Classification
  - Contextualises a recognised concept, i.e. Negation



Two Main Tasks

- Document Annotation
  - Entity Extraction and Linking
- Concept Classification
  - Negation, ...

pulmonary disease

cui	C0024115
tui	T047
type	Disease or Syndrome
source_value	lung disease
acc	1
Negated	True
Experiencer	Patient

Patient C0030705 X

PAST MEDICAL HISTORY C0262926 : The patient C0030705 denies high blood pressure C0020538 , diabetes C0011849 , heart disease C0018799 , lung disease C0024115 , thyroid C0040132 , kidney C0022646 , or bladder dysfunctions C0232841 . The patient stated C0683521 that she quit smoking C0085134 prior to C0332152 her past C1444637 childbirth C0005615 and is currently not pregnant C0549206 . The patient has C0332310 had a C-section C0007876 and also an appendectomy C0003611 .

MEDICATIONS C0013227 : Patient C0030705 currently C0521116 states C3148680 she is taking:

1. Vicodin C0483514 500 mg C0024467 two times a day C0439511 .
2. Risperdal C0592071 .
3. Zoloft C0284660 .
4. Stool C0015733 softeners.
5. Prenatal C0678804 pills C0994475 .

MedCAT: Concept Annotation (Entity Extraction and Linking)

Concept Database		
ID	*Names*	*Embedding*
1	Heart Rate, HR	V_1
2	Hour, HR	V_2
...		

The procedure took 3 hours, during which his heart rate was in the 60s.
Later in the evening the HR was in the 50s.
.
.
.

MedCAT: Training Procedure

Concept Database		
ID	*Names*	*Embedding*
1	Heart Rate, HR	V_1
2	Hour, HR	V_2
...		

The procedure took 3 hours, during which his heart rate was in the 60s.
Later in the evening the HR was in the 50s.

$$V_{cntx} = \frac{1}{2s}\left[\sum_{i=1}^{s} V_{w_{k-i}} + \sum_{i=1}^{s} V_{w_{k+1+i}}\right]$$

$$sim = max(0, \frac{V_{concept}}{\|V_{concept}\|} \cdot \frac{V_{cntx}}{\|V_{cntx}\|})$$

$$lr = \frac{1}{C_{concept}}$$

$$V_{concept} = V_{concept} + lr \cdot (1 - sim) \cdot V_{cntx}$$

MedCAT: Run Phase

The procedure took 3 hours, during which his heart rate was in the 60s.
Later in the evening the HR was in the 50s.

Concept Database		
ID	*Names*	*Embedding*
1	Heart Rate, HR	V_1
2	Hour, HR	V_2
...		

$$V_{cntx} = \frac{1}{2s}\left[\sum_{i=1}^{s} V_{w_{k-i}} + \sum_{i=1}^{s} V_{w_{k+1+i}}\right]$$

Word Embeddings

Disease -> Medication	*Disease -> Procedure*	*Symptom -> Medication*	*Symptom -> Everything*
Hypertensive disease	**Neoplastic Process (Cancer)**	**Fever**	**Hemorrhage**
Metoprolol 50 MG	Chemotherapy	Levofloxacin	Intracranial Hemorrhages
Metoprolol 25 MG	Radiosurgery	Vancomycin	Cerebellar hemorrhage
Valsartan 320 MG	FOLFOX Regimen	Vancomycin 750 MG	Postoperative Hemorrhage
Nadolol 20 MG	Chemotherapy Regimen	Azithromycin	Retroperitoneal Hemorrhage
Atenolol 100 MG	Preoperative Therapy	Levofloxacin 750 MG	Amyloid angiopathy
Enalapril 10 MG	Anticancer therapy	Dexamethasone	Internal bleeding
Oral form diltiazem	Parotidectomy	Lorazepam	Hematoma, Subdural, Chronic
nimodipine 30 MG	Resection of ileum	Acetaminophen	Intraparenchymal

Results and applications of MedCAT | NER + L*

Model	Hospital Test Site	# Annotated Examples	F1 μ	F1 SD±	F1 IQR
M1: Base - No Training	KCH	3,358	0.638	0.297	0.333
M2: Base + Self-Supervised MIMIC-III	KCH	3,358	0.840	0.109	0.150
M3: Base + Self-Supervised KCH	KCH	3,358	0.889	0.078	0.103
M4: KCH Self-Supervised + KCH Supervised	KCH	3,358	0.947	0.044	0.051
M4: KCH Self-Supervised + KCH Supervised	UCLH	499	0.903	0.103	0.112
M5: KCH Self-Supervised + KCH Supervised + UCLH Self-Supervised	UCLH	499	0.905	0.079	0.034
M6: KCH Self-Supervised + KCH Supervised + UCLH Self-Supervised + UCLH Supervised	UCLH	499	0.926	0.060	0.086
M4: KCH Self-Supervised + KCH Supervised	SLaM	1,425	0.885	0.095	0.088
M7: KCH Self-Supervised + KCH Supervised + SLaM Self-Supervised	SLaM	1,425	0.907	0.047	0.082
M8: KCH Self-Supervised + KCH Supervised + SLaM Self-Supervised + SLaM Supervised	SLaM	1,425	0.945	0.029	0.025

*- Kraljevic & Searle. 2020 (https://arxiv.org/abs/2010.01165)

The MedCAT Workflow

1 / 2: Prepare the environment

- Install MedCAT (pip install medcat)
- Prepare datasets
- Choose a vocabulary (existing, or bespoke)
- Choose a Concept DB or Subset
    - UMLS, SNOMED-CT, HPO etc. OR Build a custom one

3: Run unsupervised training

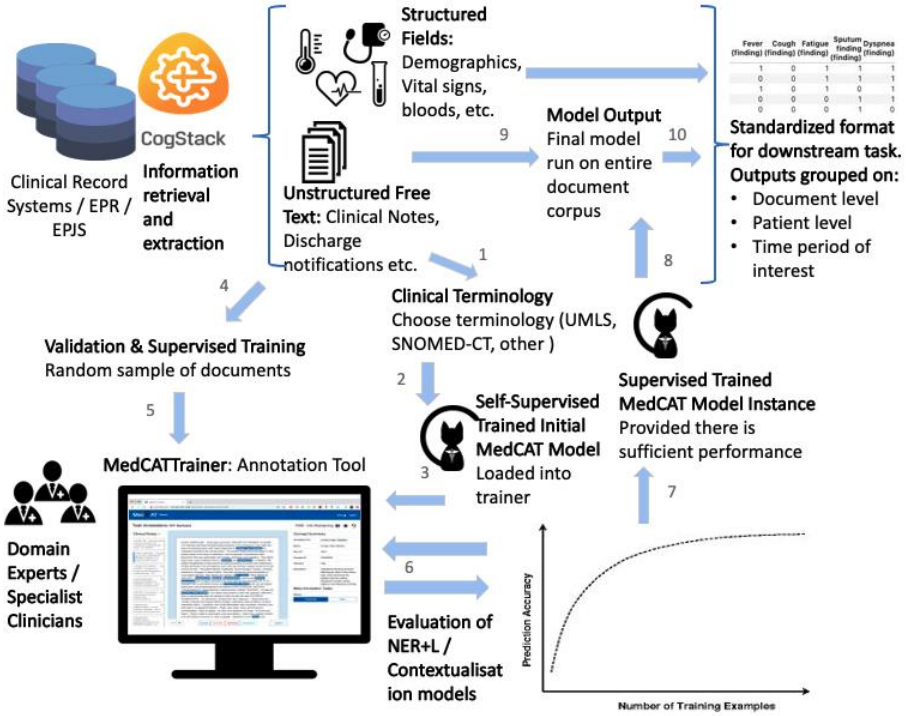
- Run over large biomedical corpora (e.g. a hospital site EPR indexed via CogStack)

4 / 5 / 6: Validate / Fine-tune using MedCATtrainer:

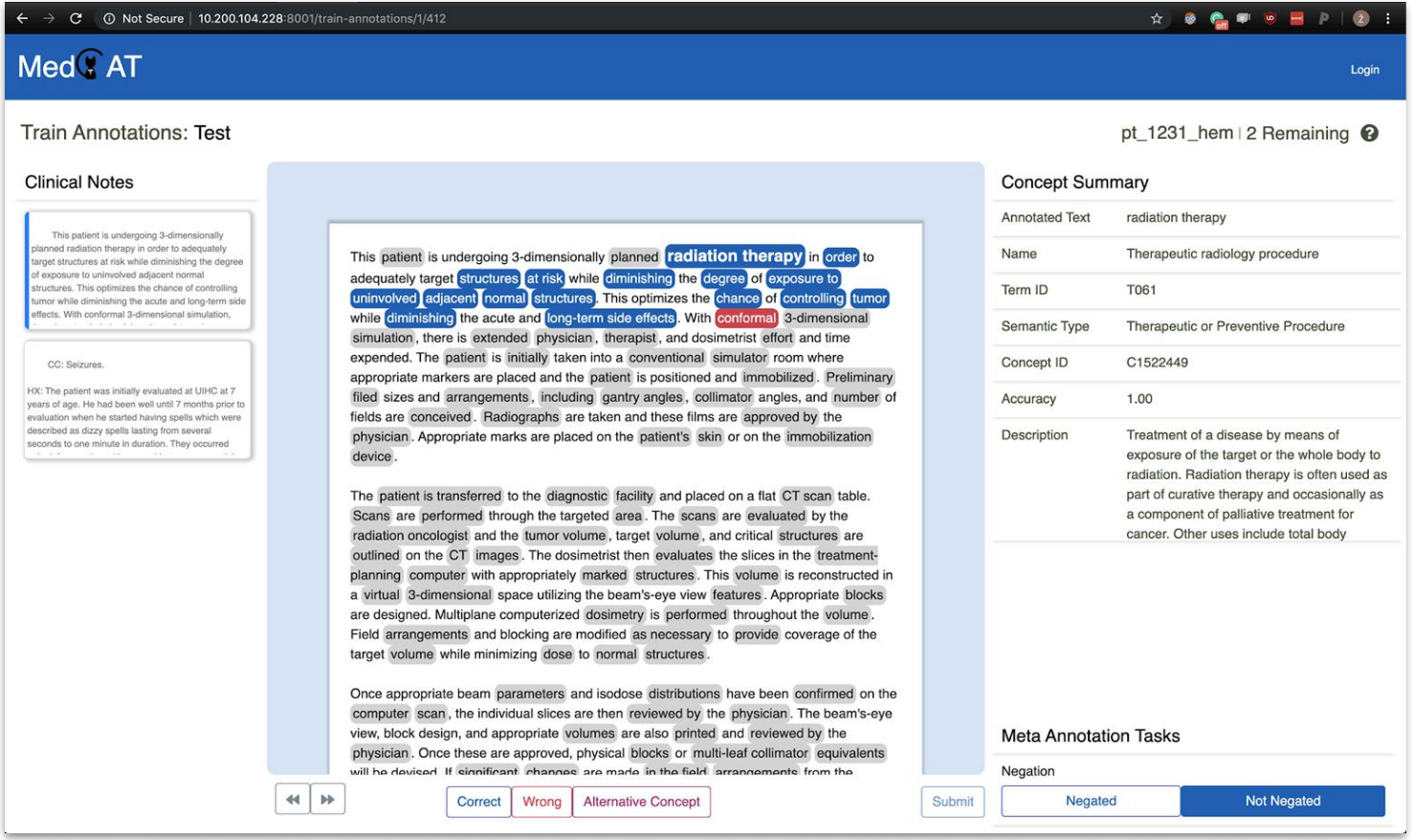
- Validate annotations with clinical teams, build model trust and measure annotation performance.
- Improve annotation performance through supervised training

7 / 8 / 9: Run final models on full dataset:

- Aggregate and structure for downstream use.



Web Application - Annotation Interface



CogStack

KING'S HEALTH PARTNERS

KING'S College LONDON

Group lead: Prof. Richard Dobson
Funding: HDR UK, NIHR Maudsley BRC

Publications:

Kraljevic & Searle et al (2020): arxiv.org/abs/2010.01165
Mascio & Kraljevic et al (2020): dx.doi.org/10.18653/v1/2020.bionlp-1.9
Bean et al (2020): doi.org/10.1002/ejhf.1924
Carr et al. (2021): doi.org/10.1186/s12916-020-01893-3
Zaker et al(2021): doi.org/10.1016/j.retram.2021.103276
Searle et al. (2020): dx.doi.org/10.18653/v1/2020.bionlp-1.8
Searle et al. (2019): dx.doi.org/10.18653/v1/D19-3024

GitHub:

- *github.com/CogStack*
- *github.com/CogStack/MedCAT*
- *github.com/CogStack/MedCATservice*
- *github.com/CogStack/MedCATtrainer*

Team

Zeljko Kraljevic
Daniel Bean
Aurelie Mascio
Lukasz Roguski
Thomas Searle
Amos Folarin
Rebecca Bendayan
James Teo
Richard Dobson

HDRUK

NHS
National Institute for Health Research

King's College Hospital NHS Foundation Trust

NHS
South London and Maudsley NHS Foundation Trust