Automated Patient Screening for Clinical Trials

Overview of the literature and challenges

Antoine Recanati with Chloé-Agathe Azencott

March, 12th 2019

Introduction : matching patients to clinical trials

Ontology + rule based feature extraction

Deep (representation) learning methods ?

Conclusion

Introduction : matching patients to clinical trials

Clinical Trials

- Procedure to assess new drug safety and efficiency
- Need to select (screen) cohort of patients satisfying *eligibility criteria*

Clinical Trials

- Procedure to assess new drug safety and efficiency
- Need to select (screen) cohort of patients satisfying *eligibility criteria*
- Screening usually done **manually**, very **time consuming** (bottleneck in the CT process)

Clinical Trials

- Procedure to assess new drug safety and efficiency
- Need to select (screen) cohort of patients satisfying *eligibility criteria*
- Screening usually done **manually**, very **time consuming** (bottleneck in the CT process)
- Generalization of **electronic health records (EHRs)** can alleviate such tasks

Typical Clinical Trial

- Title, Summary, Condition name, Interventions
- List of **inclusion** and **exclusion** criteria (free text)
- `https://clinicaltrials.gov`

Electronic Health Record (EHR)

EHRs of hospital patients typically contains

- **Structured** data (age, demographic data, treatments, physical characteristics : BMI, blood pressure, *etc.*)
- **Unstructured** (free text) data (clinical narratives, progress notes, imaging reports, discharge summaries)

Data

- Clinical trials descriptions : all on `https://clinicaltrials.gov`
- EHRs from patients : 50000 deidentified EHRs (for research, English) (without matching data)

Formalization of the matching problem

$x \in \mathcal{X}$ represents a patient's EHR

$y \in \mathcal{Y}$ represents a trial (list of criteria)

Goal :

$$\begin{array}{ll} \text{find} & f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\} \\ \text{such that} & f(x, y) = 1 \quad \text{iff} \quad x \in \textbf{Elig}(y) \quad (x \text{ is eligible for } y). \end{array}$$

Metrics ?

Given $x_1, \ldots, x_p$ patient records, $y_1, \ldots, y_T$ trials, and $M \in \{0,1\}^{p \times T}$ assignment matrix such that $M_{i,j} = 1$ if patient i participated in trial j and 0 otherwise,

$$P = \sum_{trial\ j} \frac{\sum_{patient\ i} f(x_i, y_j) M_{i,j}}{\sum_{patient\ i} f(x_i, y_j)}$$

$$R = \sum_{trial\ j} \frac{\sum_{patient\ i} f(x_i, y_j) M_{i,j}}{\sum_{patient\ i} M_{i,j}}$$

Metrics ? (ctd.)

$$R = \sum_{trial\ j} \frac{\sum_{patient\ i} f(x_i, y_j) M_{i,j}}{\sum_{patient\ i} M_{i,j}}$$

Metrics ? (ctd.)

$$R = \sum_{trial\ j} \frac{\sum_{patient\ i} f(x_i, y_j) M_{i,j}}{\sum_{patient\ i} M_{i,j}}$$

- $M_{i,j} \neq \mathbb{1}[x_i \in \mathbf{Elig}(y_j)]$; PU learning ?

Metrics ? (ctd.)

$$R = \sum_{trial\ j} \frac{\sum_{patient\ i} f(x_i, y_j) M_{i,j}}{\sum_{patient\ i} M_{i,j}}$$

- $M_{i,j} \neq \mathbb{1}[x_i \in \mathbf{Elig}(y_j)]$; PU learning ?
- Metric of interest : time spent by doctor within acceptable recall interval

Metrics ? (ctd.)

$$R = \sum_{trial\ j} \frac{\sum_{patient\ i} f(x_i, y_j) M_{i,j}}{\sum_{patient\ i} M_{i,j}}$$

- $M_{i,j} \neq \mathbb{1}[x_i \in \textbf{Elig}(y_j)]$; PU learning ?
- Metric of interest : time spent by doctor within acceptable recall interval
- Leverage common criteria across different trials ?

Formalization of the matching problem (ctd.)

Each trial = combination of inclusion / exclusion criteria.
$z \in \mathcal{Z}$ represents a criterion
$y_j = (z_j^{(1)}, \dots, z_j^{(n_j)})$ Goal :

$$\begin{array}{ll} \text{find} & \phi : \mathcal{X} \times \mathcal{Z} \to \{0,1\} \\ \text{such that} & \phi(x,z) = 1 \quad \text{iff} \quad x \in \textbf{Elig}(z) \quad (x \text{ satisfies } z). \end{array}$$

And $\tilde{M}_{i,k} = M_{i,j}$ for $k = 1, \dots, n_j$, for all trial j.

Challenges

- Division into **atomic criteria** / relation between criteria (NER)

Challenges

- Division into **atomic criteria** / relation between criteria (NER)
- Synonyms, misspellings, **equivalent formulations**

Challenges

- Division into **atomic criteria** / relation between criteria (NER)
- Synonyms, misspellings, **equivalent formulations**
- Still $\tilde{M}_{i,k} \neq \mathbb{1}[x_i \in \textbf{Elig}(z_k)]$

Challenges

- Division into **atomic criteria** / relation between criteria (NER)
- Synonyms, misspellings, **equivalent formulations**
- Still $\tilde{M}_{i,k} \neq \mathbb{1}[x_i \in \textbf{Elig}(z_k)]$
- **No matching data** *yet*. Can we still make progress using proxys ?

Intermission : ICD10 classification

International Classification of Diseases (codes with descriptive sentence to tag patients' diseases. Essentially used for billing)

- **ICD-10 Version:2016**
  - I Certain infectious and parasitic diseases
  - II Neoplasms
    - C00-C97 Malignant neoplasms
      - C00-C75 Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue
      - C76-C80 Malignant neoplasms of ill-defined, secondary and unspecified sites
        - C76 Malignant neoplasm of other and ill-defined sites
        - C77 Secondary and unspecified malignant neoplasm of lymph nodes
          - C77.0 Secondary and unspecified malignant neoplasm: Lymph nodes of head, face and neck
          - C77.1 Secondary and unspecified malignant neoplasm: Intrathoracic lymph nodes
          - C77.2 Secondary and unspecified malignant neoplasm: Intra-abdominal lymph nodes
          - C77.3 Secondary and unspecified malignant neoplasm: Axillary and upper limb lymph nodes
          - C77.4 Secondary and unspecified malignant neoplasm: Inguinal and lower limb lymph nodes
          - C77.5 Secondary and unspecified malignant neoplasm: Intrapelvic lymph nodes
          - C77.8 Secondary and unspecified malignant neoplasm: Lymph nodes of multiple regions
          - C77.9 Secondary and unspecified malignant neoplasm: Lymph node, unspecified
        - C78 Secondary malignant neoplasm of respiratory

C77.1	**Intrathoracic lymph nodes**
C77.2	**Intra-abdominal lymph nodes**
C77.3	**Axillary and upper limb lymph nodes** Pectoral lymph nodes
C77.4	**Inguinal and lower limb lymph nodes**
C77.5	**Intrapelvic lymph nodes**
C77.8	**Lymph nodes of multiple regions**
C77.9	**Lymph node, unspecified**
C78	**Secondary malignant neoplasm of respiratory and digestive organs**
C78.0	**Secondary malignant neoplasm of lung**
C78.1	**Secondary malignant neoplasm of mediastinum**
C78.2	**Secondary malignant neoplasm of pleura** Malignant pleural effusion NOS
C78.3	**Secondary malignant neoplasm of other and unspecified respiratory organs**
C78.4	**Secondary malignant neoplasm of small intestine**
C78.5	**Secondary malignant neoplasm of large intestine and rectum**
C78.6	**Secondary malignant neoplasm of retroperitoneum and peritoneum** Malignant ascites NOS
C78.7	**Secondary malignant neoplasm of liver and intrahepatic bile duct**
C78.8	**Secondary malignant neoplasm of other and unspecified digestive organs**
C79	**Secondary malignant neoplasm of other and unspecified sites**
C79.0	**Secondary malignant neoplasm of kidney and renal pelvis**

Intermission : ICD10 classification

International Classification of Diseases (codes with descriptive sentence to tag patients' diseases. Essentially used for billing)

- Well-posed classification (multilabel or multiclass) problem : input EHRs, output : ICD code (class)
- CNN works well with input text EHRs (Mullenbach et al. 2018)

How to represent (vectorize) x and z ?

- To structure or not to structure the data ?

How to represent (vectorize) x and z ?

- To structure or not to structure the data ?
- ICD10 classification : works well with CNNs to represent x but well-posed and large amount of labeled data.

How to represent (vectorize) x and z ?

- To structure or not to structure the data ?
- ICD10 classification : works well with CNNs to represent x but well-posed and large amount of labeled data.
- Here, x and z is text. Represent x and z in same space (translation-like problem ?)

How to represent (vectorize) x and z ?

- To structure or not to structure the data ?
- ICD10 classification : works well with CNNs to represent x but well-posed and large amount of labeled data.
- Here, x and z is text. Represent x and z in same space (translation-like problem ?)
- Old-fashioned NLP : use ontology + NER to extract features. Broadly used for clinical text.

Ontology + rule based feature extraction

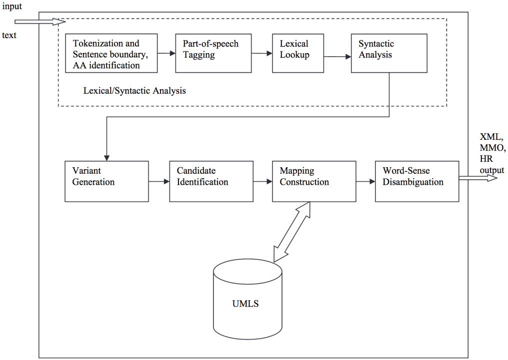
Ontologies for clinical text

- ICD10 : disease codes with descriptive sentences
- MeSH (Medical Subject Headings) : thesaurus of controlled vocabulary used for PubMed indexing. Each term has short description and relations to other terms
- SNOMED CT : hiearchical+relational structure between classes of concepts
- UMLS : “Meta-thesaurus”. Millions of concept codes associated with descriptives and relations between them

Mapping text to clinical concepts

Tools using NER and/or UMLS (parse text and map to concepts)

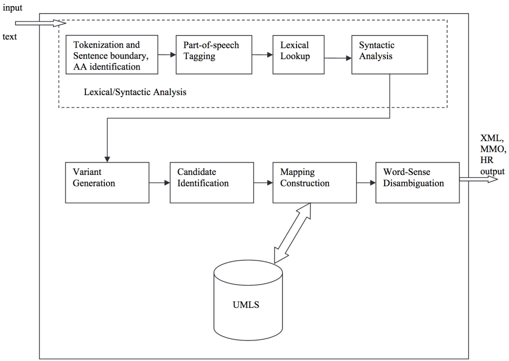
- MetaMap (`https://ii.nlm.nih.gov/Interactive/UTS_Required/metamap.shtml`)(Figure from Aronson & Lang (2010)), cTAKES, DNorm



Mapping text to clinical concepts

Tools using NER and/or UMLS (parse text and map to concepts)

- MetaMap (https://ii.nlm.nih.gov/Interactive/UTS_Required/metamap.shtml)(Figure from Aronson & Lang (2010)), cTAKES, DNorm
- ConText, NegEx : regex-based tools to find negative or context (family) in medical documents



Finding patients for clinical trials : text search

Garcelon et al. (2016)

- context of rare diseases : text search may be sufficient
- family history important (e.g. father has Crohn disease)
- Text search + negation and context (family) yields good performance

Finding patients for clinical trials : use mapping to ontology to find similar patients

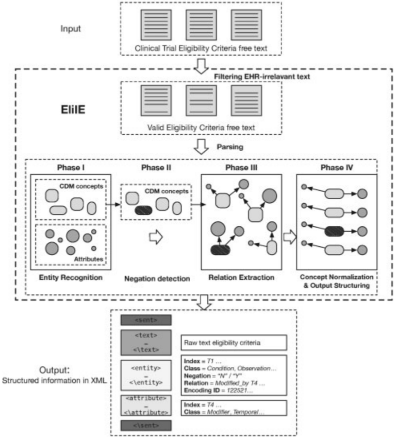
Garcelon et al. (2017)

- context of rare diseases : sparse set of relevant clinical concepts
- Method : map EHR to UMLS concepts to find representation vector of patients
- (Incorporate context and negation disambiguation)
- Given patient with rare disease, identify potentially similar patients based on their EHR

Use ontology-based mapping to extract information from clinical trials description

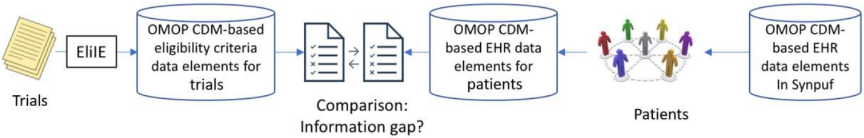
Kang et al. (2017)

- Goal : structure concepts in EC with terminology common to EHRs concepts (“normalization”)
- Specific entity recognition for eligibility criteria (relation between criteria, *etc.*)
- Fine-tuned on Alzheimer’s disease eligibility criteria



Join the dots between CT and EHRs : “the data gap”

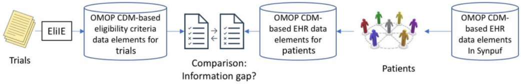
Butler et al. (2018)



Join the dots between CT and EHRs : "the data gap"

Butler et al. (2018)

- Goal : Assess intersection of concepts extracted from EC and EHRs



Join the dots between CT and EHRs : “the data gap”

Butler et al. (2018)

- Goal : Assess intersection of concepts extracted from EC and EHRs
- Involves manual unification of the clinical terms in EC before concept extraction

Table 1. Manual revision of clinical entities.

Types of Revision	Example	Times
Formatting; Typo	delerium -> delirium	207
Formatting; Plural	cancers -> cancer	253
Formatting; removal of non-informative words	heart rate measurement -> heart rate	364
Formatting; removal of abbreviations	absolute neutrophil count (ANC) -> absolute neutrophil count	1768
Simplification	asthmatic conditions -> asthma	573
Breaking down long phrases to logically-connected single phrases	basal or squamous cell carcinoma -> basal cell carcinoma or squamous cell carcinoma	445
Total		**3610**

Join the dots between CT and EHRs : “the data gap”

Butler et al. (2018)

- Goal : Assess intersection of concepts extracted from EC and EHRs
- Involves manual unification of the clinical terms in EC before concept extraction
- Also on Alzheimer’s disease data

Table 1. Manual revision of clinical entities.

Types of Revision	Example	Times
Formatting; Typo	delerium -> delirium	207
Formatting; Plural	cancers -> cancer	253
Formatting; removal of non-informative words	heart rate measurement -> heart rate	364
Formatting; removal of abbreviations	absolute neutrophil count (ANC) -> absolute neutrophil count	1768
Simplification	asthmatic conditions -> asthma	573
Breaking down long phrases to logically-connected single phrases	basal or squamous cell carcinoma -> basal cell carcinoma or squamous cell carcinoma	445
Total		**3610**

Join the dots between CT and EHRs : "the data gap"

Butler et al. (2018)

- Goal : Assess intersection of concepts extracted from EC and EHRs
- Involves manual unification of the clinical terms in EC before concept extraction
- Also on Alzheimer's disease data
- Intersection not so broad

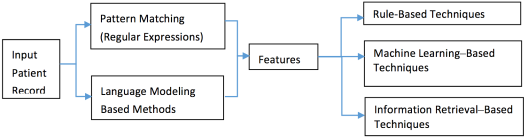
Table 4. The top 20 common SNOMED CT terms in AD trials and their prevalence in EHR dataset.

SNOMED CT Term	SNOMED-CT ID	Trial Count	Prevalence in Trials	Count of uses in EHR data for AD patients
Alzheimer's disease	26929004	972	64.29%	30,262
Mini-mental state examination	273617000	705	46.63%	**0**
Presenile dementia	12348006	599	39.62%	7,089
Disease	64572001	555	36.71%	12,029,900
Current chronological age	424144002	515	34.06%	**0**
Mental disorder	74732009	499	33.00%	505,870
Magnetic resonance imaging	113091000	482	31.88%	63,171
Cerebrovascular accident	230690007	371	24.54%	4
Global assessment of functioning - 1993 Diagnostic and Statistical Manual of Mental Disorders- ver.4th	284061009	361	23.88%	**0**
Systemic disease	56019007	353	23.35%	**0**
Disorder of nervous system	118940003	335	22.16%	780,478
Substance abuse	66214007	279	18.45%	9,466
Parkinson's disease	49049000	275	18.19%	**0**
Impaired cognition	386806002	260	17.20%	13,375
Seizure disorder	128613002	240	15.87%	28,586
Hypersensitivity reaction	421961002	218	14.42%	4,686
Schizophrenic disorders	191526005	216	14.29%	40777
History of clinical finding in subject	417662000	207	13.69%	189,543
Risk identification: childbearing family	386414004	205	13.56%	**0**
Clinical dementia rating scale	273367002	204	13.49%	**0**

Extract information from EHRs: domain specific rules

Adupa et al. (2016)

- EHR information extraction method for a given clinical trial (PARAGON)



Extract information from EHRs: domain specific rules

Adupa et al. (2016)

- EHR information extraction method for a given clinical trial (PARAGON)
- Domain specific rules (Heart Failure)

Table 4. Regular expressions for extracting LVEF-containing sentences and values.

<table>
<tr><th>S/N</th><th>Regular Expression</th></tr>
<tr><td>1</td><td>(left ventricular ejection fraction|lvef|lv ejection fraction|left ventricle ejection fraction|ejection fraction|ef |ejection fraction)[^_%\\.]*?([\\d-\\.]+)\\s*'?%</td></tr>
<tr><td>2</td><td>(left ventricular systolic function|left ventricular function|systolic function of the left ventricle|lv systolic function|left ventricular ejection fraction|ejection fraction|left ventricle)(normal|normal global|low normal|well preserved|severely reduced|moderately decreased|moderately depressed|severely decreased|severe|markedly decreased|markedly reduced|severely globally reduced|mildly decreased|mildly depressed|severely depressed)</td></tr>
<tr><td>3</td><td>(normal|normal global|low normal|well preserved|severely reduced|moderately decreased|moderately depressed|severely decreased|severe|markedly decreased|markedly reduced|severely globally reduced|mildly decreased|mildly depressed|severely depressed)</td></tr>
<tr><td>4</td><td>.*(moderate|marked|severe) (lv systolic dysfunction|left ventricular dysfunction|left ventricular systolic dysfunction).*</td></tr>
<tr><td>5</td><td>((\\d+\\s*(\\-|to)\\s*\\d+)|(\\d*\\.\\d*\\s*(\\-|to)\\s*\\d*\\.\\d*)|(\\d*\\.\\d+)|(\\d+))(?=(\\s*(\\%)))</td></tr>
<tr><td>6</td><td>\\d+(\\.\\d+)?</td></tr>
</table>

Extract information from EHRs: domain specific rules

Adupa et al. (2016)

- EHR information extraction method for a given clinical trial (PARAGON)
- Domain specific rules (Heart Failure)
- Goal : save time for prescreening with high recall

		Prescreening Gold Standard (Manual)	
		Patients Included	Patients Excluded
Classification outcome (algorithmic)	Patients included	38	6
	Patients excluded	2	152

Deep (representation) learning methods ?

- Think of Computer Vision
- Now transfer learning works with text too (BERT, ELMO, etc.)
- Unsupervised methods ? (Word2Vec)
- Yet, not always satisfying in domain-specific tasks (even in CV)

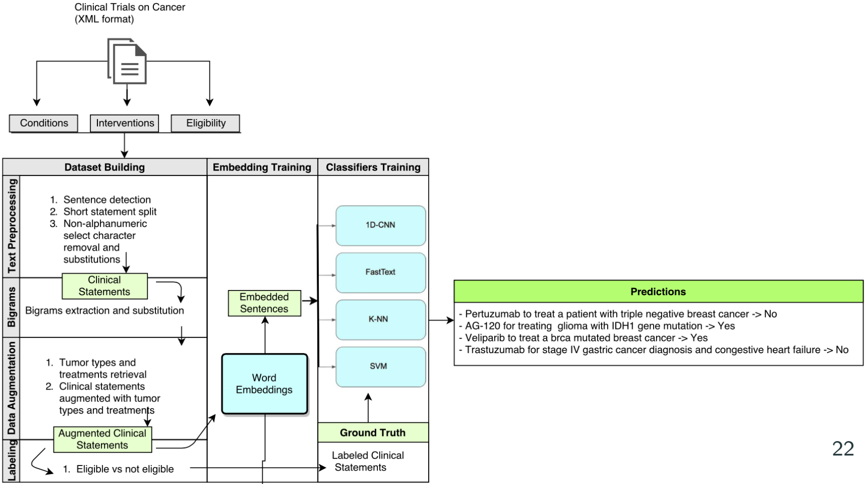
Training deep representation of clinical trials with a random classification task

Bustos & Pertusa (2018)

- Goal : train deep neural network (CNN) to obtain accurate embedding of clinical text (words)
- Task : classify statements as True or False (Eligible / Not eligible)
- Data : uses data from `clinicaltrials.gov` only) to generate data (labeling given by inclusion/exclusion, data augmentation through simple sentences)
- Belief in the magic of word embeddings

Training deep representation of clinical trials with a random classification task

Bustos & Pertusa (2018)



Training deep representation of clinical trials with a random classification task

Bustos & Pertusa (2018)

Condition	Intervention
Rectal Neoplasms	Procedure: Irreversible electroporation (IRE)

Criteria

Inclusion Criteria:

- Rectal Neoplasms diagnosed by positive biopsy or non-invasive criteria,
- Not suitable for surgical resection,
- Eastern Cooperative Oncology Group (ECOG) score of 0-1,
- A prothrombin time ratio > 50%,
- Platelet count > 80x10^9/L,
- Ability of patient to stop anticoagulant and anti-platelet therapy for seven days prior to and seven days post NanoKnife procedure,
- Able to comprehend and willing to sign the written informed consent form (ICF),
- Have a life expectancy of at least 3 months.

Exclusion Criteria:

- Cardiac insufficiency, ongoing coronary artery disease or arrhythmia,
- Any active implanted device (eg Pacemaker),
- Women who are pregnant or women of child-bearing potential who are not using an acceptable method of contraception,
- Have received treatment with an investigational agent/ procedure within 30 days prior to treatment with the NanoKnife™ LEDC System,
- Are in the opinion of the Investigator unable to comply with the visit schedule and protocol evaluations.

A. Original Source: https://clinicaltrials.gov/ct2/show/NCT02425059

Label	Clinical Statement
Eligible	*study intervention is irreversible electroporation . rectal neoplasm and not suitable for surgical resection*
Not eligible	*study intervention is irreversible electroporation . rectal neoplasm and cardiac insufficiency ongoing coronary artery disease or arrhythmia*

B. Extracted features after preprocessing

Conclusion

Summary, TODOs, challenges and open questions

- Matching unstructured text data (EHRs) to unstructured text (Clinical Trials)
- Goal : prescreen patients with high recall, and provide reasonable number of patients for manual screening
- Domain restriction allows information retrieval with specifically designed rules (*e.g.*, Alzheimer's or Heart Failure)
- Degree of precision for matching also depends on domain restriction (*e.g.*, just output patients with "Heart Failure" in their EHR ?)
- Evaluate baselines (text-search and concept mapping tools)
- Make progress without matching data (other, simpler task (*e.g.*, classification of diseases))
- Annotate data ?
- Reliably augment the matching data (*e.g.*with patient similarity, or leveraging external corpus or ontology)

References

Adupa, A. K., Garg, R. P., Corona-Cox, J., Shah, S., Jonnalagadda, S. R. et al. (2016), 'An information extraction approach to prescreen heart failure patients for clinical trials', *arXiv preprint arXiv:1609.01594* .

Aronson, A. R. & Lang, F.-M. (2010), 'An overview of metamap: historical perspective and recent advances', *Journal of the American Medical Informatics Association* **17**(3), 229–236.

Bustos, A. & Pertusa, A. (2018), 'Learning eligibility in cancer clinical trials using deep neural networks', *Applied Sciences* **8**(7), 1206.

Butler, A., Wei, W., Yuan, C., Kang, T., Si, Y. & Weng, C. (2018), 'The data gap in the ehr for clinical research eligibility screening', *AMIA Summits on Translational Science Proceedings* **2017**, 320.

Garcelon, N., Neuraz, A., Benoit, V., Salomon, R. & Burgun, A. (2016), 'Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse', *Journal of the American Medical Informatics Association* **24**(3), 607–613.

Garcelon, N., Neuraz, A., Benoit, V., Salomon, R., Kracker, S., Suarez, F., Bahi-Buisson, N., Hadj-Rabia, S., Fischer, A., Munnich, A. et al. (2017), 'Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. warehouse and the needle in the needle stack', *Journal of biomedical informatics* **73**, 51–61.

Kang, T., Zhang, S., Tang, Y., Hruby, G. W., Rusanov, A., Elhadad, N. & Weng, C. (2017), 'Eliie: An open-source information extraction system for clinical trial eligibility criteria', *Journal of the American Medical Informatics Association* **24**(6), 1062–1071.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. (2018), 'Explainable prediction of medical codes from clinical text', *arXiv preprint arXiv:1802.05695* .